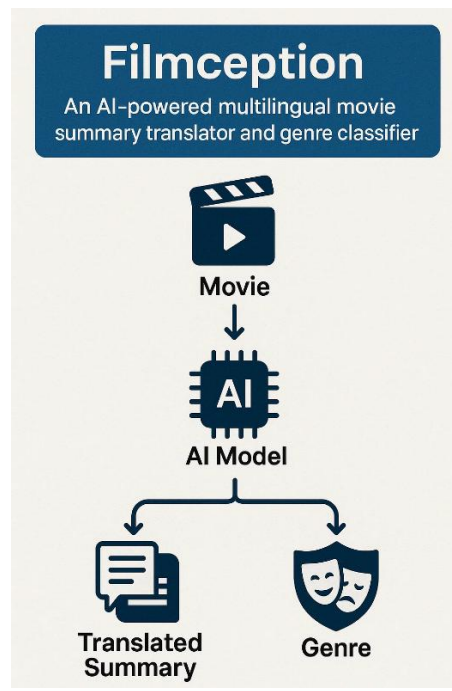# Artificial Intelligence – Spring 2025 Semester Project

Filmception

An AI-powered Multilingual movie summary translator and genre classifier

Deadline: 6th May, 2025

General Instructions:

- This assessment item follows the following CLOS:
    - **[CLO - 4]** General understanding of major concepts and approaches in knowledge representation, planning, learning, robotics and other AI areas.

    - **[CLO - 5]** Developing programming skills for AI applications.

- This project is to be done in pairs.

- The deadline will not be extended under any circumstances.

- Evaluation will be done on the basis of a demo. Anyone failing to appear for the demo or unable to give up to the mark demonstration will be awarded zero.

Department of Cyber Security

Overview

This project is designed to develop a comprehensive system for processing movie summaries, predicting movie genres, and converting movie summaries into audio formats. The goal is to create a user-interactive menu-based system where:

- A user can input a **movie summary**.

- The system will offer the option to **convert the summary into audio** in multiple languages (Arabic, Urdu and Korean).

- The system will also provide an option to **predict the movie's genre(s)** based on the summary using a machine learning model.

The system will allow users to select language preferences for audio conversion, as well as the option to classify the movie into one or more genres based on the machine learning model trained on the CMU Movie Summary dataset.

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

Project Components

1. Data Preprocessing and Cleaning

- **Summary Extraction and Cleaning**:

  - You will first preprocess the movie summaries extracted from the **CMU Movie Summary Dataset**. The preprocessing will involve the following:

    - **Removing special characters, stopwords, and redundant spaces**.

    - **Lowercasing** the text to standardize it.

    - **Tokenization** to break down the summaries into individual words or phrases.

    - **Stemming/Lemmatization** to reduce words to their base or root form.

    - **Removing non-relevant words** like numbers and punctuations that don't add value to the genre classification.

- **Metadata Extraction**:

  - From the **movie.metadata.tsv** file, extract the **genre information** for each movie. The genres are provided as multi-labels, indicating that each movie can belong to multiple genres (e.g., "Action", "Adventure", "Comedy").

- o The final output from this preprocessing should be a **new cleaned file** containing the **Movie ID**, **summary**, and **genres** (one or more). This file will be used for training the genre prediction model.

- **Train-Test Split**: You have to perform a **train-test split** on the dataset. The split should be carefully chosen to prevent data leakage and ensure robust model evaluation.

2. Text Translation and Audio Conversion

- **Text Translation**:

  - o Once the movie summaries are cleaned and prepared, **translate** each summary into **Arabic, Urdu and Korean**. This will test the capability to work with multilingual data.

  - o You can use existing translation tools or APIs such as **Google Translate**, **MarianMT** (Hugging Face), or **DeepL** to achieve this.

- **Audio Conversion**:

  - o After translation, the **translated summaries** will be converted into audio using a **Text-to-Speech (TTS)** engine (such as **gTTS**, **pyttsx3**, or **Amazon Polly**).

  - o The audio will be available for playback, and the user will be able to choose which **language** they want to listen to. This allows the user to hear the movie summaries in multiple languages based on their preference.

- **At least 50 movie summaries from the data set should be translated, converted to audios and saved.**

3. Movie Genre Prediction Model

- **Model Development**:

  - o You have to will build a **machine learning model** to predict the **genres** of movies based on their summaries. This will be a **multi-label classification** problem, where a movie can belong to more than one genre.

  - o You have the freedom to choose the model architecture. The simpler models might include **Logistic Regression** or **Random Forests**, while more complex models can include **Deep Learning models** like **LSTMs**, **CNNs**, or **Transformer-based models** (like **BERT**).

- o **Feature Extraction**: You will need to extract features from the movie summaries. This could involve using **TF-IDF**, **word embeddings** (like **Word2Vec**, **GloVe**), or even directly using **pre-trained transformers** for text representation.

- **Evaluation**:

  - o The model's performance will be evaluated using the following metrics:

    - ▪ **Accuracy**: The overall accuracy of the model.

    - ▪ **Precision, Recall, F1-Score**: These metrics will help assess the performance of the model in terms of identifying specific genres.

    - ▪ **Confusion Matrix**: To visualize the classification results for the training and test sets.

4. Interactive User Interface (Menu-Based System)

The core of this project is an interactive, **menu-driven system**. The workflow will be as follows:

1. **Input**: The user inputs a movie summary.

2. The system will display the following options:

   - o **Option 1: Convert Summary to Audio**:

     - ▪ The user selects the **language** for the audio (e.g., English, Spanish, French).

     - ▪ The system generates the **audio** and plays it for the user.

   - o **Option 2: Predict Genre**:

     - ▪ The user selects the **option to classify** the genre of the movie based on the summary.

     - ▪ The system will output the predicted genre(s) of the movie.

The system should be designed in a way that allows users to explore both features sequentially.

The detailed breakdown of the evaluation criteria is given below

Component	Criteria	Marks
Preprocessing and Data Cleaning	Deep cleaning of movie summaries (lowercasing, punctuation removal, etc.)	10
	Tokenization, stop-word removal, lemmatization	10
	Genre extraction and multi-label formatting from metadata	15
	Creation of final clean dataset (Movie ID, summary, genres)	10
Translation and Audio Conversion	Accurate translation into 2-3 different languages	20
	High-quality audio conversion of translations	20
	Proper storage and menu-based playback options	10
Genre Prediction Model	Model selection (simple or deep architectures, justified with reasoning)	25
	Feature extraction (TF-IDF, embeddings, etc.) and input processing	20
	Handling of multi-label classification	20
	Training and test set preparation with justified split ratio	5
	Evaluation metrics (accuracy, precision, recall, F1-score)	20
	Confusion matrix for training and testing sets	10
User Interface	Menu-based interaction for translation/audio/genre classification	15
	Smooth flow and user experience	5
	Error handling and output messages	5
Documentation and Code Quality	Modular, well-documented, and readable code	5
	Clear and complete explanation of workflow in a pdf report	5
	Use of visualizations (graphs, sample outputs, plots)	5
Demo	Q/A	15
Total		**250**
	Bonus: Additional functionality (e.g., A GUI to interact with your system)	20