

Probabilistic Fusion Model for Multi-Label Media Content Classification

Javier Carreno

*Department of Computer Science
Royal Holloway, University of London
Surrey TW20 0EX, United Kingdom
Javier.Carreno.2023@live.rhul.ac.uk*

Khuong An Nguyen

*Department of Computer Science
Royal Holloway, University of London
Surrey TW20 0EX, United Kingdom
Khuong.Nguyen@rhul.ac.uk*

Zhiyuan Luo

*Department of Computer Science
Royal Holloway, University of London
Surrey TW20 0EX, United Kingdom
Zhiyuan.Luo@rhul.ac.uk*

Andrew Fish

*Department of Computer Science
University of Liverpool
Liverpool L69 3BX, United Kingdom
Andrew.Fish@liverpool.ac.uk*

Abstract—One of the major challenges in multi-modal media content classification is effectively integrating different modalities (e.g., text and visuals) while simultaneously managing the dependencies between different labels. In this work, we explore several strategies to address this challenge. We found that a combination of textual and visual embeddings from state-of-the-art models (BERT for text and CLIP for image data) was insufficient. We propose a novel probability fusion algorithm that integrates model outputs, leveraging their individual strengths to enhance classification accuracy. Our algorithm achieved an Exact Match Ratio (EMR) of 56.29% for films and 54.23% for TV programmes across 2 datasets.

Index Terms—Multi-modal Learning, Multi-label Classification, Media Content Classification

I. INTRODUCTION

In today’s highly competitive media sector, with many television and entertainment providers competing for audience attention, effective recommendation systems play a key role in engaging and retaining the viewers. An accurate content classification, built upon genre combinations, is a fundamental building block to guide viewing choices and improve engagement.

The problem of media content classification, particularly for films and TV programs, is inherently multi-label, as content can belong to different genres. However, traditional content classification largely depended on single-modality models, using text-based features such as titles, synopses, and metadata. These approaches often struggled with multi-label complexity, imbalanced genre distributions, and the ambiguity of textual descriptions, which led to poor performance [1]. While recent research has explored multi-modal learning, to incorporate images, audio, and video for improving classification [2], [3], many of these methods still suffered from data alignment problems, interactions between different modalities, and computational inefficiency [4].

To address these challenges, we introduce a novel algorithm that integrates outputs from state-of-the-art BERT (for

text classification) and CLIP (for images classification) using probabilistic fusion. Our key contributions are:

- **A multi-modal classification algorithm:** We propose a probabilistic fusion approach that effectively integrates multiple modalities (i.e., texts and images).
- **Two multi-modal datasets:** We curated two media datasets that contain over 60,000 samples of films and TV programmes across different genres, for future research.
- **Empirical performance analysis:** We conducted a comprehensive empirical analysis on these datasets and compared our model’s performance against state-of-the-art single-modal and multi-modal models.

The rest of the paper is organised as follows. Section II reviews related work on multi-label multi-modal classification. Section III presents our fusion approach. Section IV details the datasets, baseline performance, and empirical results. Finally, Section V summarises our contributions and outlines future directions.

II. RELATED WORK

While traditional classification assumes a single label per instance, multi-label classification assigns multiple labels, making the task more difficult. Early methods like binary relevance [5] treated each label independently, ignoring label dependencies. Later approaches, such as classifier chains [6], accounted for inter-label relationships, though. Recent advances, including deep learning-based methods [7], leverage neural networks to model these dependencies automatically. Notwithstanding, challenges such as label imbalance, where certain labels dominate the dataset, and the scalability of algorithms with an increasing number of labels, remain persistent [8]. Addressing these issues requires innovations in architecture and loss functions specifically designed for multi-label complexity.

