Enhancing Movie Genre Classification with Transformer-Based Models and Conformal Prediction

Matthieu Liger, Kirara Kura, Zixia Huang

March 13, 2025

1 Introduction

1.1 Background and Motivation

Movie genre classification is an essential task in the field of natural language processing (NLP) and machine learning. With the increasing volume of digital media, accurately categorizing movies into their respective genres can aid in content recommendation. Text-based classification using cast, crew, keywords and descriptions has gained significant traction due to advancements in deep learning.

This project explores the use of two BERT-based prediction methods combined with two types of conformal prediction frameworks to enhance the reliability and interpretability of movie genre classification.

1.2 Problem Statement

Movie genre classification presents multiple challenges due to the inherent subjectivity and overlapping nature of genres. A single movie can belong to multiple genres, making it a multi-label classification problem rather than a simple single-label classification task. Furthermore, the complexity of natural language in movie plot summaries introduces ambiguity and variability, which traditional models struggle to handle effectively.

Bert language model is able to do more accurate and context-aware text analysis. By leveraging BERT, genre classification models can capture intricate relationships between words and phrases, leading to more robust predictions. However, a key challenge in classification tasks is ensuring reliable and well-calibrated predictions. To address this, conformal prediction techniques offer a principled way to quantify uncertainty and provide confidence measures along with predictions.

1.3 Related works

Recent advancements in natural language processing and machine learning have significantly improved the performance of text classification tasks, including movie genre classification. Several studies have explored different methodologies to enhance genre prediction accuracy.

One prominent approach involves the use of traditional machine learning techniques such as Support Vector Machines (SVM) and Random Forests, which rely on manually engineered features extracted from textual descriptions [3]. However, these methods often struggle to capture the deeper contextual meanings in movie plot summaries.

With the rise of deep learning, researchers have increasingly turned to neural network-based methods, particularly transformer architectures like BERT. DistilBERT [4], have demonstrated state-of-the-art performance in various text classification tasks.

In addition to model advancements, researchers have also explored the use of conformal prediction techniques to improve classification reliability. Conformal prediction provides a way to quantify uncertainty by generating prediction sets with confidence levels, ensuring that model outputs are statistically valid [5].

This project builds upon these existing works by integrating BERT-based genre classification models with conformal prediction techniques. By comparing two different prediction methods and two types of conformal prediction, we aim to evaluate their effectiveness in improving accuracy, confidence estimation, and overall robustness in movie genre classification.

1.4 Brief Conclusion

we explored two transformer-based approaches for multi-label movie genre classification, combining with two conformal prediction techniques to ensure reliable and well-calibrated predictions.

Method 1 utilized a DistilBERT-based text classifier combined with an alternative metric-based conformal risk control. It achieved higher score, with 0.80507 on the public leaderboard, outperforming baseline approaches by leveraging structured text formatting and a conformal risk control framework.

Method 2 introduced a hybrid DistilBERT model, incorporating both textual and numerical features alongside a ranking-based nonconformity score for conformal prediction.It achieves an evaluation score of 0.67 on the validation set.

1.5 Project Roadmap

The following sections include the dataset's statistical description, two methods implemented for classification, two conformal prediction methods, and the final conclusions drawn from the experiments.

2 Data Description

2.1 Dataset Overview

The dataset contains train and test data, which contains metadata about movies, including attributes such as title, release date, runtime, revenue, vote counts, and textual descriptions (overview). The train dataset also includes binary labels for multiple genres, making it a multi-label classification problem.

2.2 Genre Distribution and Multi-Label Characteristics

The dataset used for training the model consists of movies labeled with multiple genres. The distribution of movies across different genres is presented in Figure 1, highlighting that Drama and Comedy are the most prevalent genres, whereas categories such as TV Movie and Western are comparatively rare.

Since movies can be associated with multiple genres, understanding the distribution of genre counts per movie is essential. As illustrated in Figure 2, most movies belong to one to three genres, while only a small fraction spans six or more genres. This multi-label nature of the dataset presents a challenge for classification, requiring models to effectively capture relationships between overlapping genre labels.

