From Metadata to Meaning: A Hybrid Clustering and Interpretable Rating Analysis of the Netflix Library

[1*]Siew Mooi Lim, [2]Xue Kang Chok and [3]Qi Xiang Choo

Department of Computer Science and Data Science, Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology, 53300 Kuala Lumpur, Malaysia

email: [1*]siewmooi@tarc.edu.my, [2]chokxuekang@gmail.com, [3]choo2003xiang@gmail.com

**Corresponding author*

Received: 05 December 2025 | Accepted: 14 February 2026 | Early access: 26 March 2026
https://doi.org/ 10.33736/jcsi.11660.2026

Abstract - *The rapid growth of streaming platforms has created vast, heterogeneous content libraries, posing challenges for effective content organisation and understanding of audience preferences. This study aims to uncover multi-faceted patterns in Netflix's content structure, categorisation, and audience reception by employing a hybrid analytical framework. Three distinct clustering methodologies—Latent Dirichlet Allocation (LDA), spectral clustering, and K-Prototypes—were applied alongside IMDb rating analysis, genre inference with TF-IDF, and advanced semantic clustering enhanced by interpretable XGBoost and SHAP values. LDA topic modelling identified three distinct thematic areas in content descriptions. Spectral clustering, using Nearest Neighbours affinity and unnormalized Laplacian regularisation, distinguished three clusters based on geographical origin and content maturity. K-Prototypes clustering identified five segments characterised by distinct format, duration, and regional patterns. Furthermore, IMDb rating analysis provided external validation of content quality, genre inference with TF-IDF elucidated textual markers for genres, and semantic clustering revealed high-value genres and low-value tropes. These findings demonstrate that Netflix maintains a carefully balanced portfolio of content spanning different formats, regions, and themes, catering to diverse audience preferences. The complementary nature of the clustering approaches provides a multi-dimensional understanding of streaming content libraries, offering actionable insights for content acquisition strategies, recommendation systems, and data-driven content personalisation on streaming platforms.*

Keywords: Unsupervised learning, hybrid clustering, semantic analysis, topic modelling, Explainable AI, streaming content analytics.

1 Introduction

The rise of digital streaming platforms has significantly transformed entertainment consumption patterns (Chatterjee, 2023). As one of the leading global providers, Netflix offers an extensive, varied catalogue that spans multiple genres, languages, and regions (Fan, 2024). Managing this scale and diversity poses challenges for organising content and understanding evolving audience preferences.

To address these challenges, this study applies clustering techniques—including Latent Dirichlet Allocation (LDA), Spectral Clustering, and K-Prototypes—to segment Netflix's content based on attributes such as genre, duration, content ratings, and country of production. By uncovering underlying patterns and relationships in the data, these models can identify regional trends and latent content structures.

The analysis aims to support data-driven strategies for content acquisition, production, and personalised recommendations to enhance user experience and optimise content delivery. While structural clustering organises the library, it fails to account for audience reception. Therefore, this study extends the analysis to include semantic quality indicators and their relationship with user ratings. These strategies are further enhanced by insights derived from IMDb rating analysis, sophisticated genre inference techniques, and advanced semantic clustering, with interpretability provided through methods like XGBoost and SHAP values.

2 Literature Review

Recent advances in content clustering have demonstrated the effectiveness of multi-method approaches for streaming platform analysis. This study leverages several complementary techniques to address Netflix's heterogeneous content attributes and complex user preferences.

2.1 Latent Dirichlet Allocation

LDA has demonstrated strong performance in uncovering latent thematic structures within unstructured content, making it effective for large-scale content segmentation and recommendation systems (Blei et al., 2003). Integrating LDA with latent factor models that incorporate user and item features has been shown to improve recommendation accuracy by combining textual reviews with rating data (Aslanyan & Frasincar, 2021). Comparative analyses consistently rank LDA highly on topic coherence when appropriately preprocessed, reinforcing its suitability for modelling multilingual, heterogeneous content libraries (Krishnan, 2023). Recent applications in streaming contexts have shown particular effectiveness in identifying content narrative patterns that traditional metadata approaches miss.

2.2 Spectral Clustering

Spectral clustering is well-suited for modelling complex, graph-based relationships in content datasets (Hess et al., 2019; Luxburg, 2007). Its capacity to uncover non-linear structures supports nuanced segmentation strategies in streaming contexts. Previous applications include analysing daily water consumption patterns to reveal seasonal and behavioural variations (Guo et al., 2024), and clustering time series data by decomposing trend, seasonality, and residual components to capture underlying temporal structures (Gutiérrez et al., 2023). These capabilities can inform content release scheduling and targeted marketing strategies, particularly when geographic and temporal patterns intersect with content maturity ratings.

2.3 K-Prototypes Clustering

K-Prototypes effectively addresses mixed-type datasets, supporting nuanced segmentation strategies in contexts combining categorical and numerical attributes. It has been applied to define geostatistical domains for mineral grade estimation by integrating geological categories with numerical measurements (Hernández et al., 2023), and to cluster shared-ride taxi requests by modelling user profiles that mix demographic and preference data (Mohd et al., 2024). In streaming contexts, this approach enables simultaneous consideration of format, duration, and categorical attributes like genre and region, revealing content portfolio structures that inform strategic decisions.

2.4 IMDb Rating Analysis

The integration of external rating sources like IMDb has become crucial for enriching content analysis on streaming platforms. Studies often leverage IMDb data for Exploratory Data Analysis (EDA) to uncover trends and audience preferences related to genres, release years, and runtime in conjunction with user ratings (Dixit et al., 2020). Furthermore, predictive models, such as Gradient Boosting, have been successfully applied to forecast IMDb ratings for movies and TV shows, utilizing features like genre, cast, director, and release year to estimate potential content success (Lash et al., 2016). These analyses are pivotal for enhancing recommendation systems by providing a more comprehensive understanding of user preferences and content quality beyond internal platform metrics.