Multi-modal learning, which combines data modalities (e.g., text, images, audio), has gained significant attention in recent years. Advances in models like CLIP [9] and LLaVA [10]

have shown promising results in understanding relationships across modalities. However, combining modalities presents new challenges, such as aligning data to ensure features match across modalities; managing interactions to capture dependencies and complementary information; and addressing computational complexity, particularly when applied to large-scale or real-time media systems. Previous research shows that basic fusion methods often struggle to capture the subtle dependencies between modalities, stressing the need for more advanced architectures or ensemble-based techniques [11].

Transformer-based multimodal models, such as MMBT [12], LXMERT [13], and FLAVA [14], represent advances in early fusion, jointly encoding image and text inputs via cross-attention or shared spaces. While effective in tasks like visual question answering and captioning, these methods rely on tightly aligned input pairs and require considerable resources. Late fusion has also been explored in earlier work, including algebraic fusion of deep CNNs [15] and cascaded classifiers for multimedia indexing [16]. Our approach extends these ideas by introducing engineered probabilistic features that capture modality disagreement and inter-genre dependencies.

Automatic genre classification has long been a central task in media analysis. Early systems relied heavily on text-based features such as synopses and metadata, but more recent work has adopted multimodal frameworks that incorporate text, images, audio, and even full video analysis [2], [3]. Despite these advances, key challenges persist—particularly genre ambiguity, overlapping category boundaries, and the need to model dependencies among co-occurring genres. Although multimodal benchmarks exist [17]–[19], few target multi-label media genre classification at recommendation scale. To address this, we curated two new datasets of films and TV programmes.

III. MULTI-MODAL MULTI-LABEL MEDIA CONTENT CLASSIFICATION

Without loss of generality, let C be a media item (e.g., film or TV show) with a feature set X comprising multimodal information such as text (e.g., synopsis, title, metadata), images (e.g., poster), and potentially audio (e.g., soundtrack) or video (e.g., trailer). Given a genre set G , our objective is to predict the binary vector $y = (y_1, y_2, \dots, y_{|G|})$, where each $y_i \in \{0, 1\}$ for $i = 1, 2, \dots, |G|$. From a training set of pairs (X, y) , we aim to learn a function $f : X \rightarrow y$ to predict the genre labels of C .

To address this challenge, we propose a general multi-modal classification framework (see Figure 1) designed to integrate predictions and probabilities from multiple modality-specific models. Rather than relying on a unified multi-modal architecture, our approach processes each modality independently and fuses their outputs at the prediction level using a probabilistic stacking mechanism. This modular design enables flexibility in the choice of models per modality and promotes robust decision-making by combining complementary views of the data.

In this study, we instantiate the framework using:

- **BERT (text modality):** We use BERT (Bidirectional Encoder Representations from Transformers) to process textual media data, including title, synopsis, director, and cast. As a state-of-the-art NLP model, BERT effectively captures semantic and contextual nuances in text [20].
- **CLIP (image modality):** We employ CLIP (Contrastive Language–Image Pre-Training) to process visual media, specifically posters. CLIP learns joint representations of text and images, embedding them in a shared space to link visual features to textual descriptions [9].

As each modality offers complementary information, we hypothesise that our fusion strategy would provide a more comprehensive understanding of the media content. Conceptually, our method captures the relationships between genres, rather than treating them as separate tasks (*flat classification*) [21], using a multi-label classification framework that accounts for dependencies. We apply a probability fusion method integrating newly engineered features, including *model agreement metrics*, *probability-based measures*, and *inter-genre relationships*, enabling each model to contribute its strengths while ensuring consistency across modalities.

- **Agreement-Based Features.**
 - **Prediction Count (PC):** This metric quantifies the total number of positive genre predictions made across all modalities for a given sample. A higher PC indicates a more permissive (potentially overconfident) prediction strategy, while a lower PC reflects a more conservative and selective approach. It is computed as:

$$PC = \sum_{g \in G} \sum_{m \in M} \hat{y}_{g,m} \quad (1)$$

where:

G is the set of genres, M is the set of modalities, $\hat{y}_{g,m} \in \{0, 1\}$ is the binary prediction of genre g by modality m for the current sample.

