Movie Genre Prediction through Multi-label Text Classification

By

Yaswanth Reddy, Nalamalapu

Abstract

This project addresses the challenge of automatically predicting movie genres based on plot descriptions using natural language processing and multi-label text classification techniques. By analyzing a dataset of 34,886 movies, the research demonstrates the efficacy of machine learning approaches in categorizing films across multiple genres simultaneously. Movie plots were preprocessed using tokenization, lemmatization, and POS tagging to create meaningful feature representations. A TF-IDF vectorizer was implemented to transform text data, and three multi-label classification models were compared: One-vs-Rest Logistic Regression, Multi-Output Classification, and Classifier Chain. The final model achieved a Hamming Loss of 0.1004, Exact Match Accuracy of 0.3514, and Jaccard Score of 0.4041. Analysis revealed distinctive linguistic patterns associated with different genres and identified common genre co-occurrences such as comedy-drama, comedy-romance, and crime-drama. This system provides a valuable tool for automatic content categorization in film databases, streaming platforms, and recommendation systems.

1. Introduction

1.1 Overview of the Project

The ability to correctly identify movie genres based on plot descriptions has significant applications in both content organization and recommendation systems. This project addresses the challenge of automating genre prediction using natural language processing (NLP) and multi-label classification techniques. The solution leverages machine learning algorithms to analyze and categorize movie plots across multiple genres simultaneously, facilitating more efficient content organization and improved user recommendations.

Movie genre classification presents unique challenges due to its inherently multi-label nature - films often span multiple genres simultaneously. Additionally, the textual information in plot descriptions contains subtle linguistic patterns that distinguish different genres. This project explores these patterns and evaluates various classification approaches to develop an effective, automated genre prediction system.

1.2 Motivation

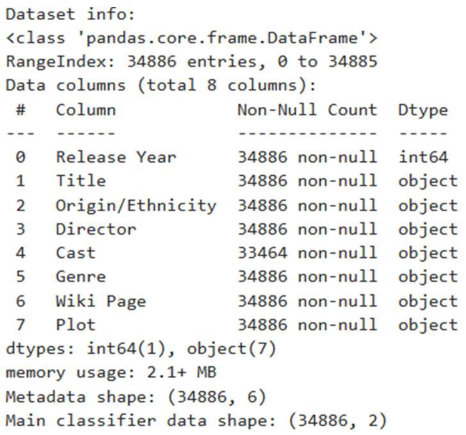
Accurate genre classification is a critical task across multiple domains in the film industry. Content providers require reliable genre tagging to organize vast libraries of movies, while recommendation systems depend on genre information to suggest relevant content to users. Manual genre classification is both time-consuming and subject to inconsistencies, creating a clear need for automated solutions.

Existing genre classification systems often struggle with multi-label classification, treating each genre independently without considering their relationships. This project addresses these limitations by implementing multi-label classification techniques that recognize the interdependence of genres. By using the textual data from plot descriptions, the system eliminates the need for expensive feature extraction from visual or audio content, making it more accessible and cost-effective for implementation across various platforms.

1.3 Dataset

This project utilizes a comprehensive movie dataset containing 34,886 entries. Each entry includes:

- Plot description (the primary feature for genre prediction)
- Genre labels (the target for classification)
- Additional metadata including release year, title, director, cast, and origin



```
Dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34886 entries, 0 to 34885
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Release Year      34886 non-null  int64
 1   Title             34886 non-null  object
 2   Origin/Ethnicity  34886 non-null  object
 3   Director          34886 non-null  object
 4   Cast              33464 non-null  object
 5   Genre             34886 non-null  object
 6   Wiki Page         34886 non-null  object
 7   Plot              34886 non-null  object
dtypes: int64(1), object(7)
memory usage: 2.1+ MB
Metadata shape: (34886, 6)
Main classifier data shape: (34886, 2)
```

The dataset presents several challenges: genre labels are inconsistent in format and granularity, with 6,083 movies labeled as "unknown" genre. The plots vary significantly in length, detail, and writing style. The genre distribution is imbalanced, with drama (7,913 instances) being the most common, followed by unknown (6,083) and comedy (5,799), while many genres appear less than 10 times.

After preprocessing and standardization, the final dataset contained 25,381 movies categorized across 9 major genres:

drama, comedy, action, romance, thriller, crime, horror, fantasy, and family.

This provides sufficient data for training while maintaining a manageable number of target classes.