2.5 Genre Inference with TF-IDF

Genre inference from textual content, particularly movie descriptions, is a well-established area of research that frequently employs TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction. This method effectively quantifies the importance of words within a movie description, allowing machine learning models to identify discriminative terms for specific genres (Dai et al., 2015). After preprocessing steps such as tokenisation and stop-word removal, TF-IDF vectors are commonly used as input for various classification algorithms, including Naive Bayes, Logistic Regression, and Support Vector Machines, to accurately assign genres to unseen content (Bhandarkar et al., n.d.). These techniques are essential for automating content categorisation and improving search and recommendation functionalities on large streaming platforms.

2.6 Semantic Clustering

Semantic clustering of movie descriptions aims to group films based on their underlying thematic meanings rather than just keywords. This involves advanced Natural Language Processing (NLP) techniques for text preprocessing and feature extraction, often utilizing word embeddings like Word2Vec, GloVe, or Sentence-BERT to capture semantic relationships between words and sentences (Reimers & Gurevych, 2019). Various clustering algorithms, including K-means and Latent Dirichlet Allocation (LDA), are then applied to these semantic representations to uncover nuanced categories that transcend traditional genre classifications. Such approaches are vital for enhancing movie recommendation systems and content organisation by providing a deeper understanding of narrative similarities (Bamman et al., 2013).

2.7 XGBoost for Cluster Interpretation

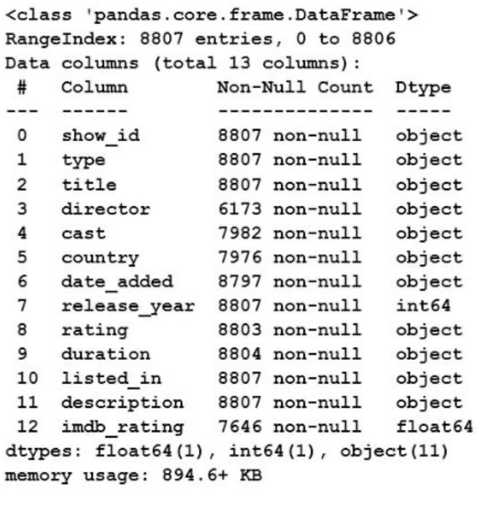
While XGBoost is primarily a supervised learning algorithm, it can be effectively leveraged for interpreting results from unsupervised methods like clustering. By training an XGBoost model to predict cluster assignments based on the features used for clustering, feature importance metrics, particularly SHAP (SHapley Additive exPlanations) values, can reveal the key drivers behind each cluster's formation. SHAP values provide a unified measure of feature importance, indicating how much each feature contributes to the prediction for individual instances, thus offering transparent and interpretable insights into the characteristics that define different movie description clusters (Lundberg & Lee, 2017). This approach allows researchers to understand not just what the clusters are, but why specific content items belong to them.

3 Dataset and Methodology

This section describes the dataset and methodology used to analyse Netflix's content library. The approach includes data preprocessing, exploratory data analysis, and a multi-stage analytical process that progresses from baseline predictive models to advanced semantic clustering. The subsections below explain each methodological component and the dataset's specific attributes.