- **Full Agreement (FA):** This metric counts how many genres have unanimous agreement across all modalities — meaning all modalities predicted the same value (either all 1s or all 0s) for that genre. It measures the consistency of predictions across modalities. It is defined as:

$$FA = \sum_{g \in G} \mathbb{I}(\hat{y}_{g,m_1} = \hat{y}_{g,m_2} = \dots = \hat{y}_{g,m_{|M|}}) \quad (2)$$

where:

$\mathbb{I}(\cdot)$ is the indicator function, returning 1 if all predictions for genre g are equal, and 0 otherwise.

- **Probability-Based Features.**
 - **Max Probability Difference (MPD):** It measures the disagreement between modalities by looking at how spread out the predicted probabilities are for a specific genre. A high MPD means the modalities strongly disagree about the likelihood of that genre. It is defined as:

$$MPD_g = \max_{m \in M} P_{g,m} - \min_{m \in M} P_{g,m} \quad (3)$$

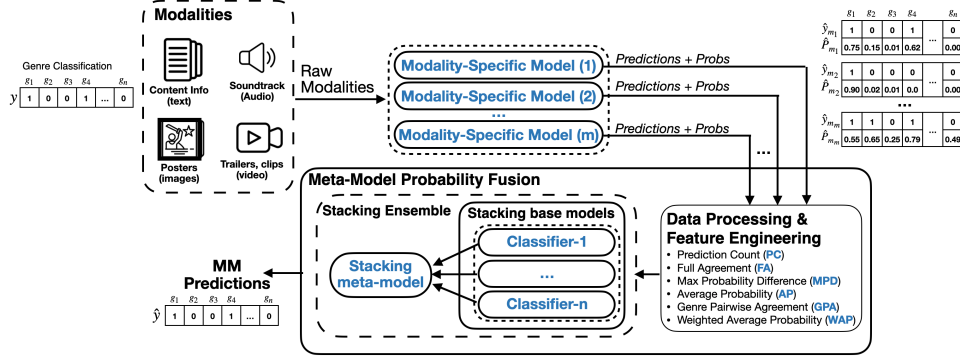


Fig. 1. Overview of our multi-modal classification framework.

where:

$g \in G$ is a specific genre,

$P_{g,m} \in [0, 1]$ predicted probability that genre g is present, according to modality m .

- **Average Probability (AP):** The mean predicted probability across modalities for a given genre, reflecting overall confidence even when individual predictions vary. It is defined as:

$$AP_g = \frac{1}{|M|} \sum_{m \in M} P_{g,m} \quad (4)$$

where:

$|M|$ is the number of modalities.

- **Inter-Genre Agreement Features.**

- **Genre Pairwise Agreement (GPA):** Measures whether all modalities make identical binary predictions for both genres in a given pair, capturing cross-modal consistency in interpreting genre combinations. It is defined as:

$$GPA_{g_i, g_j} = \mathbb{I} \left(\hat{y}_{g_i, m_1} = \hat{y}_{g_i, m_2} = \dots = \hat{y}_{g_i, m_{|M|}} \right) \cdot \mathbb{I} \left(\hat{y}_{g_j, m_1} = \hat{y}_{g_j, m_2} = \dots = \hat{y}_{g_j, m_{|M|}} \right) \quad (5)$$

where: g_i and g_j are two different genres.

In addition to feature engineering, modality-specific EMR (Exact Match Ratio) scores are computed on the training set and normalised to yield weights $\hat{w}_m$. These weights are then used to compute the *Weighted Average Probability (WAP)*, a genre-level feature that combines predicted probabilities across modalities, depending on how well each modality performs during training.

$$WAP_g = \sum_{m \in M} \hat{w}_m \cdot P_{g,m} \quad (6)$$

where:

$\hat{w}_m \in [0, 1]$ is the normalised EMR-based weight for modality m , computed using ground truth labels y .