2.3 Genre Correlations

Certain genres tend to co-occur more frequently than others. Figure 3 presents a correlation matrix of genre relationships, indicating which genres are commonly associated with each other.

3 Method 1: DistilBERT Model with Alternative Metric-Based Conformal Risk Control

This solution leverages Hugging Face's `DistilBertModel` with custom text preprocessing, featuring structured template formatting, strategic token separation.For the conformal prediction method, we used an alternative Metric-Based (provided from the competition [2]) conformal risk control. We discuss model selection, parameter choices, and the rationale behind key architectural decisions.

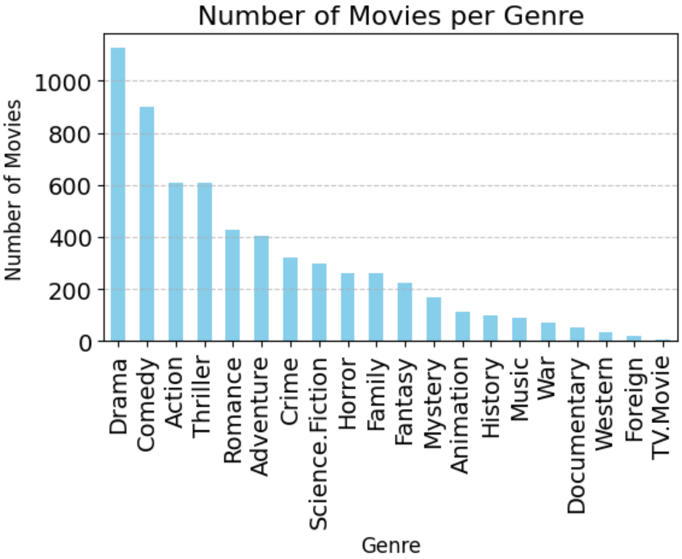

Figure 1: Number of Movies per Genre.

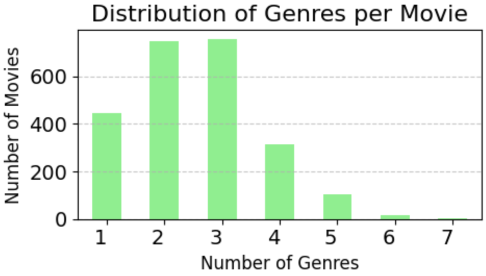

Figure 2: Distribution of Genres per Movie.

3.1 Models and Methodology

The classification model was implemented using DistilBERT as the backbone, leveraging its efficiency advantages—retaining approximately 95% of BERT's performance while using 40% fewer parameters.

The classification model consists of a DistilBERT encoder, followed by a dropout layer and a fully connected layer that outputs logits for each genre label. A sigmoid activation function is applied to convert these logits into probabilities, allowing multi-label classification.

The architecture of the model follows a sequential structure. It begins with a pre-trained DistilBERT encoder, which processes input text and generates contextualized word embeddings. The [CLS] token's embedding is extracted and passed through a dropout layer to mitigate overfitting. The output is then fed into a fully connected linear layer, which maps the hidden representation to genre-specific logits. Finally, a sigmoid activation function transforms these logits into probability scores for each genre.

During training, the model optimizes a binary cross-entropy loss function with class weights to address genre imbalance. The AdamW optimizer is used with weight decay for regularization. A learning rate scheduler with warm-up steps is employed to improve convergence stability. To prevent overfitting, early stopping is implemented based on validation performance. Performance metrics, including precision, recall, and F1-score, are tracked throughout training, and the best model state is saved for inference.

3.2 Text Preprocessing and Tokenization Strategy

The dataset, comprising training, validation, and test sets, undergoes preprocessing to ensure consistency with the evaluation set. Genre labels are extracted, and missing values are replaced with 'no data.' Text inputs are structured by concatenating key attributes using a special separator token ('SEP') to explicitly distinguish different aspects of the movie. Tokenization is performed with the DistilBERT tokenizer for model compatibility, as illustrated in Figure 4.