3.1 Data Collection

The dataset used in this study was obtained from Kaggle (Bansal, 2023) and contains information about movies and TV shows available on Netflix. To further enrich the dataset, an 'imdb_rating' column was integrated from IMDb, providing an external measure of content popularity and critical reception. The dataset overview is shown in Figure 1. While most columns are self-explanatory (such as title, director, cast), several key features warrant explanation: 'duration' represents either the runtime in minutes for movies or number of seasons for TV shows; 'listed_in' contains genre information as comma-separated values; and 'description' provides a text summary of the content, which is used for our topic modelling, genre inference, and semantic clustering analysis.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 13 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
 12  imdb_rating   7646 non-null   float64
dtypes: float64(1), int64(1), object(11)
memory usage: 894.6+ KB
```

Figure 1: Dataset overview

3.2 Data Preprocessing

Data preprocessing focused on optimising feature representation for clustering analysis. The 'listed_in' column was renamed to 'genre' for clarity. Missing values were strategically handled: 'duration' and 'rating' nulls were sourced from Netflix's official database to maintain data integrity; 'date_added' entries were imputed using release year information to preserve temporal patterns; and 'director', 'cast', and 'country' nulls were categorised as "Unknown" to maintain categorical structure. Date standardisation converted string formats to datetime objects for temporal analysis. Outlier analysis using IQR methods identified extreme release years, which were validated against source data and retained as legitimate historical content, preserving the dataset's full temporal scope for clustering analysis.

3.3 Data Analysis and Visualisation

The Exploratory Data Analysis (EDA) of Netflix's content catalogue reveals insights into video types, release patterns, genres, durations, IMDb ratings patterns, and more. These insights are organised into separate sections for movies and TV shows where appropriate due to their inherent differences.

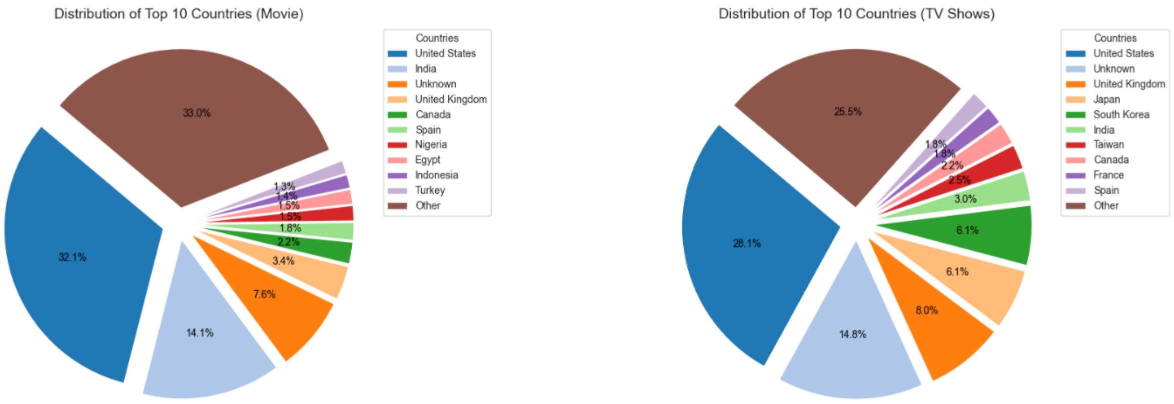


Figure 2: Distribution of country of production for movies and TV shows

Netflix's content catalogue reveals movies constituting 67.8% of the library. As shown in Figure 2, the United States dominates production, followed by India and the United Kingdom, reflecting Netflix's global expansion strategy and local content investment priorities (Fan, 2024).

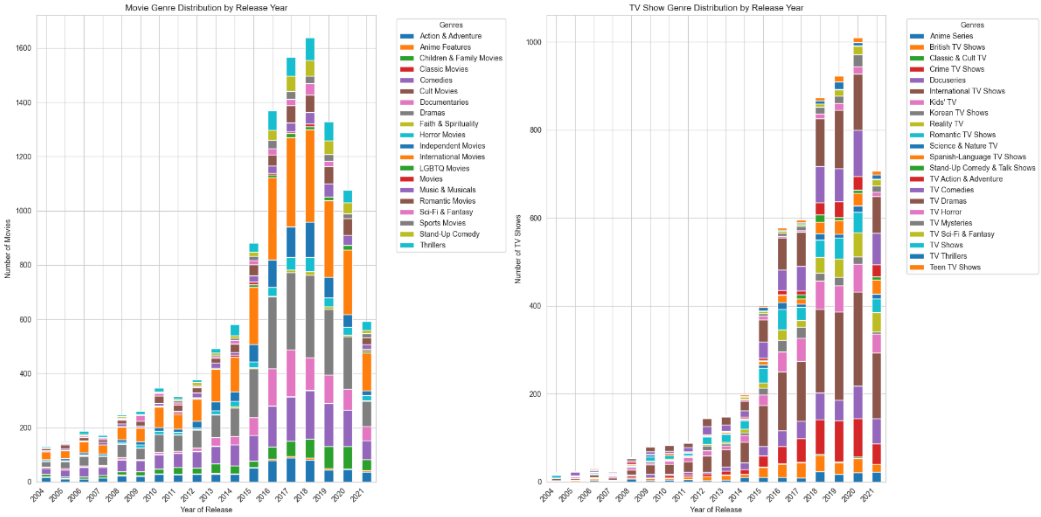


Figure 3: Distribution by release year and genres from 2004 to 2021

As illustrated in Figure 3, both movies and TV shows experience massive growth from mid-2010s onward, likely reflecting streaming-era expansion (Chatterjee, 2023), and 2020 appears to be the production peak for both categories. Genre diversification increases significantly over time, with many small genre segments, such as Korea TV Shows, appearing only in later years.

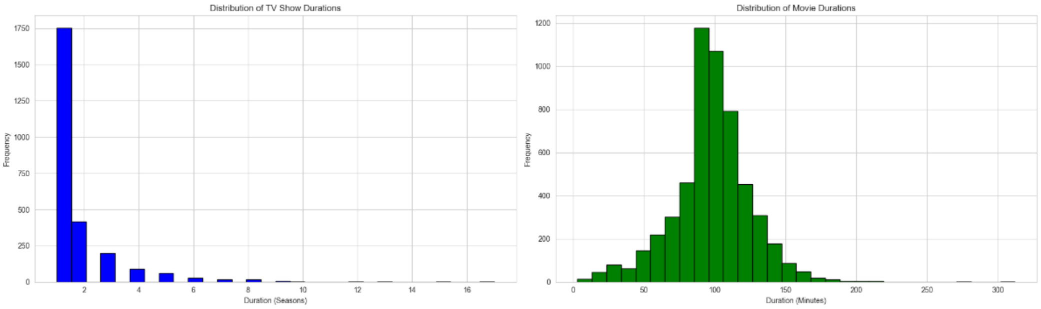


Figure 4: Video durations in seasons (TV shows) and minutes (movies)

Based on Figure 4, duration analysis reveals strategic content portfolio management: TV shows cluster around 1-2 seasons, optimising for binge-watching consumption patterns prevalent on streaming platforms (Matrix, 2014), while movies peak at 100-120 minutes, consistent with standard theatrical runtime conventions (Cutting et al., 2010).

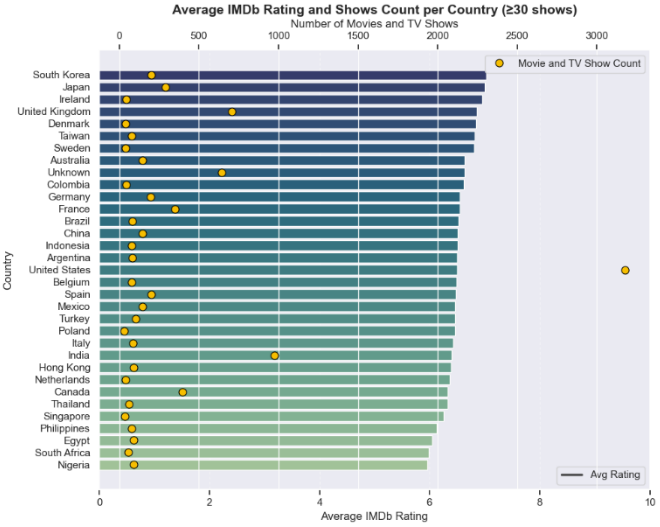


Figure 5: Average IMDb rating of movies and TV shows for each country with more than 30 entries

As depicted in Figure 5, South Korea, Japan, Ireland, United Kingdom, Denmark, Taiwan, and Sweden appear among the highest average ratings (around 7.0). The United States has significantly more content than all other countries, but its average rating is around mid-range.

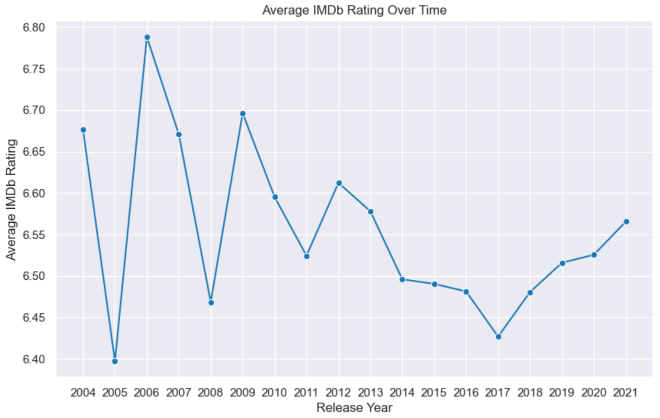


Figure 6: Average IMDb rating for shows released over the years

Based on Figure 6, the trends of IMDb ratings are not linear. Ratings were higher and more variable for the shows released in the mid-2000s. After 2014, ratings stabilise around the mid-6.4 to 6.5 range, with a slight upward shift after 2018.

Love
Christmas
Secret
Black
World
Story
Movie
Life
Day
Time
Man
Last
King
War
Girl
American
House
Game
Boy
family
life
find
love
world
new
take
friend
woman
young
man
home
become
group
team
help
live
one

(a) Titles

(b) Description

Figure 7: Thematic analysis of titles and descriptions