Using these features, we train a multi-output classifier to predict the genre labels (see Algorithm 1). This fusion framework combines outputs from multiple models, capturing interdependencies between genres and enhancing flexibility. By understanding the relationships between the features and target labels, the model can generalise better, resulting in more accurate predictions.

Algorithm 1 Multimodal Probabilistic Fusion for Multi-Label Classification

Input: Training set $y, \{\hat{y}_m, \hat{P}_m\}_{m \in M}$

Output: M_{meta} trained multi-output classifier

Steps:

- 1) For each modality $m \in M$, obtain out-of-fold predictions $\hat{y}_m^{oof}$ and probabilities $\hat{P}_m^{oof}$
 - 2) Compute modality performance based on EMR (7), and derive normalised weights $\hat{w}_m$
 - 3) Compute engineered features using out-of-fold predictions:
 - Prediction Count (PC) (1)
 - Full Agreement (FA) (2)
 - Max Probability Difference (MPD) (3)
 - Average Probability (AP) (4)
 - Genre Pairwise Agreement (GPA) (5)
 - Weighted Average Probability (WAP) (6)
 - 4) Train stacking base models on combined modality outputs and engineered features
 - 5) Train final meta-classifier M_{meta} using outputs from stacking base models
-

IV. EMPIRICAL RESULTS

This section details the performance results of our model on two datasets and compare its performance to baseline state-of-the-art models.

A. Datasets

To evaluate our multi-modal model, we curated 2 datasets in films and TV programmes (from *FilmAffinity* and *IMDB*), as no pre-existing datasets met our multi-modal needs. Initially, genre combinations were kept in the order assigned by the source, where a genre's position indicated its primary classification based on their prevalence. To ensure consistency, we reordered them alphabetically, treating variations like 'Drama,

Comedy’ and *‘Comedy, Drama’* as equivalent. Genre combinations appearing only once were excluded for *Films*, and those with fewer than 10 instances were removed for *TV Programmes* to reduce noise and enhance model robustness.

The final *Films* dataset comprises 56,637 instances across 88 genre combinations derived from 8 distinct genres: ‘*Action*’, ‘*Animation*’, ‘*Comedy*’, ‘*Drama*’, ‘*Horror*’, ‘*Kids*’, ‘*Romance*’, and ‘*Thriller*’. Similarly, the *TV Programmes* dataset consists of 9,110 instances across 94 genre combinations from 20 genres: ‘*Action*’, ‘*Adventure*’, ‘*Animation*’, ‘*Biography*’, ‘*Comedy*’, ‘*Crime*’, ‘*Documentary*’, ‘*Drama*’, ‘*Family*’, ‘*Fantasy*’, ‘*Game-Show*’, ‘*History*’, ‘*Music*’, ‘*Mystery*’, ‘*News*’, ‘*Reality-TV*’, ‘*Romance*’, ‘*Sci-Fi*’, ‘*Sport*’, and ‘*Talk-Show*’ (Table I).

TABLE I
SUMMARY OF THE FILM AND TV PROGRAM DATASETS

Feature	Films	TV
Instances	56,637	9,110
Genres	8	20
Genre Combinations	88	94
Source	filmaffinity.com	imdb.com

Example entries from the Film dataset are shown in Table II. Table III provides a summary of the key features for each dataset entry. Further details and access to the datasets can be found at <https://github.com/carrenyo/media-multi-label>.

B. Evaluation metrics

Standard metrics treat each label independently, which can misrepresent performance in the presence of label imbalance and inter-label dependencies [1], [22]. This is especially problematic in media content, where genres often co-occur in skewed and dependent ways. For example, a model might predict frequent genres like “Drama” while missing more relevant ones, yet still appear to perform well. We propose alternative evaluation metrics:

- **Exact Match Ratio (EMR):** The fraction of samples where predicted labels exactly match true labels, also known as Exact Match Accuracy.

$$\text{EMR} = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \mathbb{I}(y_i = \hat{y}_i) \quad (7)$$

where: $\mathbb{I}(y_i = \hat{y}_i)$ returns **1** if the predicted label set $\hat{y}_i$ exactly matches the true set y_i for sample i , and **0** otherwise.