3.3 Cross-Validation and Ensemble Strategy

To maximize the utility of limited data while improving generalization and reducing overfitting, we employed a 5-fold cross-validation (CV) strategy. In each fold, the dataset was split into 80% training data and 20%

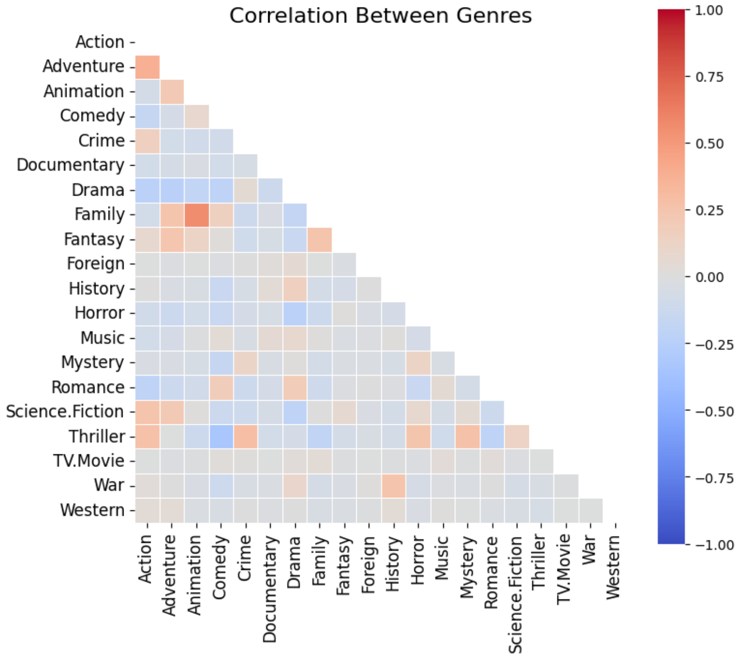

Figure 3: Correlation Between Genres.

Title: L.I.E.[SEP]
Overview: In this biting and disturbing coming-of-age tale ...[SEP]
Tagline: On the Long Island Expressway... [SEP]
Release Date: 2001-01-20[SEP]
Runtime: (no data)[SEP] ...

Figure 4: Example Text

validation data, ensuring that every sample contributed to both training and validation across different iterations. This iterative approach allowed for a more robust estimation of model performance and mitigated bias introduced by a single train-test split.

After training, we constructed an ensemble model using the five trained models. The final prediction was determined through majority voting, where each model contributed equally to the decision. This approach leveraged the diversity among the CV-trained models, reducing variance and improving overall prediction stability compared to relying on a single model.Figure 5 illustrates the overall structure of the ensemble method.

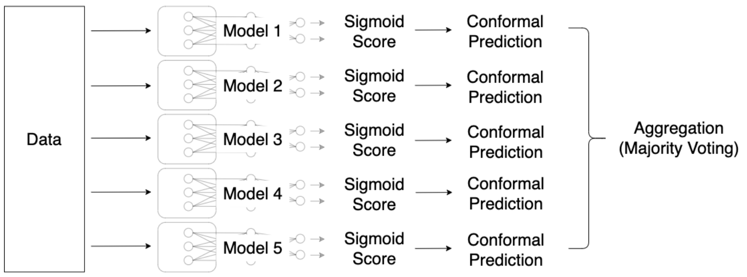

Figure 5: Ensemble strategy with cross-validation and conformal prediction.

3.4 Parameter Selection

The batch size of 16 was chosen as a compromise between memory efficiency and stable gradient estimates. The model was trained for 5 epochs, ensuring sufficient updates while preventing overfitting. The AdamW optimizer was utilized with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e-08$, and a learning rate of $3e-5$. A weight decay of 0.01 was incorporated to introduce L2 regularization, preventing excessive model weight growth. The learning rate was scheduled with a warm-up phase, where the first 10 percent of training steps used an increasing learning rate before decay.

3.5 Prediction with Alternative Metric-Based Conformal Risk Control

In this method, we employ Conformal Risk Control (CRC) as introduced in [1], which enables distribution-free control over expected loss under any chosen loss function. Unlike traditional conformal prediction, which guarantees coverage, CRC allows us to specify and control a broader class of risks, making it more adaptable to different predictive objectives. We experimented with two loss function: one that controls the false negative rate (FNR) as used in [1] and miscoverage rate. The algorithm for miscoverage is shown in Algorithm1, and we chose $\alpha = 0.1$, meaning we allowed miscoverage of up to 10%.