2. Background

Genre classification systems have evolved significantly with advances in natural language processing and machine learning. Traditional approaches relied on keyword matching and rule-based systems, which struggled with the nuanced language of plot descriptions and the overlapping nature of film genres.

Current state-of-the-art solutions for text-based genre classification utilize various neural network architectures, including RNNs, LSTMs, and transformer models like BERT. These approaches have achieved impressive results but often require substantial computational resources and large labeled datasets. Alternative approaches using TF-IDF features with classical machine learning algorithms provide more lightweight solutions that can be effective for specific applications.

Multi-label classification presents unique challenges compared to traditional multi-class problems. While multi-class classification assumes each instance belongs to exactly one class, multi-label problems allow instances to belong to multiple classes simultaneously. This requires specialized algorithms that can model label dependencies and handle class imbalance effectively.

This project aims to address these challenges by implementing and comparing several multi-label classification techniques using TF-IDF features extracted from movie plot descriptions. The focus is on developing a system that balances accuracy with computational efficiency, making it practical for real-world applications.

3. Approach

3.1 Data Preprocessing

The data preprocessing pipeline consisted of several key steps designed to transform raw movie plot descriptions into structured features suitable for machine learning:

1. **Text Cleaning:** HTML tags, special characters, and URLs were removed from plot descriptions using regular expressions. All text was converted to lowercase to standardize the input.

2. **Tokenization**: The NLTK library was used to split plot descriptions into individual tokens, providing the foundation for more sophisticated NLP techniques.

3. **Stopword Removal:** Common English stopwords (e.g., "the", "and", "is") were removed as they typically don't contribute meaningful information for genre classification.

4. **Part-of-Speech Tagging:** Each token was tagged with its grammatical part of speech (noun, verb, adjective, adverb) to provide additional context for lemmatization and feature engineering.

5. **Lemmatization:** Words were reduced to their base form using NLTK's WordNetLemmatizer with POS information to improve accuracy (e.g., "running" → "run").

6. **Short Word Filtering:** Words with fewer than three characters were removed to further reduce noise.

The preprocessing function successfully transformed plot descriptions into clean token lists, as demonstrated by this example:

Original plot: "A bartender is working at a saloon, serving drinks to customers. After he fills a stereotypically Irish man's bucket with beer, Carrie Nation and her followers burst inside and assault the Irish man..."

Processed tokens: ['bartender', 'work', 'saloon', 'serve', 'drink', 'customer', 'fill', 'stereotypically', 'irish', 'man', 'bucket', 'beer', 'carrie', 'nation', 'follower', 'burst', 'inside', 'assault', 'irish', 'man']

3.2 Genre Analysis and Standardization

Initial analysis revealed over 700 unique genre labels in the dataset, many appearing only a few times. To create a more manageable and meaningful set of genres, several standardization steps were implemented:

1. **Frequency Analysis:** Genre distribution was analyzed, revealing drama (7,913), unknown (6,083), and comedy (5,799) as the most common labels.

2. **Threshold Filtering:** Genres appearing fewer than 100 times were filtered out, reducing the set to 31 major genres.

3. **Hierarchical Mapping:** A two-level mapping approach was implemented to further consolidate similar genres:

- First level: Mapped specialized genres to their parent categories (e.g., "sci-fi" → "science fiction")
- Second level: Further consolidated into final categories (e.g., "science fiction" → "fantasy")

4. **Multi-label Handling:** Movies with multiple genres were preserved as multi-label instances.

The final genre distribution consisted of 9 major categories: drama (9,716), comedy (7,612), action (3,821), romance (2,412), thriller (2,374), crime (1,464), horror (1,450), fantasy (1,328), and family (1,089).

3.3 Feature Engineering

Several feature engineering techniques were applied to create a rich representation of the plot descriptions:

1. **TF-IDF Vectorization:** A TF-IDF vectorizer was implemented with the following parameters:

- Maximum 10,000 features
- Minimum document frequency of 5
- Maximum document frequency of 0.7
- N-gram range of (1, 2) to capture both individual words and common phrases
- Sublinear TF scaling to dampen the effect of high-frequency terms

2. **POS Distribution Features:** Four additional features were created for each plot:

- Percentage of nouns (average: 64.6%)
- Percentage of verbs (average: 16.7%)
- Percentage of adjectives (average: 10.4%)
- Percentage of adverbs (average: 2.1%)

These features provide additional context about the writing style and structure of the plot descriptions that might correlate with specific genres.