- **Hamming Loss.** Measures the fraction of incorrect labels in multi-label classification, with lower values indicating better performance.

$$\text{Hamming Loss} = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \sum_{j=1}^L \mathbb{I}(y_{ij} \neq \hat{y}_{ij}) \quad (8)$$

where n_{samples} is the number of samples, L is the total number of labels, y_{ij} and $\hat{y}_{ij}$ represent the j -th true and predicted labels for sample i , and $\mathbb{I}(y_{ij} \neq \hat{y}_{ij})$ is

an indicator function which returns **1** if the true and predicted labels differ, otherwise **0**.

- **AUC-ROC for Multi-Label Classification.** Measures the model’s ability to distinguish between classes, providing an overall evaluation of ranking performance. Higher values indicate better performance

$$\text{AUC-ROC} = \frac{1}{L} \sum_{j=1}^L \text{AUC}(\text{ROC}_j) \quad (9)$$

where L is the total number of labels, and $\text{AUC}(\text{ROC}_j)$ is the AUC for the ROC curve of label j .

C. Empirical performance

In addition to evaluating our model’s performance against BERT and CLIP, we include BERT-CLIP, a third model formed by concatenating their embeddings.

- **Text Feature Extraction and Classification (BERT):** The *bert-base-uncased* model extracted 768-dimensional embeddings from combined text attributes (*title*, *synopsis*, *director*, *cast*). Text was tokenised and padded using the *BERT tokenizer*. A custom classifier with a linear output layer generated logits. Fine-tuning employed *BCEWithLogitsLoss* and *AdamW* (lr: 0.001) over 20 epochs (batch size: 64) with early stopping on validation loss.
- **Image Feature Extraction and Classification (CLIP):** The *openai/clip-vit-base-patch32* model extracted 512-dimensional features from resized film posters (224x224 pixels), preprocessed with *CLIPProcessor* for consistency and normalisation. A custom feed-forward neural network performed multi-label classification with a linear output layer for logits. The model was trained using *BCEWithLogitsLoss* and *Adam* (lr: 0.001) for 20 epochs with a batch size of 64.
- **Text and Image Feature Extraction and Classification (BERT-CLIP):** This model combined 768-dimensional text features from BERT and 512-dimensional image features from CLIP. The features were standardised and concatenated into a single vector, which was passed to a custom feed-forward neural network. The model was fine-tuned with *BCEWithLogitsLoss* and *Adam* optimizer (lr: 0.001) over 20 epochs with a batch size of 64.

Our meta-model (MM) with probabilistic fusion was built using a *MultiOutputClassifier* with *Random Forest*, *LightGBM*, and *Extra Trees* trained on engineered features. A *Logistic Regression meta-model* was trained using out-of-fold predictions from the base models. Training followed a *5-fold Cross-Validation* approach to ensure robustness, preventing data leakage. The final MM predictions were obtained by aggregating test set predictions across folds.

All experiments were conducted using Python 3.10.13. Model implementation and training were performed with PyTorch (*torch*, *torch.nn*, *torch.optim*), and pre-trained transformer models were accessed via the Hugging Face Transformers library (*BertModel*, *CLIPModel*, *BertTokenizer*, *CLIPProcessor*). Image preprocessing used the *PIL* library, and