Lexical analysis reveals thematic content strategies through recurring narrative elements. As depicted in Figure 7, "Love", "World", and "Life" dominate titles and descriptions, indicating Netflix's emphasis on universally appealing, emotionally resonant content that transcends cultural boundaries.

3.4 Methodologies

3.4.1 Latent Dirichlet Allocation

LDA analysis focused on Netflix's content descriptions to identify latent thematic structures. Text preprocessing applied standard NLP techniques: special character removal, tokenisation (minimum 3 characters), stop-word elimination, and lemmatisation. TF-IDF weighting enhanced term importance representation for sparse document matrices.

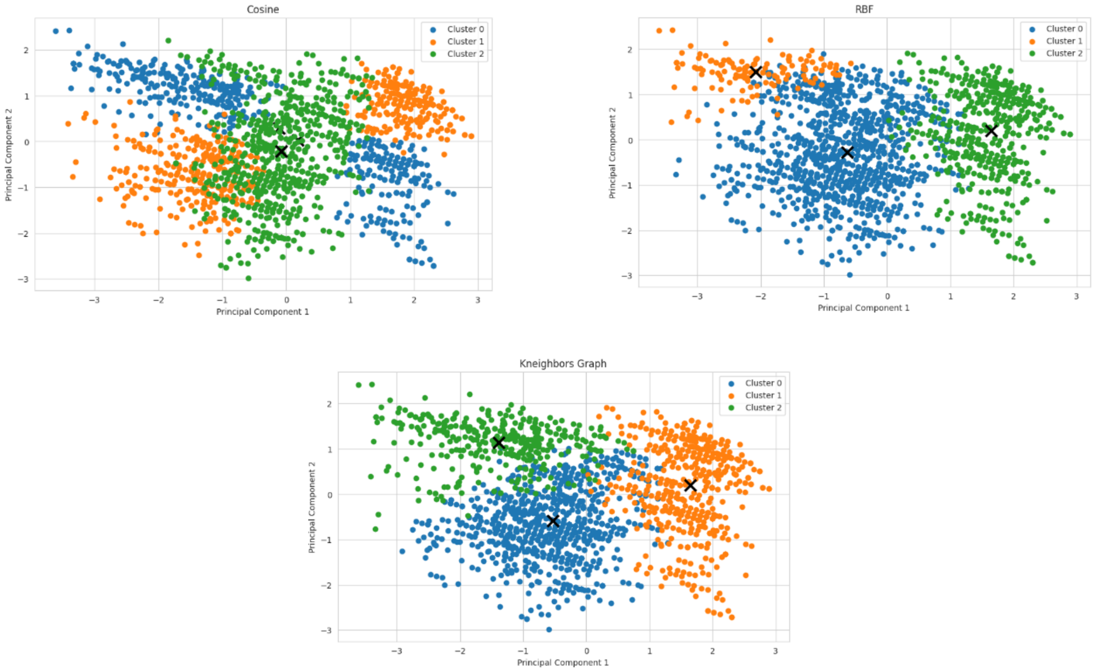
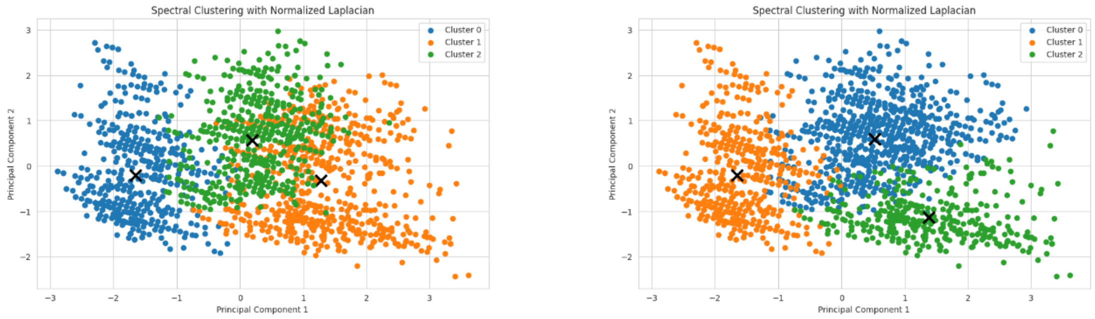


Figure 8: Comparison of different affinity matrices: Cosine (top left), RBF (top right), and KNeighbours Graph (middle)

Parameter optimisation then used grid search with 5-fold cross-validation, testing TF-IDF features (500, 1000, 2000) and topic numbers (3, 5, 7). Model evaluation incorporated Log Likelihood, Perplexity, and CV coherence measures, while Gensim conversion enabled advanced topic coherence analysis, providing comprehensive quality assessment for interpretability.

3.4.2 Spectral Clustering

Spectral clustering utilised mixed-type features: 'video_type', 'country', 'year_release', 'rating', and 'genres'. Preprocessing included label encoding for categorical variables, random sampling for computational efficiency, and standard scaling for distance-based similarity computation. Cluster optimisation used silhouette analysis (2-5 clusters), with 3 clusters achieving optimal performance. As shown in Figure 8, the affinity matrix comparison evaluated cosine, RBF, and nearest neighbours approaches, with nearest neighbours achieving superior Davies-Bouldin Index (1.51), Calinski-Harabasz Index (748.89), and silhouette score (0.28).



(a) Normalised Laplacian

(b) Unnormalised Laplacian

Figure 9: Comparison of normalised and unnormalised Laplacians

Graph regularisation, which modifies the similarity matrix before clustering, was examined using both normalised and unnormalised Laplacians. As illustrated in Figure 9, the unnormalised Laplacian slightly outperformed the normalised variant, producing a higher silhouette score and a lower DBI, indicating better clustering quality.

3.4.3 K-Prototypes Clustering

K-Prototypes preprocessing addressed mixed-type data complexity: numerical features (release_year, duration) were maintained; categorical features underwent one-hot encoding (video_type, rating); multi-value features (genres, countries) were transformed into binary indicators. This approach enabled simultaneous Euclidean/Hamming distance computation across data types. Cluster number selection used the elbow method (k=1-9), indicating k=2 as optimal for variance minimisation; however, k=5 was selected for enhanced interpretability and practical content strategy applications, balancing statistical optimality with actionable business insights.

3.4.4 Predictive Modelling and Genre Inference

The initial analysis focused on establishing a baseline for predicting IMDb ratings using linear regression and Ridge regression models. This step helped to understand the basic relationships between features and ratings and provided a starting point for more complex models, with a grid search performed on the Ridge regression model to find the optimal alpha value. Building on this baseline, a genre inference model was developed using TF-IDF to explore the relationship between content descriptions and genres. This approach allowed for the identification of words and phrases that are most indicative of specific genres. To contrast statistical keyword frequency with contextual meaning, we employed a two-step approach: first utilising TF-IDF for lexical genre inference, followed by sentence embeddings for semantic clustering.

3.4.5 Semantic Clustering and Interpretable Models

The limitations of the TF-IDF-based genre inference, which primarily identifies statistical distinctiveness rather than semantic meaning, led to the adoption of a more advanced approach. To move beyond keyword matching and understand the underlying "vibe" or "theme" of the content, semantic clustering was performed using sentence embeddings from the SentenceTransformer library (`all-MiniLM-L6-v2`) and K-Means clustering. This allowed for the grouping of content based on the semantic similarity of their descriptions. To interpret these clusters and identify "High-Value Genres" and "Low-Value Tropes," an XGBoost model was trained to predict cluster membership. SHAP (SHapley Additive exPlanations) values were then used to provide granular interpretability, highlighting the specific features (words, bigrams) that most strongly influence cluster assignment and, by extension, audience ratings.