3.4 Model Selection and Training

Three multi-label classification approaches were implemented and compared:

1. **One-vs-Rest (OvR) Classifier:**

- Base classifier: Logistic Regression with C=1.0 and liblinear solver
- Trains independent binary classifiers for each genre
- Simple approach that doesn't model genre dependencies

2. **Multi-Output Classifier:**

- Base classifier: Logistic Regression with C=1.0 and liblinear solver
- Similar to OvR but with a slightly different implementation
- Also treats each genre independently

3. **Classifier Chain:**

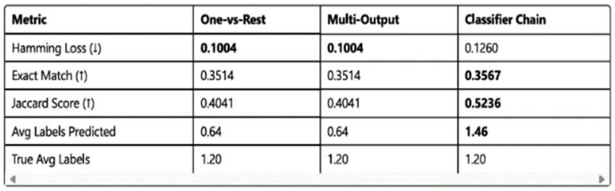
- Base classifier: Logistic Regression with balanced class weights
- Builds a chain of classifiers, each incorporating predictions from previous models
- Chain order based on genre frequency: drama → comedy → action → romance → thriller → crime → horror → fantasy → family
- Models the dependencies between genres

All models were trained on 80% of the data (20,304 samples) and evaluated on a 20% test set (5,077 samples).

4. Results

4.1 Model Performance Comparison

The three models were evaluated using multiple metrics designed for multi-label classification:



Metric	One-vs-Rest	Multi-Output	Classifier Chain
Hamming Loss (↓)	**0.1004**	**0.1004**	0.1260
Exact Match (↑)	0.3514	0.3514	**0.3567**
Jaccard Score (↑)	0.4041	0.4041	**0.5236**
Avg Labels Predicted	0.64	0.64	**1.46**
True Avg Labels	1.20	1.20	1.20

Key findings from the performance comparison:

1. **One-vs-Rest and Multi-Output** produced identical results across all metrics, suggesting their implementations are effectively equivalent for this dataset.

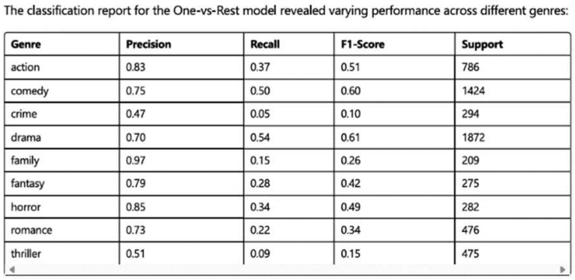
2**. Classifier Chain** achieved the highest Exact Match accuracy (0.3567) and significantly higher Jaccard Score (0.5236), indicating better overall prediction quality.

3. **Label Density:** The One-vs-Rest and Multi-Output models predicted fewer genres per movie (0.64) than the true average (1.20), while the Classifier Chain predicted more (1.46).

4. **Hamming Loss:** Despite higher Jaccard scores, the Classifier Chain had higher Hamming Loss (0.1260), indicating more individual label errors but better overall set predictions.

4.2 Per-Genre Performance Analysis

The classification report for the One-vs-Rest model revealed varying performance across different genres:



The classification report for the One-vs-Rest model revealed varying performance across different genres:

Genre	Precision	Recall	F1-Score	Support
action	0.83	0.37	0.51	786
comedy	0.75	0.50	0.60	1424
crime	0.47	0.05	0.10	294
drama	0.70	0.54	0.61	1872
family	0.97	0.15	0.26	209
fantasy	0.79	0.28	0.42	275
horror	0.85	0.34	0.49	282
romance	0.73	0.22	0.34	476
thriller	0.51	0.09	0.15	475

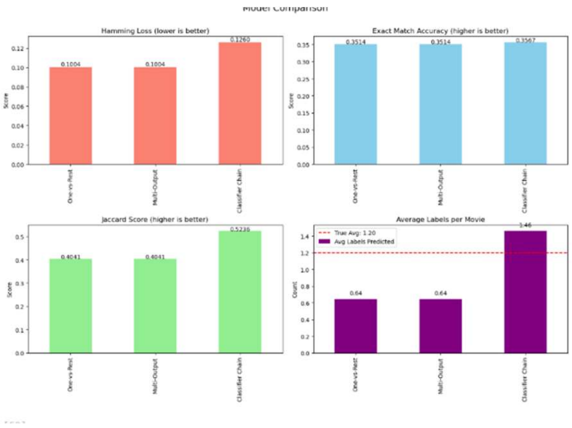
Key observations:

1. **High Precision:** Most genres showed good precision, especially family (0.97), horror (0.85), and action (0.83).

