Movie Recommendation System Using Transformer Models and Plot Summaries

Team #19
Members:

1. Arvind Chary Padala (ap2522),
2. Ram Sampreeth Budireddy(rb1424),
3. Tanishq Sharma (ts1266)

Project Description:

This project aims to develop a plot-based movie recommendation system by leveraging plot summaries and metadata from the “CMU Movie Summary Corpus”. The CMU dataset provides over 42,306 movie plot summaries, along with rich metadata, including genre, release date, language, and character information. This dataset will allow us to explore and recommend movies to users based on genre and plot similarity.

In our system, users can express their interest in watching a specific type of movie, with a summary of the desired movie plot. We will address two main tasks in this project:

1. Genre Classification: Classify the genre of a user’s input query to understand the desired movie genre.
2. Plot Similarity Matching: Once the genre is identified, recommend similar movies by finding other movies in the same genre with similar plot structures.

NLP Tasks Involved:

1. Classification: For genre classification, we will classify user input into a genre based on plot keywords and expressions.
2. Sequence Similarity Matching: For plot matching, we’ll generate embeddings for plot summaries and match the user’s genre preferences to similar plot structures.

3. Named Entity Recognition: If the user query includes some specific entity(Actor, Location, or character), to handle such requests NER would be utilized.
4. Recommendation based on user input: Based on the genre classification and sequence similarity, we recommend watching movie titles (ranked).

Dataset: link

The “CMU Movie Summary Corpus” will serve as the primary dataset for this project. It includes:
- Plot summaries for 42,306 movies.
- Genre labels, which provide the target labels for our genre classification task.
- Additional metadata, such as movie release date, language, and runtime, will help refine recommendations further.

Proposed Models and Approach:

We will compare our results against baseline models and then implement more sophisticated neural network architectures:

1. Baseline Models:
   a. Naive Bayes or Logistic Regression: We’ll use traditional methods like Naive Bayes and Logistic Regression trained on tokenized plot summaries for genre classification.
   b. TF-IDF Vector Similarity: For initial plot similarity matching, we will use TF-IDF vectors to find movies with similar plot structures within the chosen genre.
2. Neural Network Models:
   Pre-trained Transformer Model(BERT or DistilBERT): We will fine-tune a transformer model for genre classification. By training on plot summaries with genre labels, the model can classify user input text into the appropriate genre as well as recommend titles based on the user input plot. This will allow the model to capture more context and nuance than traditional models. If we are not able to get promising results, we may further try using encoder-decoder models (BART, T5, or Sentence-BERT).

Model Training Plan:

1. We will use 80% of the dataset for training, 10% for validation, and 10% for testing to ensure our models generalize well to unseen data.
2. The pre-trained transformer model will be fine-tuned on genre labels, and movie plots using the cleaned plot summaries as input.
3. After generating embeddings for each plot summary, we will store them for real-time similarity calculations within the predicted genre.
4. Dropout, regularization, and learning rate adjustments will be applied to improve model performance.

Evaluation Metrics:

1. Genre Classification: We will use Precision, Recall, F1, and, ChrF scores to measure the accuracy of genre predictions.
2. Plot Similarity Matching: Cosine similarity will be used to measure relevance, and recommendations will be validated by qualitative analysis (e.g., user feedback) or relevance scoring within genre categories.

Research Papers:

We will refer to and build on the methodologies discussed in these papers:

1. L. Cai, Y. Song, T. Liu, and K. Zhang, "A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification," in IEEE Access, vol. 8, pp. 152183-152192, 2020, doi:10.1109/ACCESS.2020.3017382.
2. Yarullin, R., Serdyukov, P. (2021). BERT for Sequence-to-Sequence Multi-label Text Classification. In: van der Aalst, W.M.P., et al. Analysis of Images, Social Networks, and Texts. AIST 2020. Lecture Notes in Computer Science(), vol 12602. Springer, Cham. https://doi.org/10.1007/978-3-030-72610-2_14.

This project will contribute to understanding how movie plot and genre information can be used effectively for classification and recommendations. We anticipate learning about the impact of neural embeddings and transformer fine-tuning on the accuracy and relevance of plot-based movie recommendations.