4 Results and Discussions

The evaluation of our clustering approaches requires different metrics due to their distinct methodological foundations and data inputs, as outlined in Table 1. Direct comparison using uniform metrics is not appropriate because: (1) LDA is a generative topic model optimised for textual thematic discovery rather than traditional clustering, making silhouette scores inapplicable; (2) spectral and K-Prototypes clustering operate on mixed-type data where traditional distance-based metrics apply. Each method's evaluation strategy aligns with its specific objectives and data characteristics, providing complementary insights into Netflix's content structure.

Table 1: Methodological comparison and evaluation metrics

Method	**Data Type**	**Primary Purpose**	**Evaluation Metrics Used**
LDA Topic Modelling	Text (descriptions)	Topic discovery and thematic analysis	Log Likelihood, Perplexity, Topic Coherence
Spectral Clustering	Mixed (categorical + numerical)	Non-linear cluster discovery	Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index
K-Prototypes Clustering	Mixed (categorical + numerical)	Mixed-type data segmentation	Elbow method, Within-cluster variance, Interpretability assessment
Semantic Clustering	Text (embeddings)	Semantic grouping and quality Inference	Silhouette, SHAP values

Table 2: Parameter optimisation results

Method	**Parameter**	**Values Tested**	**Optimal Value**	**Performance Metrics**
LDA Topic Modelling	Topics (k)	3, 5, 7	3	Coherence: 0.412, Perplexity: 678.56
	TF-IDF Features	500, 1000, 2000	500	Log Likelihood: optimised via grid search
Spectral Clustering	Clusters (k)	2, 3, 4, 5	3	Silhouette: 0.28, DBI: 1.51, CH Index: 748.89
K-Prototypes Clustering	Clusters (k)	1, 2, 3, 4, 5, 6, 7, 8, 9	5	Elbow method indicated k=2, but k=5 selected for interpretability
Semantic Clustering (K-Means)	Clusters (k)	Range 2-15	14 (positive rating) and 12 (negative rating)	Silhouette: 0.0450 (positive rating) and 0.0329 (negative rating)

Table 2 presents the parameter optimisation results for each method, showing the systematic evaluation of different parameter values where applicable. While traditional clustering methods (spectral and K-Prototypes) use k-values with silhouette-based evaluation, LDA employs topic coherence metrics across different topic numbers. This methodologically diverse approach ensures optimal parameter selection tailored to each algorithm's specific characteristics and Netflix's content structure.

Table 3: Comparison of different algorithms on Netflix content

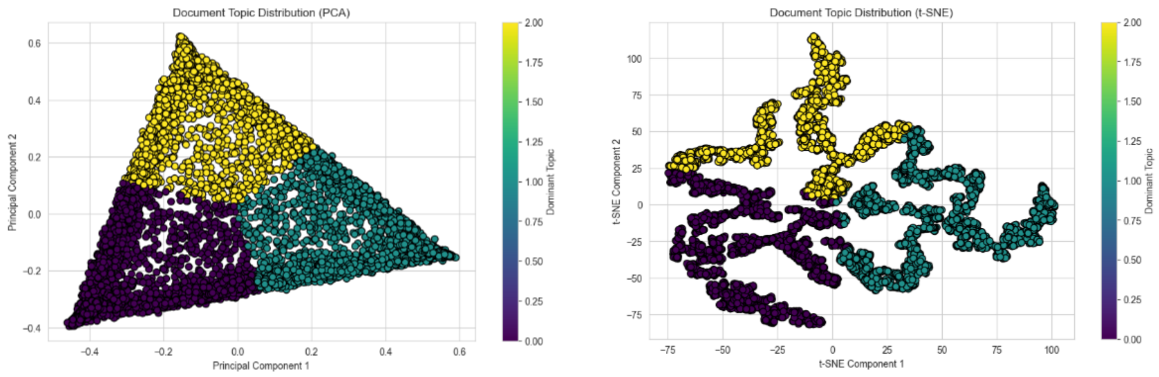
Method	Features	Cluster	Characteristics
LDA Topic Modelling	Description	Topic 1	Documentary-style content and family narratives (key terms: documentary, family, life, father, love)
		Topic 2	Action and crime-related content (key terms: force, murder, crime, power)
		Topic 3	Entertainment and series content (key terms: special, school, team, friend)
Spectral Clustering	Video type, Country, Year release, Rating, Genres	Cluster 0	International films and mature content (circa 2018)
		Cluster 1	U.S. TV shows with mature content (circa 2020)
		Cluster 2	U.S. movies, particularly restricted dramas (circa 2017)
K-Prototypes Clustering	Video type, Country, Year release, Rating, Genres, Duration	Cluster 0	TV Shows (95%), avg. year 2016.6; international TV content, U.S. TV dramas
		Cluster 1	Movies only, avg. 94 mins, year 2013.9; international and U.S. dramas
		Cluster 2	Older movies (avg. year 2004.5), longer duration (avg. 160 mins); Indian content
		Cluster 3	Shorter movies (avg. 61 mins), year 2015; documentaries and children's content
		Cluster 4	Medium duration movies (120 mins), year 2012.6; U.S. and Indian content
Semantic Clustering with K-Means	Description	High rating	Documentaries, Intellectual & Educational Content, Prestige drama
		Low rating	Generic holiday rom-coms, teen movies, derivative horror films

4.1 Structural Portfolio Analysis