This method ensures that predicted genre sets are sufficiently large to cover ground truth labels while balancing coverage and efficiency. This method led to better coverage (above 90%) while maintaining a reasonable set size. Consequently, we selected this as our final approach.

Algorithm 1 Conformal Risk Control for Miscoverage with Expectation

```
Require: Validation Sigmoid scores $S_{\text{val}}$, Validation labels $Y_{\text{cal}}$ (genre binary label), Test Sigmoid scores $S_{\text{test}}$, Validation sample size $n$
Ensure: Conformal prediction sets $C(X_{\text{test}})$ satisfying non-coverage risk constraints
  Step 1: Initialize parameters
  Compute risk target:
    $R_{\text{target}} = \frac{n+1}{n}\alpha - \frac{1}{n+1}$
  We use $\alpha$=0.1, and the above inflates the risk to ensure we do not underestimate the risk
  Step 2: Compute empirical expectation of risk function
  Define empirical expectation as:
    $E[\ell_{\text{miscoverage}}(C_\lambda(X_{\text{test}}), Y_{\text{test}})] \approx \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{miscoverage}}(C_\lambda(X_i), Y_i)$
  Where:
    $\ell_{\text{miscoverage}}(C_\lambda(X_i), Y_i) = 1 - \mathbb{1}\{Y_i \in C_\lambda(X_i)\}$        ▷ Miscoverage loss
  Step 3: Find the optimal threshold $\lambda_{\text{hat}}$
  Use a root-finding algorithm (Brent's method) to find $\lambda_{\text{hat}}$ such that:
    $E[\ell_{\text{miscoverage}}(C_{\lambda_{\text{hat}}}(X_{\text{test}}), Y_{\text{test}})] \leq \alpha$
  This requires iteratively evaluating the empirical miscoverage risk $R(\lambda)$ as follows:
    (a) Define the prediction set for validation data: $C_{\text{val}}(\lambda) = \{i : S_{\text{val},i} \geq \lambda\}$
    (b) Compute empirical expectation of miscoverage: $E[\ell_{\text{miscoverage}}] = \frac{1}{n}\sum_{i=1}^{n}(1 - \mathbb{1}\{Y_i \in C_{\text{val}}(\lambda)\})$
  Step 4: Construct final prediction sets
  Define the prediction set for validation data as: $C(X_{\text{test}}) = \{j : S_{\text{test},j} \geq \lambda_{\text{hat}}\}$
  Step 5: Ensure no empty prediction sets
  For $C(X_{\text{test}})$ with no genre assigned, assign all genres
  Step 6: Return the final prediction sets
  return $C(X_{\text{test}})$
```

3.6 Results and Evaluation

The models attained an average evaluation metric score of 0.79741 on the validation set and 0.80507 on the public leaderboard, indicating strong performance in the balanced classification task.

Training was conducted on Kaggle's dual NVIDIA T4 GPUs, with each fold taking approximately 16 minutes, resulting in a total cross-validation training time of 80 minutes. As an example, Figure 6 illustrates the overall training dynamics. The validation loss (red line) remains relatively stable, with a slight downward trend during the first two epochs.

In addition, the performance metrics plot in Figure 6 shows steady improvement in precision, recall, and F1 score across epochs. The validation loss versus F1 score scatter plot indicates that as validation loss decreases, F1 score generally improves, with the highest performance achieved in epoch 5. This confirms that the model effectively learns useful patterns in the dataset over time, while early stopping helps prevent overfitting. The step-wise training loss plot further illustrates fluctuations in loss reduction, emphasizing the importance of checkpoint selection for ensuring model stability. This structured approach provides two key improvements.

First it emphases on targeted Feature, the explicit text templating with `sep_token` separation improved accuracy compared to naive concatenation of raw text fields in ablation studies. This structured formatting directs the model's attention to more categorical descriptions while maintaining contextual relationships with other information.