TABLE II
EXAMPLE ENTRIES FROM THE FILM DATASET

Title	Director	Cast	Synopsis	Genre(s)
Where the Crawdads Sing	Olivia Newman	Daisy Edgar-Jones, Taylor John Smith, Harris Dickinson	A woman who raised herself in the marshes of the Deep South becomes a suspect in the murder of a man with whom she was once involved.	Drama, Romance
Beyond the Limits	Olaf Ittenbach	Xenia Seeberg, Darren Shahlavi	A rookie journalist seeks answers from a mortician, unveiling a tale of violence and evil.	Horror, Thriller
Girlfriend, Girlfriend	Satoshi Kuwahara		A honest young man is compelled to accept every girl who asks to be romantically involved with him.	Animation, Comedy, Romance
Chaos Theory	Marcos Siega	Ryan Reynolds, Emily Mortimer	Frank Allen, a time-obsessed efficiency expert, discovers the unpredictability of life.	Comedy, Drama
L'affaire SK1	Frédéric Tellier	Raphaël Personnaz, Nathalie Baye	Young inspector unravels a decade-long serial murder case through perseverance and setbacks.	Thriller

TABLE III
SUMMARY OF THE FILM AND TV PROGRAMME DATASET FEATURES

Feature	Description
<i>title</i>	Title of the film or TV programme.
<i>director</i>	Name of the director(s).
<i>cast</i>	Main cast members featured in the media content.
<i>synopsis</i>	A brief summary or description of the content’s plot.
<i>genre</i>	Categories or genres of the content (e.g., <i>Horror</i>).

BERT uses: *title*, *director*, *cast*, and *synopsis*. CLIP uses: poster images.

numerical computations were supported by *numpy*. For ensemble learning and meta-model construction, we used *scikit-learn*’s *RandomForestClassifier*, *ExtraTreesClassifier*, *Multi-OutputClassifier*, and *LogisticRegression*, along with *LGBM-Classifer* from the *lightgbm* library. Data was split 80–20 into training and test sets, and a fixed random state (42) was used throughout to ensure consistency across runs.

As shown in Table IV, our MM outperformed the baseline models on all metrics except for the AUC-ROC score for TV, which was slightly lower than that of BERT-CLIP and BERT.

Although BERT and CLIP performed well individually, BERT-CLIP failed to fully leverage their combined strengths, resulting in suboptimal performance. On the TV dataset, EMR dropped 3.29% compared to BERT, while in the film dataset, it improved by 5.26%. This highlighted the challenges of modality fusion [23], which our MM handled more effectively through probability fusion.

Genre performance varied by *model*, *data distribution*, and *multimodal learning*. The film dataset (56,637 instances, 8 genres) had a somewhat balanced but still skewed distribution towards genres like ‘*Drama*’ (3,468) and ‘*Comedy*’ (1,627). MM delivered the most consistent results, particularly in genres where text and image together improve classification (see Figure 2).

‘*Drama*’ is a clear example of how posters influence classification. While BERT-CLIP, MM, and CLIP achieve 83–87% accuracy, BERT struggles with only 60%, relying solely on text, which results in misclassifications caused by misleading

keywords and limited context. Table V highlights these misclassifications in the baseline models, with the corresponding posters shown in Figure 3.

The TV dataset (9,110 instances, 20 genres) includes 94 genre combinations, with only 31 exceeding 10 instances, posing a challenge (Figure 4). While MM outperforms the others, its performance is highly variable, alternating strong results for common combinations and significant drops for rare or complex ones.

BERT outperforms BERT-CLIP in the TV dataset, with CLIP’s EMR dropping by 30.1%. This may be due to TV posters often featuring simple snapshots (see Figure 5), unlike the detailed artwork in films. Documentaries, with more distinctive visuals, allow CLIP to outperform other models, while the simplicity of other genres hinders genre differentiation for baseline models.

V. CONCLUSION AND FUTURE WORK

In today’s highly competitive media arena, effective recommendation systems are essential for engaging and retaining users. Since the decision to watch or skip content is deeply personal and hinges on classification, systems should go beyond conventional genre labels, integrating user behaviour and strategic categorisation to enhance content recommendations.

