Plot Twists Over Time: How Movie Stories Have Changed Over 95 Years

Ansel Cheung [*1] **Alessio Villa** [*2] **Bartol Markovinović** [*3] **Martín López de Ipiña** [*4] **Niklas Abraham** [*5]

Abstract

Cinema serves as a cultural archive that reflects evolving societal values and narrative preferences across generations. Understanding how movie stories have changed over time provides insights into cultural evolution, but quantitative analysis of narrative structures at scale has remained challenging. We address this gap by analyzing semantic evolution in cinema through embedding movie plot summaries from 1930 to 2024 into a unified semantic space. Using distance distributions, novelty scores, and Kolmogorov-Smirnov tests, we quantify temporal shifts in narrative structures across nearly a century of filmmaking. Our analysis reveals overall semantic stability in the broader corpus, suggesting that fundamental narrative patterns remain consistent over time. However, we also identify distinct emergence patterns in specific subgenres, demonstrating how thematic categories evolve and consolidate in response to cultural and historical contexts.

1. Introduction

Cinema provides a rich archive of narrative structures that encode evolving societal values across generations. Previous approaches employ keyword search or topic modelling (Dubourg et al., 2023) to explore temporal trends in movie plots. While these older methods might be insightful but challenging, recent computational work (Xu et al., 2020) has revealed hidden cultural patterns in large narrative corpora. One such analysis conducted on the musical domain uses high dimensional embeddings to observe changes to structural properties over time (Di Marco et al., 2025). We build upon these foundations by leveraging advances in large language models (LLMs) to embed movie plot summaries (Sreenivasan, 2013) into a unified semantic space. Using novel statistical methods, we quantitatively analyze how movie narratives evolve over time.

2. Data and Methods

2.1. Data Collection

The initial dataset was constructed by querying Wikidata, an open knowledge graph structured as a network of items, connected together by properties. We used its SPARQL API to extract items classified as films with release dates ranging from 1930 to 2024. In order to adhere to Wikidata's query size limitations, we gathered the data iteratively by release year. We first retrieved QIDs of items describing feature films, then processed them in small batches to gather the rest of relevant features including title, release date, duration, genres, directors, actors, English Wikipedia link, and very importantly links to external movie databases TMDb and IMDb. During collection of QIDs, items classified as short film or television series episode were excluded, and final filtering of non-feature films was performed during data cleaning.

Second, we enriched the dataset with TMDb statistics, retrieving vote counts, vote averages, and popularity for each film. These served as proxies for audience engagement and informed later filtering and weighting.

The third stage, the most data-intensive, focused on obtaining full-text plot summaries. Leveraging Wikipedia sitelinks from Wikidata, we accessed each film's Wikipedia page to extract the plot section. Wikipedia's editorial standards ensure relatively uniform and neutral plot descriptions, facilitating standardized comparative semantic analysis. This step used the Wikipedia API for article retrieval, section extraction, and text normalization, transforming metadata into the dense textual data required for downstream embedding.

The last enrichment stage addressed limitations in TMDb voting statistics. Many movies in the corpus lacked TMDb ratings or vote counts, and when available, these counts were often substantially lower than those reported by other platforms. To address this issue, we enriched the dataset with IMDb vote averages and vote counts, obtained from IMDb's non-commercial data files and merged using the

[*]Equal contribution [1]Matrikelnummer 7274374, MSc Machine Learning [2]Matrikelnummer 7306912, MSc Computer Science [3]Matrikelnummer 7324790, MSc Machine Learning [4]Matrikelnummer 7293076, MSc Machine Learning [5]Matrikelnummer 7307188, MSc Machine Learning. Correspondence to: AC <ansel-heng-yu.cheung@uni-tuebingen.de>.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201). Style template based on the ICML style files 2025.

IMDb title identifier. The inclusion of IMDb data ensures broader coverage and higher vote volumes, resulting in a more stable measure of audience reception.

All data sources are open and appropriately licensed. Wikidata (Wikimedia Foundation, 2024a) is released under CC0 1.0 Universal (public domain). Wikipedia (Wikimedia Foundation, 2024b) is under CC BY SA 4.0, and TMDb (TMDb, 2024) under CC BY NC 4.0, allowing non-commercial research with attribution. This ensures reproducibility and legal compliance.