Secondly, the conformal prediction method is more accurate in terms of the evaluation metrics. Comparing to conformal prediction methods that controls the false negative rate (FNR), we directly optimize based on prediction set size and coverage. This method ensures that predicted genre sets are sufficiently large to cover ground truth labels while balancing coverage and efficiency.

Our method's 0.80507 score substantially outperforms baseline approaches. This demonstrates that the `sep_token` strategy and conformal prediction approach provide greater performance benefits than simply using a larger model architecture.

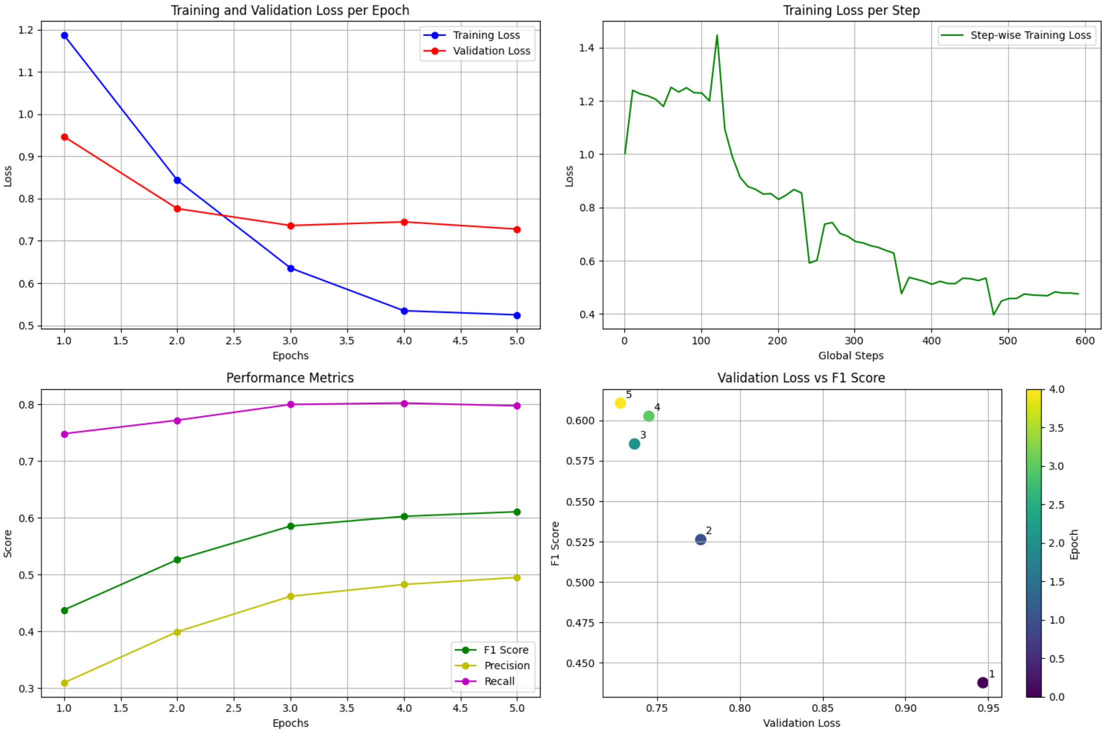

Figure 6: Training and validation loss in fold 1.

4 Method 2: Hybrid DistilBERT Model with Ranking-Based Nonconformity Score

This method presents a transformer-based movie classification system featuring a novel tokenizer retraining approach. Using DistilBERT as the base model, we demonstrate how domain-specific tokenizer adaptation combined with structured text formatting achieves superior performance on the movie classification task. Later, we also apply a large regularization to the model, this implementation highlights custom vocabulary training, parameter optimization strategies, and their impact on model efficacy.

4.1 Models and Methodology

The second approach introduces a Hybrid DistilBERT Genre Classifier, which incorporates both textual and numerical features. Unlike the first model, which relies solely on text embeddings from DistilBERT, this method combines the [CLS] token representation with structured numerical attributes (e.g. movie revenue, vote, budget) to enhance predictive power. The architecture consists of a DistilBERT encoder to extract textual feature representations, followed by a concatenation layer that combines text embeddings with numerical features. The fully connected feed-forward network includes a first hidden layer with 2048 neurons and ReLU activation, a second hidden layer with 128 neurons, and a final output layer with 20 neurons, one for each genre. Dropout layers with a rate of 0.1 are applied at multiple stages for regularization.