However, genre classification is inherently subjective, with overlaps that complicate strict categorisation. The same content is often labelled with different genres across platforms, reflecting variations in perception. Thus, in this paper, we introduced a novel multi-modal classification algorithm based on probabilistic fusion. Our meta-model showed good performance across two datasets. Our model demonstrated the potential of multi-modal, multi-label classification to uncover audience insights. A detailed analysis of viewing patterns revealed key viewer preferences, engagement triggers, and behavioral trends, offering useful intelligence for content providers [24].

Future work will explore personalisation, broader datasets, and additional modalities to enhance adaptability, as well as label powerset and ranking-based methods to capture inter-genre

TABLE IV
PERFORMANCE COMPARISON OF CLIP, BERT, BERT-CLIP, AND META-MODEL

Metric	Film Dataset				TV Dataset			
	CLIP	BERT	BERT-CLIP	MM	CLIP	BERT	BERT-CLIP	MM
EMR	0.3977	0.4597	0.5123	0.5629	0.2783	0.5060	0.4731	0.5423
Hamming Loss	0.1158	0.1033	0.0902	0.0785	0.0577	0.0351	0.0389	0.0325
AUC-ROC	0.8941	0.9137	0.9250	0.9357	0.7847	0.8979	0.9109	0.8808

Random state = 42; 80-20 train-test split; OOF with 5-fold cross-validation.

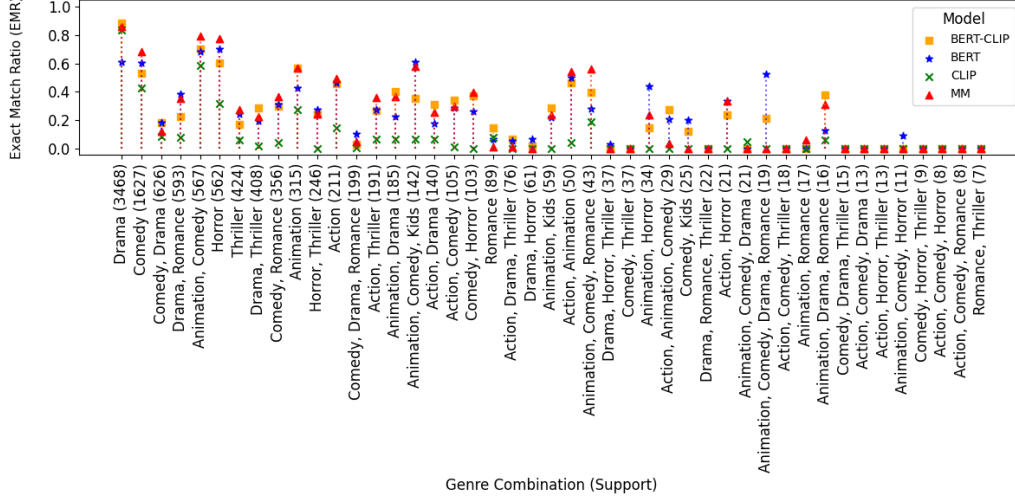


Fig. 2. EMR and Support for the Most Frequent Genres (Film Dataset).

TABLE V
MISCLASSIFICATION OF *Drama* FILMS IN THE BASELINE MODELS

Film	CLIP	BERT	BERT-CLIP	MM
All Good	Drama	Comedy, Drama	Drama	Drama
Donnie Brasco	Drama	Action, Comedy	Action, Drama, Thriller	Drama
Schindler's List	Drama	Comedy, Drama	Drama	Drama
Anne of Green Gables	Drama	Animation, Kids	Drama	Drama
A Mother's Son	Drama	Drama, Thriller	Drama, Thriller	Drama
Night Train	Drama	Thriller	Drama	Drama
Henry IV	None	Comedy, Drama	None	Drama



Fig. 3. Example posters of misclassified drama films, caused by challenging keywords and limited context, as shown in Table V.