After data processing, our final dataset consisted of 141,119 movies with 81% average feature coverage.

2.2. Data cleaning

After collecting the raw movie data from Wikidata, TMDb and Wikipedia, we first ensured that our dataset does not contain any duplicates with respect to Wikidata QIDs and Wikipedia links, and then applied several filtering steps described in Figure 1.

The most critical filtering step was the removal of movies with low-entropy plots. Raw dataset contained a significant number of incomplete or overly brief plots (e.g. this). To identify and remove such movies we employed a filtering method inspired by (Wenzek et al., 2019), who used perplexity of a Large Language model to filter out low quality documents. While (Wenzek et al., 2019) utilized perplexity of a 5 gram language model trained on high quality data, we tokenized the plots with the BGE-M3 tokenizer and computed the Shannon entropy of the token distribution for each plot. To determine the optimal entropy threshold, we sampled 150 movies from the borderline entropy region of $[4.0, 5.5]$ and manually annotated them as either *good* or *bad* quality. The threshold of 4.8398 was chosen to maximize the $F\beta$ score with $\beta = 0.5$ prioritizing precision over recall.

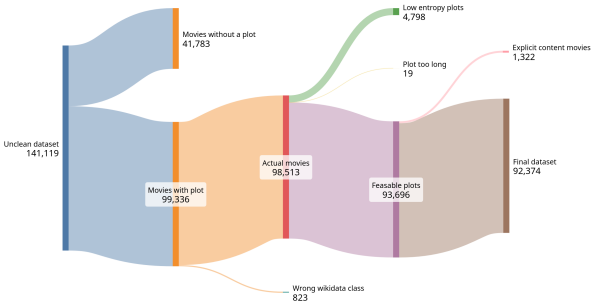

Figure 1. Data cleaning pipeline: number of movies retained after each filtering step.

Embedding of the movie plot summaries. For embedding, we selected the BGE-M3 model (Chen et al., 2024) based on its strong showing (28th place) on the Massive Text Embedding Benchmark (MTEB) leaderboard[1]. BGE-M3 is compact (0.5B parameters), supports 8,192-token context length, and can embed entire movie plots in a single vector.

We used a single static model for all time periods to ensure unified representation in a common latent space. Since Wikipedia summaries are modern English (maintained since 2001), linguistic drift is negligible. After evaluating chunking and pooling methods (Devlin et al., 2018; Gong et al., 2019; Raffel et al., 2023), we selected CLS Token (details: (Group, 2025b)).

Internal data validation. To check that the embeddings meaningfully represent plot similarity, we measured cosine distances within and outside major movie franchises (Harry Potter, Star Wars, James Bond). Movies from the same franchise consistently clustered closer together (e.g., Harry Potter: 0.21 inner-group distance vs. 0.57 to random movies), while the global average distance between any two movies was 0.52. This confirms that the embeddings capture true semantic similarity.

Genre Taxonomy. The raw dataset included 975 unique genre labels, many of which were redundant or highly similar. To simplify and standardize the taxonomy, we first removed genres that appeared only once, reducing the set to 463. We then embedded the Wikipedia descriptions (available for 359 genres) using the BGE-M3 model, and clustered these vectors with $k = 20$ using k-means. Each cluster was manually labeled based on thematic similarity, resulting in 20 coherent genre categories used in further analyses.

3. Methodology

With data collected and cleaned, we embed movie plot summaries into a semantic space and analyze the resulting embeddings using three complementary approaches: distance-based drift analysis, novelty scoring, and Kolmogorov-Smirnov tests.

Distance analysis. Once movie plots are embedded into a unified semantic space, quantitative analysis of their geometric relationships becomes possible through distance metrics. The cosine distance between embeddings provides a natural measure of semantic dissimilarity, enabling the construction of cumulative distribution functions over pairwise distances within defined subsets of the corpus.