This hybrid architecture aims to leverage both structured and unstructured data to improve classification accuracy. The numerical attributes complement textual information, helping the model better distinguish between genres that may share similar descriptions but differ in structured metadata.

We show the visualizations of saliency maps for the movie "2001: A Space Odyssey" in figure 7.

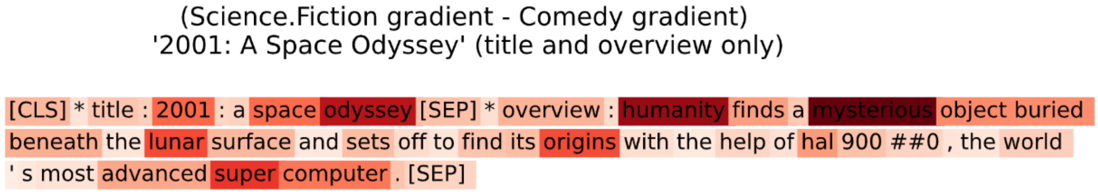

Figure 7: 2001: A Space Odyssey

4.2 Ranking-based Nonconformity Score Conformal Prediction

In this method, we utilize a ranking-based nonconformity score, which assigns higher uncertainty to movies where the model struggles to confidently rank the correct genres. The core idea is to determine the nonconformity score based on the worst-ranked true label.

For a given input x_i, the model produces a probability vector $f(x_i)$, where each entry corresponds to the predicted probability of a particular genre. The steps to compute the nonconformity score are shown in algorithm 2. This can be thought of as a conservative extension of the classic nonconformity score which consists of add probabilities up the one true label in the case of single-class data. The coverage obtained was found to match the expectation of $1-\alpha$ but resulted in wider prediction sets than other methods in this document.

Once the nonconformity scores are computed, we use them to form conformal prediction sets for new test samples. The key idea is to select genres where the cumulative probability surpasses a calibrated threshold.

Algorithm 2 Nonconformity Score Based on the Worst-Ranked True Label

```
Require: A calibration dataset $\{(x_i, y_i)\}_{i=1}^{n}$, with $y_i \in \{0,1\}^K$ with $K = 20$, the number of classes.
Require: A model $f$ returning probabilities $f(x_i) = (f_1(x_i), \ldots, f_K(x_i))$ for each $x_i$.
  for $i = 1 \to n$ do (i.e. each $x_i$ in the calibration set)
      Compute $\big(f_1(x_i), \ldots, f_K(x_i)\big)$
      Sort in descending order:
          $f_{(1)}(x_i) \geq f_{(2)}(x_i) \geq \ldots \geq f_{(K)}(x_i),$
  and let rank$(k)$ be the position of class $k$ in this sorted list (where rank$(k) = 1$ for the largest probability).
      Let worstRank$_i$ be the highest rank (lowest probability) among the true labels:
          $\text{worstRank}_i \ = \ \max\{\text{rank}(k) : y_i[k] = 1\}.$
      Define the nonconformity score
          $s(x_i, y_i) \ = \ \sum_{j=1}^{\text{worstRank}_i} f_{(j)}(x_i).$
  end for
  return $\{s(x_i, y_i)\}_{i=1}^{n}$
```

Given a significance level α, the conformal prediction set for a new instance $x_{n+1}$is constructed as:

$$S(x_{n+1}) = \{k : \sum_{j=1}^{rank(k)} f_{(j)}(x_{n+1}) \leq Q(1-\alpha)\} \tag{1}$$

where where $Q(1-\alpha)$ is the quantile threshold derived from the calibration set.

This approach ensures that the true genres are included in the prediction set with a probability of at least $(1-\alpha)$.

4.3 Parameter Selection

Given the hybrid architecture integrating textual and numerical features, it was essential to optimize both the DistilBERT fine-tuning and the fully connected layers.