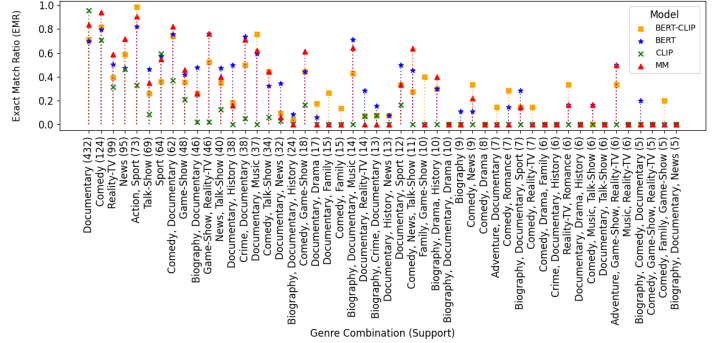


Fig. 4. EMR and support for the most frequent genres (TV Dataset).



Fig. 5. Examples of TV program posters from 'Talk-Show', and 'Game-Show' to 'Reality-TV' and 'News' - illustrating their similarity across genres.

relationships. We also plan to benchmark our method against multimodal transformers (e.g., MMBT, LXMERT, FLAVA) to highlight its advantages in modularity and efficiency.

REFERENCES

- [1] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.
- [2] A. Yadav and D. K. Vishwakarma, "A unified framework of deep networks for genre classification using movie trailer," *Applied Soft Computing*, vol. 96, p. 106624, 2020.
- [3] Z. Zhang, Y. Gu, B. A. Plummer, X. Miao, J. Liu, and H. Wang, "Movie genre classification by language augmentation and shot sampling," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7275–7285.
- [4] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and trends in multimodal machine learning: Principles, challenges, and open questions," *arXiv preprint arXiv:2209.03430*, 2022.
- [5] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, pp. 64–74, 2008.
- [6] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, pp. 333–359, 2011.
- [7] A. N. Tarekegn, M. Ullah, and F. A. Cheikh, "Deep learning for multi-label learning: A comprehensive survey," *arXiv preprint arXiv:2401.16549*, 2024.
- [8] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognition*, vol. 118, p. 107965, 2021.
- [9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [10] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.
- [11] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [12] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.
- [13] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.
- [14] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 638–15 650.
- [15] K. Aderghal, J. Benois-Pineau, K. Afdel, and C. Gwenaëlle, "Fuseme: Classification of smri images by fusion of deep cnns in 2d+ ϵ projections," in *Proceedings of the 15th international workshop on content-based multimedia indexing*, 2017, pp. 1–7.
- [16] B. Safadi and G. Quénot, "Active learning with multiple classifiers for multimedia indexing," *Multimedia Tools and Applications*, vol. 60, no. 2, pp. 403–417, 2012.
- [17] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.
- [18] R. B. Mangolin, R. M. Pereira, A. S. Britto Jr, C. N. Silla Jr, V. D. Feltrim, D. Bertolini, and Y. M. Costa, "A multimodal approach for multi-label movie genre classification," *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 19 071–19 096, 2022.
- [19] P. Cascante-Bonilla, K. Sitaraman, M. Luo, and V. Ordonez, "Moviescope: Large-scale analysis of movies using multiple modalities," *arXiv preprint arXiv:1908.03180*, 2019.
- [20] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [21] A. Nichol, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.
- [22] S. Nazmi, X. Yan, A. Homaifar, and E. Doucette, "Evolving multi-label classification rules by exploiting high-order label correlations," *Neurocomputing*, vol. 417, pp. 176–186, 2020.
- [23] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.
- [24] J. Carreno, K. An Nguyen, Z. Luo, and A. Fish, "Unlocking viewer insights in linear television: A machine learning approach," in *International Conference on Business Informatics Research*. Springer, 2024, pp. 53–67.