2. **Low Recall:** All genres suffered from low recall, particularly crime (0.05) and thriller (0.09).

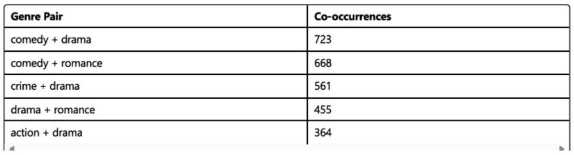
3. **Balanced Performance:** Drama and comedy achieved the most balanced precision-recall trade-off.

4. **Confusion Matrices:** Analysis of confusion matrices showed that drama and comedy were most frequently predicted, while crime and thriller were rarely predicted even when they should have been.



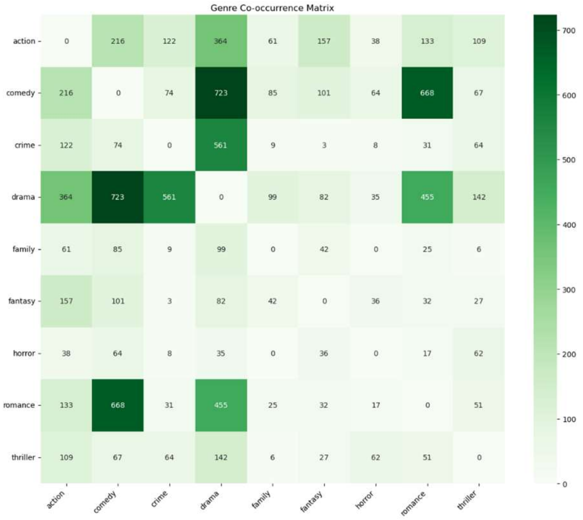
4.3 Genre Co-occurrence Analysis

The analysis of genre co-occurrences in the dataset revealed strong relationships between certain genres:



Genre Pair	Co-occurrences
comedy + drama	723
comedy + romance	668
crime + drama	561
drama + romance	455
action + drama	364

These co-occurrences highlight natural genre combinations in cinema and explain some of the classification challenges, as films rarely belong to a single distinct genre.



	action	comedy	crime	drama	family	fantasy	horror	romance	thriller
action	0	216	122	364	61	157	38	133	109
comedy	216	0	74	723	85	101	64	668	67
crime	122	74	0	561	9	3	8	31	64
drama	364	723	561	0	99	82	35	455	142
family	61	85	9	99	0	42	0	25	6
fantasy	157	101	3	82	42	0	36	32	27
horror	38	64	8	35	0	36	0	17	62
romance	133	668	31	455	25	32	17	0	51
thriller	109	67	64	142	6	27	62	51	0

4.4 Test Case Predictions

To assess practical performance, the model was tested on several example plot descriptions:

1. **"A detective investigates a series of murders in a small town."**

   Predicted: crime (0.3747)

2. **"Two people from different social backgrounds fall in love despite their families' objections."**

   Predicted: drama (0.7057), romance (0.4508)

3. **"A spaceship crew discovers an alien life form that starts killing them one by one."**

   Predicted: fantasy (0.3823)

4. **"A comedian struggles with addiction while trying to revive his failing career."**

   Predicted: drama (0.5063), comedy (0.3523)

5. **"A group of friends plan the perfect heist to rob a bank."**

   Predicted: comedy (0.4443), crime (0.3455)

These examples demonstrate the model's ability to identify appropriate genres from concise plot descriptions and to predict multiple genres when appropriate.

5. Discussion

5.1 Interpretation of Results

The performance comparison between models reveals important insights about multi-label genre classification:

1. **Label Dependencies:** The Classifier Chain significantly outperformed the independent classifiers (One-vs-Rest and Multi-Output) in terms of Jaccard Score, highlighting the importance of modeling genre dependencies. Genres often co-occur in predictable patterns (e.g., comedy-romance, crime-drama), and leveraging these relationships improves prediction quality.

2. **Label Density:** Both One-vs-Rest and Multi-Output models consistently under-predicted the number of genres per movie (0.64 vs. 1.20 true average), indicating a tendency toward conservative predictions. This suggests these models are optimizing for precision at the expense of recall. In contrast, the Classifier Chain slightly over-predicted (1.46), which contributed to its higher Jaccard Score despite increased Hamming Loss.