As movie genres provide a meaningful taxonomy with potential temporal evolution patterns, we examine semantic drift across different time periods of an arbitrary number of years. To this end, embeddings are first grouped by genre g into discrete time periods τ, forming the set $\mathcal{M}_g^{(\tau)}$ of plot embeddings. For each group, two alternative representative

[1] https://huggingface.co/spaces/mteb/leaderboard

embeddings are computed: the **centroid** (arithmetic mean) $\bar{\mathbf{e}}_g^{(\tau)}$ and the **medoid** (cosine distance minimizer embedding) $\tilde{\mathbf{e}}_g^{(\tau)}$. We employ the following metrics to analyse the drift dynamics across the groups: (i) genre drift and acceleration, and (ii) inter-genre distance.

Genre drift measures the cosine distance between representative embeddings of consecutive periods, capturing how much a genre's semantic center evolves over two time periods. Acceleration quantifies the change in drift between consecutive periods.

Inter-genre distance determines cosine distance between representatives of each pair of genres for each time period, enabling pairwise comparison between specific genres.

Since our dataset contains significantly more movies in later years, the medoid and centroid embedding estimators are heavily affected by sampling variance. To mitigate this, we downsample to ensure equal sampling error across all genre-year groups and apply bootstrapping to estimate confidence intervals.

Novelty analysis. To investigate the claim that movies are becoming less novel over time, we propose a novelty metric defined as the minimal cosine distance between a specific movie's plot embedding and the embeddings of all movies in the dataset released prior to it. This can be formally written as:

$$\text{Novelty}(m_i) = \min_{j:year_j<year_i} \left(1 - \frac{E(m_i)\cdot E(m_j)}{\|E(m_i)\|\|E(m_j)\|}\right) \tag{1}$$

where $E(m)$ denotes the embedding vector of a movie's plot. Intuitively, a higher novelty score indicates that the movie's plot is more dissimilar from prior movies, while a lower score implies existence of a very similar movie released earlier. To compute these scores, we leverage the Faiss library (Douze et al., 2024) to efficiently find the preceding nearest neighbors of movie plot embeddings based on the cosine distance.

Kolmogorov-Smirnov test. To compare distance distributions across movie subsets (e.g., different decades or genres), we employ the Kolmogorov-Smirnov test (Massey, 1951). We construct ϵ-balls ($\epsilon \in [0.24, 0.30]$) around anchor movies representing specific themes, collecting all movies within the distance threshold to define local semantic neighborhoods. We compare distance distributions and temporal CDFs (cumulative distribution functions of release years) within these neighborhoods to a control group (global mean embedding). Divergence indicates temporal shifts: if semantic structure remains stable, distributions should be similar across time periods.

4. Results

In this section, we present and interpret the main empirical findings of our analysis on the embedding space of movie plot summaries.

4.1. General Spatial Analysis

We begin by examining the global structure of the embedding space through pairwise cosine distances between movies. These distances are approximately normally distributed, with a mean of $\mu = 0.5195$ and a standard deviation of $\sigma = 0.0624$, capturing the typical dissimilarity between movie plots and serving as a baseline for further analysis.

To evaluate whether genre labels map to distinct regions, we compared intra-genre (mean: 0.5042) and inter-genre (mean: 0.5268) distances across 19 genres. The resulting separation gap of 0.0226, ratio of 1.0448, and a silhouette score of -0.0334 indicate considerable overlap, with genres only weakly separated in the embedding space. This suggests genre boundaries are porous, consistent with the hybrid and overlapping nature of film categories.

4.2. Genre drift analysis

After normalization, the cosine distance of each time group representative embedding with respect to the previous one exhibited only random fluctuation with no clear trend; inter-genre analysis yielded the same result. These outcomes suggest that genres may be too broad as analytical categories, with any underlying patterns likely obscured by noise.

4.3. Spread analysis

We analyzed the spread of movies per year using three metrics: mean L2 norm, Frobenius norm, and spectral norm of each movie embedding relative to its yearly centroid. All metrics were computed on centered yearly embeddings (i.e. yearly centroid was subtracted from each movie embedding before computing the norms) (Yamagiwa & Shimodaira, 2024). As with the drift analysis, no interpretable patterns emerged: mean L2 norm and Frobenius norm stayed relatively constant at 0.7 and 12.4 respectively, whereas spectral norm had a slight increase from 2.1 to 2.8. This indicates that the overall spread of movies remains relatively constant over the years with outliers becoming more polarizing. The nature of these polarizing axes proved difficult to characterize, as they changed yearly and were a combination of multiple dimensions.