Spectral clustering, with three clusters using Nearest Neighbours affinity and unnormalised Laplacian regularisation, identified coherent groupings including international content, mature U.S. TV shows, and restricted U.S. dramas, as summarised in Table 3. These results highlight geographic and maturity-based patterns that can inform demographic-focused recommendations. Complementing this, K-Prototypes clustering identified five distinct clusters, effectively separating TV shows from movies while subdividing movies by duration and genre. The model revealed key portfolio segments such as contemporary TV series, standard-length international films, longer-format cinema, short-form documentaries, and medium-length dramatic content, as detailed in Table 3.

4.2 Thematic Discovery

While Spectral and K-Prototypes organised content by metadata, LDA revealed the latent vocabulary of Netflix. The LDA model was optimised via grid search, selecting three topics and 500 maximum TF-IDF features. It achieved a coherence score of 0.412 and a low perplexity of 678.56. As shown in Figure 10, dimensionality reduction visualisations using PCA and t-SNE show clear topic separation, revealing distinct themes in Netflix's content description texts, as described in Table 3.



(a) PCA visualisation

(b) t-SNE visualisation

Figure 10: Dimensionality reduction visualisations of LDA results

4.3 Semantic and Qualitative Analysis

The analysis of the Netflix dataset was conducted in a multi-stage process, beginning with baseline predictive models, followed by genre inference, and culminating in advanced semantic clustering. This section presents the results and discusses the insights gained at each stage of this analytical progression.

4.3.1 Baseline Predictive Modelling

The initial analysis established a baseline for predicting IMDb ratings. A linear regression model was first trained, followed by a Ridge regression model, for which a grid search identified an optimal alpha of 3.0. The Ridge model provided more interpretable results, highlighting words and phrases with the most significant positive and negative correlations to IMDb ratings. For instance, words like 'documentary,' 'docuseries,' and 'drama' were positively correlated with higher ratings, while words like 'team,' 'commit,' and 'quickly' were negatively correlated.

4.3.2 Genre Inference and Its Limitations

To understand the relationship between content descriptions and genres, a genre inference model using TF-IDF was developed. This model identified key terms that were highly indicative of specific genres. For example, 'documentary' was a strong indicator for the 'Documentaries' genre, and 'drama' for 'TV Dramas'. While this approach was effective for basic categorisation, it offered limited semantic insight, often revealing tautological relationships rather than deeper thematic connections. This limitation highlighted the need for a more nuanced approach to understand the underlying themes and "vibe" of the content.

4.3.3 Semantic Clustering and Interpretable Insights

To move beyond simple keyword matching, semantic clustering was employed using sentence embeddings and K-Means clustering. This approach grouped content based on the semantic similarity of their descriptions, revealing several distinct thematic clusters. For example, one cluster of highly-rated content included words like 'prison sentence,' 'true story,' and 'murder suspect,' suggesting a theme of true-crime documentaries. In contrast, a cluster of low-rated content included words like 'special force,' 'man find,' and 'love life,' pointing towards generic action and romance tropes.

To interpret the drivers behind these clusters, the XGBoost model analysis utilising SHAP values reveals distinct semantic patterns. As illustrated in Figure 11, the term 'documentary' serves as the strongest positive predictor of content quality, alongside terms like 'explore' and 'world,' which suggest audience preference for educational and realist narratives. Conversely, the negative side of the spectrum is dominated by terms such as 'team,' 'rescue,' and 'christmas.' The clear separation of these terms in Figure 11 confirms the 'Trope Penalty,' where content descriptions relying on formulaic action or holiday clichés are statistically penalized by audiences.

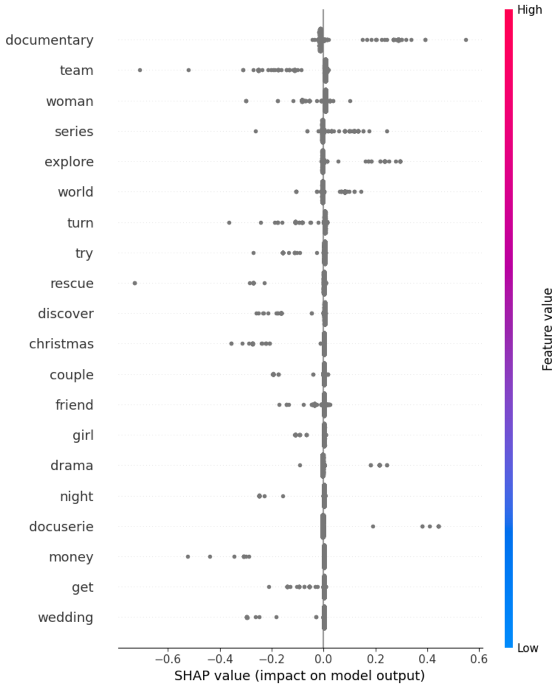


Figure 11: SHAP summary plot: The impact of textual features on IMDb ratings. Features on the right drive higher ratings (e.g., 'documentary'), while features on the left drive lower ratings (e.g., 'team', 'rescue').

4.3.4 The Low R^2 Score: Prediction vs. Inference