3. **Genre-Specific Performance:** The models performed best on well-defined genres with distinctive language patterns (horror, family, action), while struggling with broader or more nuanced genres (thriller, crime). This suggests that certain genres have more identifiable linguistic markers in plot descriptions.

4. **Confusion Matrix Analysis:** The confusion matrices revealed that rare genres were

frequently missed (false negatives), while common genres occasionally generated false positives. This pattern reflects the class imbalance in the dataset and suggests that more sophisticated balancing techniques might improve performance for underrepresented genres.

5.2 Implications

The findings from this project have several implications for content categorization and recommendation systems:

1. **Automated Content Tagging:** The models demonstrate sufficient accuracy to assist in preliminary genre classification, potentially reducing the workload for manual content taggers and improving consistency across large content libraries.

2. **Multi-Genre Recognition:** The system's ability to predict multiple genres simultaneously aligns with the reality of modern content, which often spans traditional genre boundaries. This multi-label approach provides a more nuanced view of content than single-label classification.

3**. Content Discovery:** By identifying genre patterns in text descriptions, the system could be extended to improve content discovery by recognizing emerging sub-genres or cross-genre trends.

4. **Text-Based Recommendation:** The NLP techniques applied here could be integrated into recommendation systems that leverage plot similarities in addition to traditional collaborative filtering approaches.

5.3 Limitations

Despite the promising results, several limitations should be acknowledged:

1. **Textual Limitations:** The models rely solely on plot descriptions, missing visual, audio, and structural elements that contribute to genre. Certain genres (e.g., musicals, action) have defining characteristics that may not be fully captured in text.

2. **Genre Subjectivity:** Genre classification inherently involves subjective judgment, and the standardization process necessarily simplified the rich diversity of film genres. Some nuance is lost in condensing hundreds of genre labels to nine categories.

3. **Data Quality:** The quality and consistency of plot descriptions varied across the dataset. Some were detailed and comprehensive, while others were brief summaries that provided limited information for classification.

4. **Class Imbalance:** Despite efforts to address it, class imbalance remained a challenge, particularly for less common genres. This contributed to the low recall for genres like thriller and crime.

5. **Language Model Constraints:** The TF-IDF approach, while efficient, lacks the semantic understanding of more advanced language models. It may miss thematic elements that require deeper contextual understanding.

5.4 Future Work

Several directions for future research could address these limitations and further improve genre prediction:

1. **Advanced Language Models:** Implementing transformer-based models like BERT or GPT could enhance semantic understanding of plot descriptions and potentially capture more subtle genre indicators.

2. **Hierarchical Classification:** Developing a hierarchical approach to genre prediction could

better reflect the nested nature of film genres (e.g., romantic comedy as a subset of comedy).

3. **Multi-Modal Analysis:** Incorporating additional data sources such as movie posters, trailers, or soundtrack information could provide a more comprehensive view for genre classification.

4. **Temporal Analysis:** Exploring how genre definitions and language evolve over time could improve classification for films from different eras.

5. **Attention Mechanisms:** Implementing attention-based models could help identify which parts of plot descriptions are most indicative of specific genres.

6. **Cross-Cultural Analysis:** Expanding the analysis to include non-English language films could reveal cultural differences in genre conventions and descriptions.

6. Conclusion

This project successfully developed and evaluated a multi-label classification system for predicting movie genres from plot descriptions. By preprocessing text data, engineering relevant features, and implementing various classification models, the system achieved promising results in automatically categorizing films across multiple genres.

The comparative analysis of three classification approaches—One-vs-Rest, Multi-Output, and Classifier Chain—revealed the importance of modeling label dependencies in multi-label classification tasks. The Classifier Chain demonstrated superior performance in terms of Jaccard Score and Exact Match accuracy, highlighting its effectiveness for this application.

The analysis also uncovered significant patterns in genre co-occurrence and genre-specific linguistic markers, providing insights into the structure of film categorization. These findings have practical implications for content organization, recommendation systems, and automated metadata generation in the film industry.

While limitations exist, particularly regarding the reliance on textual data alone and challenges with class imbalance, the framework established here provides a solid foundation for future work. The system offers a practical, accessible approach to movie genre classification that could be integrated into various applications, from film databases to streaming platforms.

7. References

1. Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 45(9), 3084-3104.

2. Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. Machine Learning, 85(3), 333-359.

3. Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 26(8), 1819-1837.