Batch size was set to 16, as it provides a balance between training stability and GPU memory limitations. Number of training epochs was set to 5 based on empirical validation results.We employed the AdamW optimizer with hyperparameters $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1e-8$, and a learning rate of 4e-5. A weight decay of 0.01 was applied to introduce L2 regularization, and early stopping was applied with a patience of 2 epochs, reducing the risk of overfitting.

In the conformal prediction stage, we set the risk control threshold λ^* using the 90th percentile quantile of nonconformity scores. This ensures that at least 90% of true genres are captured in the prediction set, maintaining high coverage while keeping set sizes manageable.

4.4 Results and Evaluations

Training was conducted on Kaggle's dual NVIDIA T4 GPUs, with each fold taking approximately the same amount of time as method 1. As an example, Figure 8 illustrates the overall training dynamics. The training loss (solid blue line) shows an overall downward trend with high variance, indicating some noise in gradient updates but consistent improvement over steps. The validation loss (red line) has a downward trend during the first five epochs.

In addition, the performance metrics plot in Figure 8 shows steady improvement in precision, recall, and F1 score across epochs. The validation loss versus F1 score scatter plot indicates that as validation loss decreases, F1 score generally improves, with the highest performance achieved in epoch 5. This confirms that the model effectively learns useful patterns in the dataset over time, while early stopping helps prevent overfitting. The step-wise training loss plot further illustrates fluctuations in loss reduction, emphasizing the importance of checkpoint selection for ensuring model stability. With the "worst-rank" nonconformity score, the performance metric obtained was up to 0.67, while with the element-wise inverted probability, the performance metric was up to 0.79.

There are two main differences between method 1 and method 2, Method 2 leverages the ranking-based nonconformity score, which provides a more structured and informative uncertainty estimate. By considering the worst-ranked true label, this method ensures that movies with uncertain genre assignments receive larger prediction sets, whereas confidently classified movies receive smaller sets. This leads to better trade-offs between coverage and specificity.

The second difference is that the textual feature representations and numerical features. This hybrid architecture leverages both structured and unstructured data to improve classification accuracy. The numerical attributes complement textual information, helping the model better distinguish between genres that may share similar descriptions but differ in numerical features.

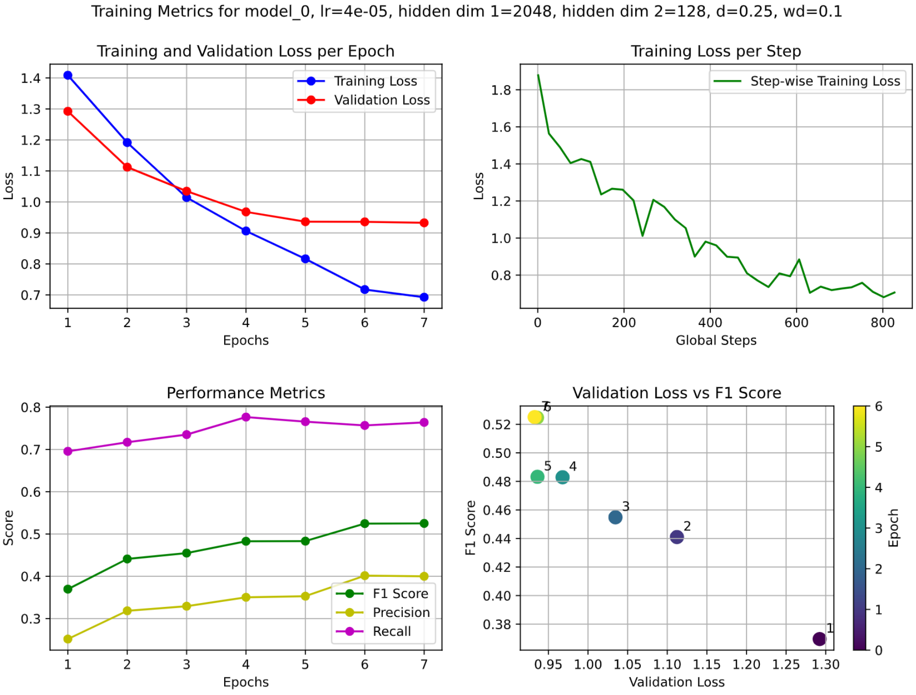

Figure 8: Fold 1

5 Discussions and Limitations

Conformal prediction guarantees marginal but not conditional coverage, meaning the expected coverage across all test points is at least $1-\alpha$, but individual prediction sets may not achieve this level. Consequently, the final prediction sets may not always attain the highest evaluation metric score.