The discrepancy between the textual description and the rating suggests that while description features capture the narrative premise (the 'latent potential'), they fail to capture the execution quality (direction, acting, production value). The low coefficient of determination ($R^2 \approx 0.02$) confirms that narrative themes are distinct from production quality. The reason for this is that the description outlines the plot, while the rating reflects the execution. A description of a "grilled steak with potatoes" could apply to both a Michelin-star meal and a burnt diner steak. The textual data lacks information about the director, cast, budget, and cinematography, all of which are critical to the final rating.

However, a low R^2 score does not mean the analysis is without value. While the model fails as a prediction engine, it succeeds in the context of inference. Even with a weak correlation, the direction is consistent. The model indicates that a documentary has a statistically better chance of a high score than a generic Christmas movie. In essence, the low R^2 score proves that in the world of content, the text is cheap, and the rating is everything.

4.4 Strategic Implications

The insights from this multi-stage analysis offer several actionable strategies. The semantic clusters can be used to enhance recommendation algorithms by suggesting content with similar thematic "vibes," rather than just matching keywords (Aslanyan & Frasincar, 2021). The identification of "High-Value Genres" (e.g., crime realism) and "Low-Value Tropes" (e.g., action fantasy) can guide content acquisition and production strategies, informed by the predictive patterns observed in IMDb rating analysis (Lash & Zhao, 2016). Marketing can leverage the key terms from high-rated clusters for more effective copywriting (Liu et al., 2016). These insights provide a more nuanced and data-driven approach to content strategy and personalisation.

5 Conclusion

This study employed three complementary clustering approaches to analyse Netflix's content library, each revealing distinct facets of the platform's content structure. LDA topic modelling uncovered three clear thematic areas in content descriptions with minimal overlap, suggesting distinct vocabulary patterns in how Netflix describes different types of content. Spectral clustering, utilising Nearest Neighbours affinity matrix and unnormalised Laplacian regularisation, identified three clusters that effectively categorised content based on geographical and maturity characteristics—international films and mature content, U.S. TV shows with mature themes, and U.S. dramas with restricted content—particularly highlighting the temporal and regional patterns in Netflix's content strategy. K-Prototypes clustering, focusing on numerical and categorical features rather than text, identified five distinct segments that revealed clear patterns in content format, duration, and origin, from short-form documentaries to lengthy Indian cinema, and from international TV shows to standard-format movies.

Beyond structural clustering, the extended analysis incorporating IMDb rating data, genre inference, and semantic clustering provided deeper insights into content quality and audience reception. IMDb rating analysis offered an external validation of content quality, with regression models highlighting features influencing audience scores. Genre inference with TF-IDF provided a robust method for automated content categorisation, revealing key textual markers for different genres. Semantic clustering, further enhanced by XGBoost and SHAP for interpretability, allowed for the identification of high-value genres and low-value tropes based on textual descriptions, offering actionable insights into narrative elements that resonate with or deter audiences.

Together, these complementary analyses demonstrate that Netflix maintains a carefully balanced and diverse content library serving various audience preferences. The text-based analyses reveal how content is described and marketed, while spectral clustering and K-Prototypes clustering show how the library is structured in terms of geographical, temporal, and technical features. These insights could be valuable for content strategy, recommendation systems, and understanding viewing patterns on streaming platforms. Future research could extend this work by performing network analysis based on shared attributes (e.g., actors, directors) to uncover complex collaborative relationships, or by incorporating temporal dynamics to track how content strategies evolve over time.

Acknowledgements