4.4. Novelty score

In order to assess if temporal trends of novelty scores exist, we plot the average novelty score per year alongside

scattered individual movie scores in Figure 2.

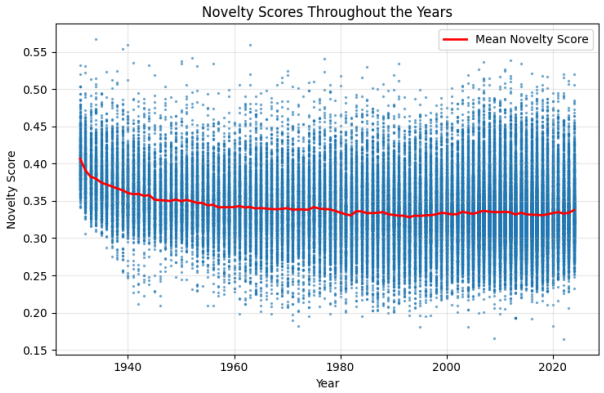

Figure 2. Novelty scores of movies over time. The blue line represents the average novelty score per year, while individual movie scores are shown as scattered points.

4.5. Kolmogorov-Smirnov test

We illustrate our approach using James Bond films as anchor movies to compare their distance distributions against all other movies. The cumulative distribution of cosine distances from Bond anchors rises steeply only for a small set of closely related films, while most movies remain more distant. This contrasts with the global mean embedding, which is closer on average to all movies, as it represents an average narrative rather than a specific subgenre.

To examine the temporal dimension, we constructed cumulative distribution functions of release years for movies within the epsilon ball and compared them to the control group. Figure 3 shows that the temporal distributions differ markedly. The left panel reveals a divergence beginning approximately in the 1960s, suggesting that the spy movie subgenre represented by the Bond anchor exhibits a distinct temporal emergence pattern compared to the broader corpus. The right panel displays normalized histograms of movie counts per year for both groups, confirming that the temporal distribution of spy-themed films diverges from the overall temporal distribution of cinema. This temporal divergence indicates that the spy film subgenre experienced a period of increased production and thematic consolidation that is not representative of general cinematic trends during the same period.

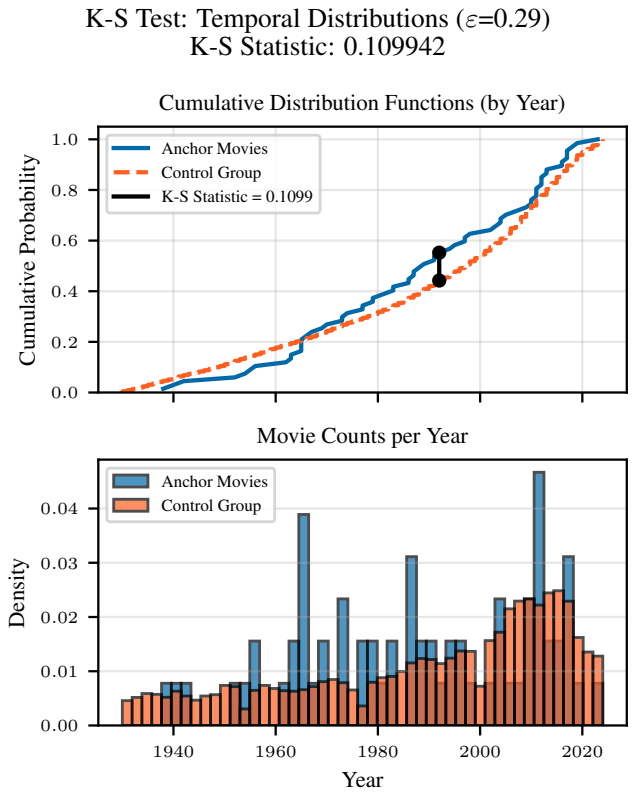

Figure 3. KS test of temporal distributions for James Bond epsilon ball ($\varepsilon = 0.29$) versus control group, showing temporal divergence. Left: Cumulative distribution functions. Right: Normalized yearly counts.

5. Discussion & Conclusion

We have utilized modern embedding methods and multiple statistical tools to analyze the evolution of movie plot embeddings over nearly a century of cinema. Our findings indicate that while the overall semantic structure of movie plots remains relatively stable, specific thematic subgenres exhibit distinct temporal emergence patterns.