Secondly, as shown in Figure 1, genres are unevenly distributed, leading to lower sigmoid scores for less frequent genres. This imbalance also results in higher false negative rates for these genres, as illustrated in Figure 9. To mitigate this, future improvements could involve creating training datasets with a more balanced genre distribution, potentially reducing disparities in false negative rates across genres. Further future work could apply genre-specific conformal prediction to improve coverage calibration and reduce uncertainty disparities.

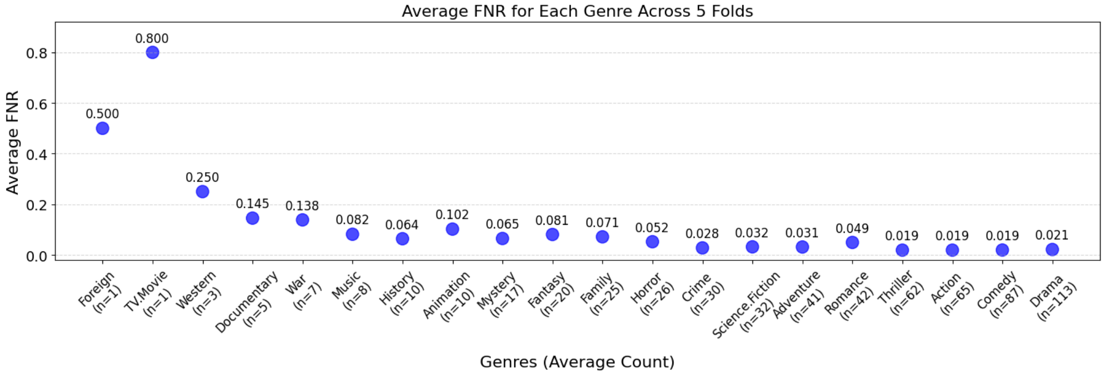

Figure 9: False negative rate (FNR) per genre in method 1

6 Conclusion

In this project, we explored two transformer-based approaches for multi-label movie genre classification, combining with two conformal prediction techniques to ensure reliable and well-calibrated predictions. Method 1 utilized a DistilBERT-based text classifier combined with an alternative metric-based conformal risk control, while Method 2 introduced a hybrid DistilBERT model, incorporating both textual and numerical features alongside a ranking-based nonconformity score for conformal prediction.

The results demonstrated that both methods were effective in genre classification, but each had its strengths and trade-offs. Method 1 achieved higher score, with 0.80507 on the public leaderboard, outperforming baseline approaches by leveraging structured text formatting and a conformal risk control framework. Method 2 achieves a score of 0.67 on the validation set. It provided more structured uncertainty estimates, ensuring that movies with uncertain genre assignments received larger prediction sets while confidently classified movies had smaller sets. The inclusion of numerical features in Method 2 contributed to better genre distinction, particularly for movies that shared similar textual descriptions but differed in numerical attributes.

Several potential improvements could further enhance the performance and reliability of the model. A larger transformer model such as BERT-large or T5 could be explored to improve text representation and genre classification performance.Graph-based learning could be used to model relationships between genres, particularly for movies that belong to multiple highly correlated genres. More advanced conformal prediction methods, such as class-conditional conformal prediction, could refine uncertainty quantification and make prediction sets more adaptive. Expanding the dataset with additional metadata such as user reviews, critic ratings, or cast and crew information could provide valuable contextual signals that further improve classification accuracy.

References

[1] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[2] ArmeenTaeb and Xiaozhu Zhang. Win25 stat 528 kaggle competition 2. `https://kaggle.com/competitions/win-25-stat-528-kaggle-competition-2`, 2025. Kaggle.

[3] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[5] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.