Multi-Label Movie Genre Classification from Plot Summaries

Jawwadul Islam
1010038095, jawwad.islam@mail.utoronto.ca
GitHub Repository

Introduction

Movies typically belong to multiple genres simultaneously: *Inception* combines science fiction, thriller, and action. Automatically predicting multiple genres from plot summaries addresses practical needs in streaming platforms and recommendation systems, where accurate categorization directly impacts user experience and content discovery. This project develops a deep learning model for multi-label genre classification using natural language processing. The challenge lies in distinguishing genres like romance versus drama, thriller versus horror, from narrative text like summaries. Traditional approaches using Term Frequency-Inverse Document Frequency (TF-IDF) representations with linear classifiers fail to model contextual dependencies in text [1; 2]. Deep learning methods, particularly Long Short-Term Memory (LSTM) networks, are good at learning distributed representations that encode these relationships [3; 4], making them well-suited for this task.

Illustration

Figure 1 shows the model architecture. Plot summaries are tokenized and converted to integer sequences, then embedded into d-dimensional vector representations. These embeddings are processed through a bidirectional LSTM encoder, and the final hidden states are passed to a classification layer with activation (sigmoid function), producing independent probability scores for each genre. The threshold for activation is set to 0.5, so anything above that will be predicted in the outcome.

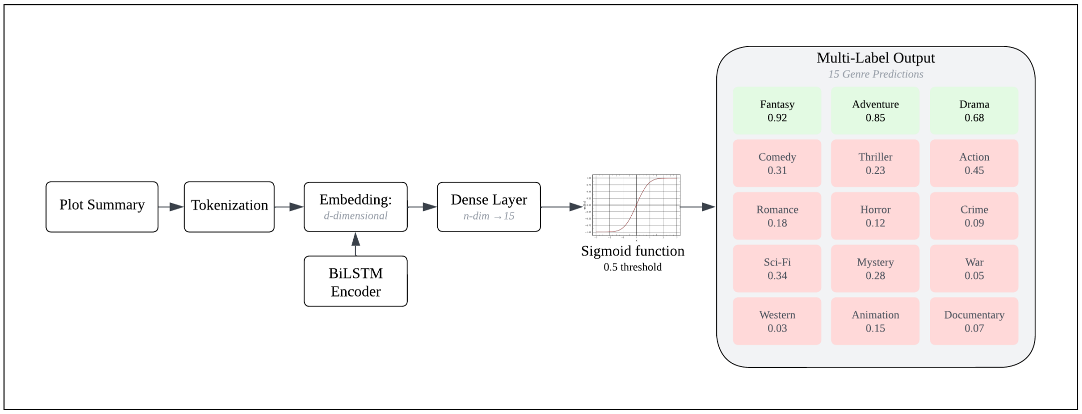

Figure 1: Multi-label genre classification architecture.

Background & Related Work

Multi-label text classification has evolved from bag-of-words features with linear classifiers [1] to neural approaches that learn contextual representations. LSTMs [3] address vanishing gradients in standard Recurrent Neural Networks (RNNs), enabling models to capture long-range dependencies critical for understanding plot summaries. Kar et al. [4] demonstrated that Bidirectional LSTM

(BiLSTM) networks outperform models using simpler emotion-flow features for predicting multi-label movie tags from plot synopses. Hoang [5] similarly applied neural networks to movie synopsis classification, achieving competitive F1-scores. While pre-trained embeddings like GloVe [6] can provide useful initialization, recent sensitivity analyses suggest that tuning embeddings specifically to the target task is essential for optimizing performance on sentence-level classification [2]. This project applies BiLSTM sequence modeling with task-specific embeddings for multi-label movie genre prediction.

Data Processing

Dataset: CMU Movie Summary Corpus available on Kaggle [7], containing 42,306 movie plot summaries with genre annotations. The dataset spans films released from 1893 to 2013 [8].
Filtering: Select only English-language films released between 2000-2013, focusing on the 15 most frequent genres (Drama, Comedy, Thriller, Action, Romance, Horror, Crime, Adventure, Science Fiction, Fantasy, Mystery, War, Western, Animation, Documentary).
Preprocessing steps: (1) Convert to lowercase, remove URLs and special characters; (2) Tokenize using Natural Language Toolkit (NLTK) word tokenizer; (3) Build vocabulary from most frequent tokens; (4) Truncate sequences to a low number of tokens; (5) Encode genres as binary vectors.
Training, Validating, Testing: Train on films from 2000-2009, validate on 2010-2011, test on 2012-2013. All preprocessing will be done in Python.

Architecture

Embedding layer: Maps vocabulary tokens to vectors using PyTorch's `nn.Embedding`, initialized randomly and trained end-to-end with the model. This approach learns task-specific representations optimized for genre classification.
Sequence encoder: Bidirectional LSTM with a number of hidden units per direction and dropout between layers.
Classification head: Fully connected layer maps n-dim vector to 15 logits with activation for independent genre predictions.
Training: Binary cross-entropy loss, Adam optimizer, early stopping on validation macro-F1 score. Target performance: 70-76% macro-F1 based on related work [4].

Baseline Model

TF-IDF + Logistic Regression: Plot summaries are vectorized using Term Frequency-Inverse Document Frequency (TF-IDF) with 5,000 features, capturing word importance while downweighting common terms. For each of the 15 genres, a separate binary logistic regression classifier is trained using a one-vs-rest strategy [1]. This transforms multi-label classification into 15 independent binary problems. This baseline represents standard practice before deep learning and isolates the contribution of neural sequence modeling: it captures which words appear but not their order or context. Expected performance based on similar tasks: 60-65% macro-F1 [2].

Ethical Considerations

Dataset bias: The CMU corpus draws from Wikipedia, reflecting historical biases in film documentation. Older genres (Western, War) may be overrepresented while emerging international genres are underrepresented. Models may learn temporal or cultural associations rather than inherent narrative features.
Subjective labeling: Genre classification is culturally dependent: "Horror" versus "thriller" varies by region. Model predictions give us one perspective, not ground truth.
Potential misuse: Automated genre filtering in recommendation systems could create filters limiting content diversity.
Limitations: The model will struggle with genre-bending films, culturally specific genres, and new genres absent from training data.

References

[1] G. Tsoumakas and I. Katakis. *Multi-Label Classification: An Overview*. International Journal of Data Warehousing and Mining, 3(3):1–13, 2007.

[2] Y. Zhang and B. Wallace. *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. Proceedings of IJCNLP, 2015.

[3] S. Hochreiter and J. Schmidhuber. *Long Short-Term Memory*. Neural Computation, 9(8):1735–1780, 1997.

[4] S. Kar, S. Maharjan, and C. Solorio. *Folksonomication: Predicting Tags for Movies from Plot Synopses using Emotion Flow Encoded Neural Network*. Proceedings of COLING, 2018.

[5] Q. Hoang. *Predicting Movie Genres Based on Plot Summaries*. arXiv preprint arXiv:1801.04813, 2018.

[6] J. Pennington, R. Socher, and C.D. Manning. *GloVe: Global Vectors for Word Representation*. Proceedings of EMNLP, 2014.

[7] M. Safi. *Movies Genre Dataset - CMU Movie Summary*. Kaggle, 2021. Available at: `https://www.kaggle.com/datasets/msafi04/movies-genre-dataset-cmu-movie-summary`

[8] D. Bamman, B. O'Connor, and N.A. Smith. *Learning Latent Personas of Film Characters*. Proceedings of ACL, 2013.