The authors thank the reviewers for their insightful comments and suggestions, which improved this work. Special thanks to Tunku Abdul Rahman University of Management and Technology (TAR UMT) for providing the necessary resources and facilities that made this research possible.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aslanyan, T. K., & Frasincar, F. (2021). Utilizing textual reviews in latent factor models for recommender systems. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing (SAC '21)* (pp. 1931–1940). Association for Computing Machinery. https://doi.org/10.1145/3412841.3442065

Bamman, D., O'Connor, B., & Smith, N. A. (2013). Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. *1*, pp. 352–361). Association for Computational Linguistics. https://aclanthology.org/P13-1035

Bansal, S. (2023). *Netflix movies and TV shows* [Dataset]. Kaggle. https://www.kaggle.com/datasets/shivamb/netflix-shows/data

Bhandarkar, S., Wolff, M., & Webb, A. (n.d.). *Genre classifications using book and film descriptions* [Stanford CS224N project report]. Retrieved November 24, 2025, from https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-169839493.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://jmlr.org/papers/volume3/blei03a/blei03a.pdf

Chatterjee, K. (2023, June 12). *The rise of streaming platforms: A revolution in entertainment*. Medium. https://medium.com/@kasturichatterjee1108/the-rise-of-streaming-platforms-a-revolution-in-entertainment-3553b094d799

Cutting, J. E., DeLong, J. E., & Nothelfer, C. E. (2010). Attention and the evolution of Hollywood film. *Psychological Science, 21*(3), 432–439. https://doi.org/10.1177/0956797610361679

Dai, A. M., Olah, C., & Le, Q. V. (2015). *Document embedding with paragraph vectors*. arXiv. https://doi.org/10.48550/arXiv.1507.07998

Dixit, P., Hussain, S., & Singh, G. (2020). Predicting the IMDB rating by using EDA and machine learning algorithms. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 6*(4), 441–446. https://doi.org/10.32628/cseit206481

Fan, H. (2024). Leader in the digital entertainment market: Netflix's continued success in a fiercely competitive environment. *Advances in Economics, Management and Political Sciences, 73*(1), 61–65. https://doi.org/10.54254/2754-1169/73/20231226

Guo, H., Liu, X., & Zhang, Q. (2024). Identifying daily water consumption patterns based on K-means clustering, agglomerative hierarchical clustering, and spectral clustering algorithms. *AQUA - Water Infrastructure, Ecosystems and Society, 73*(5), 870–887. https://doi.org/10.2166/aqua.2024.294

Hernández, H., Alberdi, E., Goti, A., & Oyarbide-Zubillaga, A. (2023). Application of the k-prototype clustering approach for the definition of geostatistical estimation domains. *Mathematics, 11*(3), 740. https://doi.org/10.3390/math11030740

Hess, S., Duivesteijn, W., Honysz, P., & Morik, K. (2019). The SpectACl of nonconvex clustering: A spectral approach to density-based clustering. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 3788–3795. https://doi.org/10.1609/aaai.v33i01.33013788

Krishnan, A. (2023). *Exploring the power of topic modeling techniques in analyzing customer reviews: A comparative analysis*. arXiv. https://doi.org/10.48550/arXiv.2308.11520

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems, 33*(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969

Liu, S. X., Yin, J., Wang, X., Cui, W., Cao, K., & Pei, J. (2016). Online visual analytics of text streams. *IEEE Transactions on Visualization and Computer Graphics, 22*(11), 2451–2466. https://doi.org/10.1109/tvcg.2015.2509990

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Matrix, S. (2014). The Netflix effect: Teens, binge watching, and on-demand digital media trends. *Jeunesse: Young People, Texts, Cultures, 6*(1), 119–138. https://doi.org/10.1353/jeu.2014.0002

Mohd, A., Teoh, L. E., & Khoo, H. L. (2024). Passengers' requests clustering with k-prototype algorithm for the first-mile and last-mile (FMLM) shared-ride taxi service. *Multimodal Transportation, 3*(2), 100132. https://doi.org/10.1016/j.multra.2024.100132

Palacios, G. A., Valencia, D. J. L., & Villeta, L. M. (2023). Time series clustering using trend, seasonal and autoregressive components to identify maximum temperature patterns in the Iberian Peninsula. *Environmental and Ecological Statistics, 30*(3), 421–442. https://doi.org/10.1007/s10651-023-00572-9

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT-networks*. arXiv. https://doi.org/10.48550/arXiv.1908.10084

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z

Wang, S., Yabes, J. G., & Chang, C.-C. H. (2021). Hybrid density- and partition-based clustering algorithm for data with mixed-type variables. *Journal of Data Science, 19*(1), 15–36. https://doi.org/10.6339/21-jds996