We must acknowledge the limitations that arise from our data sources. Wikipedia plot summaries, while standardized, may not fully capture the nuances of original narratives, potentially introducing bias. Additionally, our reliance on a single embedding model, while ensuring a unified semantic space, may overlook temporal linguistic shifts. More importantly, the evolution of cinema is not only reflected in plot summaries but also in cinematography, direction, music, feeling, acting and other non-textual elements. Future work could explore multimodal embeddings that integrate visual and auditory features alongside textual data.

All code and data used in this project are openly available in our GitHub repository(Group, 2025b) and as a dedicated dataset on Hugging Face(Group, 2025a).

Contribution Statement

Contribution Statement:

- **Ansel Cheung:** Performed genre classification analysis, classification of movie plots into genres, and conducted genre drift and PCA analysis of the movie plots.
- **Alessio Villa:** Developed and maintained the IMDb and TMDb API pipelines, and contributed to the related work research and methods background sections.
- **Bartol Markovinović:** Defined the data pipeline cutoff and carried out resulting data cleaning, managed the integration of Wikidata, and conducted novelty score analysis.
- **Martín López de Ipiña:** Carried out genre drift statistical analysis on the general embedding space, performed general spatial analysis of embeddings, and analyzed the cosine distance distributions.
- **Niklas Abraham:** Performed embedding model selection and evaluation, analyzed chunking methods, and performed KS test and distance distribution analysis.

Overall, all authors contributed equally to the project. This is reflected in the various analysis sections throughout the report, where each member's work formed an integral and balanced part of the final study.

References

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024. URL https://arxiv.org/abs/2402.03216.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Di Marco, N., Loru, E., Galeazzi, A., Cinelli, M., and Quattrociocchi, W. Decoding musical evolution through network science, 2025. URL https://arxiv.org/abs/2501.07557.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library. 2024.

Dubourg, E., Mogoutov, A., and Baumard, N. Is cinema becoming less and less innovative with time? using neural network text embedding model to measure cultural innovation. In Šeļa, A., Jannidis, F., and Romanowska, I. (eds.), *Proceedings of the Computational Humanities Research Conference 2023*, volume 3558 of *CEUR Workshop Proceedings*, pp. 676–686, Paris, France, December 2023. URL https://ceur-ws.org/Vol-3558/paper7806.pdf.

Gong, L., He, D., Li, Z., Qin, T., Wang, L., and Liu, T. Efficient training of BERT by progressively stacking. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2337–2346. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/gong19a.html.

Group, D. L. Movie plot embeddings dataset. https://huggingface.co/datasets/NiklasAbraham/MoviePlotEmbeddingsDataset, 2025a. Licensed under MIT License.

Group, D. L. Our github repository. https://github.com/NiklasAbraham/GroupDataLiteracy, 2025b. Licensed under MIT License.

Massey, F. J. The kolmogorov–smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46 (253):68–78, 1951.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Sreenivasan, S. Quantitative analysis of the evolution of novelty in cinema through crowdsourced keywords. *Scientific Reports*, 3(2758), 2013. doi: 10.1038/srep02758. URL https://www.nature.com/articles/srep02758.

TMDb. The movie database (tmdb). https://www.themoviedb.org, 2024. Licensed under CC BY-NC 4.0 for non-commercial use.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. Ccnet: Extracting high quality monolingual datasets from web crawl data, 2019. URL https://arxiv.org/abs/1911.00359.

Wikimedia Foundation. Wikidata. https://www.wikidata.org, 2024a. Licensed under CC0 1.0 Universal (Public Domain).

Wikimedia Foundation. Wikipedia, the free encyclopedia. https://www.wikipedia.org, 2024b. Licensed under CC BY-SA 4.0.

Xu, H., Zhang, Z., Wu, L., and Wang, C.-J. The cinderella complex: Word embeddings reveal gender stereotypes in movies and books. *PLOS ONE*, 15(e0225385), 2020. doi: 10.1371/journal.pone.0225385.

Yamagiwa, H. and Shimodaira, H. Norm of mean contextualized embeddings determines their variance, 2024. URL https://arxiv.org/abs/2409.11253. Accepted to COLING 2025.