Data Literacy 2025 Project Report

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

Abstract

Cultural narratives encode and transmit evolving societal values, yet quantifying how meanings change over time remains methodologically challenging. This project investigates semantic evolution in cinema by analyzing how genres and thematic clusters shift within a unified semantic space across multiple decades. By representing movies as embeddings and tracking their geometric trajectories measuring velocity, acceleration, and curvature we distinguish periods of gradual semantic drift from moments of structural reorganization in cinematic history. This framework provides a quantitative foundation for understanding cultural change at scale and tests whether established linguistic laws of semantic evolution extend to film as a cultural medium.

1. Introduction

Cinema provides a rich archive of narrative structures that encode evolving societal values across generations. Stories serve not only to entertain but to instruct, and those narratives that align with existing social values are more likely to survive and propagate through collective memory. Recent computational work has revealed hidden cultural patterns in large narrative corpora. (Xu et al., 2020) used word embeddings to uncover systematic gender stereotypes in movie synopses, revealing the "Cinderella complex" where female characters' happiness depends asymmetrically on male characters. (Matthews & Glitre, 2021) applied topic modeling to investigate genre structure and temporal evolution, demonstrating that lexical features capture meaningful genre conventions and showing how genres shift in composition over time. These studies establish that quantitative methods can illuminate cultural phenomena at scales beyond traditional close reading, revealing patterns that operate across thousands of narratives.

However, measuring semantic change in cultural narratives over extended historical periods remains methodologically challenging. Previous approaches have examined genre structure at specific moments or through discrete topic models that capture lexical shifts but not the continuous geometric evolution of semantic categories. Can we characterize not merely that genres change, but how they change whether through gradual drift, sudden discontinuities, or cyclical patterns? Furthermore, while linguistic corpora have been analyzed for semantic drift using diachronic word embeddings, these methods require temporal alignment procedures that introduce potential artifacts when comparing meanings across decades.

We address these questions by constructing a unified semantic space from a large corpus of film plot summaries spanning multiple decades. By embedding all narratives into a single static vector space, we eliminate temporal alignment requirements while preserving fine grained semantic relationships. Within this space, we represent genres and thematic clusters as centroids and track their trajectories over time. By computing geometric properties of these trajectories including velocity, acceleration, and curvature we can distinguish periods of continuous semantic evolution from moments of structural reorganization where genres undergo fundamental conceptual shifts. This geometric analysis reveals not just that meanings change, but the dynamics of how they change, providing quantitative measures of cultural evolution.

2. Related Work

There have been several notable works in the area of movie plot analysis using natural language processing. (Sreenivasan, 2013) makes use of SBert to create text embeddings for their movie dataset, which was also curated from Wikipedia and IMDB. They devised a "Innovation Score" which was mean cosine distance. This paper's interesting results include Awarded movies having higher IS, and that

---
[*]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: Initials1 <ansel-heng-yu.cheung@uni-tuebingen.de>, Initials2 <alessio.villa@student.uni-tuebingen.de>, Initials3 <bartol.markovinovic@student.uni-tuebingen.de>, Initials4 <martin.lopez-de-ipina-munoz@student.uni-tuebingen.de>, Initials5 <niklas-sebastian.abraham@student.uni-tuebingen.de>.

IS over time is somewhat related (28% explained variance) to negative quadratic relationship which could be explained by how the movie space has reached saturation of ideas and any more new movie ideas would probably be similar to another movie which came out before.

(Dubourg et al., 2023) is another relevant paper which uses crowdsourced keywords from the IMDB as a window into the contents of films and prescribe novelty scores for each film based on occurrence probabilities of individual keywords and keyword pairs. They devised a method to assign a novelty score to each film on the basis of the keywords associated with it and the keywords appearing in all films that were released prior to it. The results show that there was a surge in novel movies about the time when Hollywood went into the golden age. They also explored how novelty affected film revenue.

However little to none of the papers analyzes text embeddings from LLMs in a temporal manner.

3. Data and Methods

3.1. Data Collection

We constructed our movie corpus using a multi stage pipeline that systematically integrated three complementary sources: Wikidata, The Movie Database (TMDb), and Wikipedia. This approach combines rich structured metadata with the detailed textual content required for semantic analysis.

Initial dataset was constructed by querying Wikidata for movies released from 1930 to 2024. In order to adhere to Wikidata's query size limitations, we iterated through the years and first acquired QIDs of all Wikidata items which have a Wikidata class that is an indirect subclass of film and have a first publication date in the given year. During this step we removed QIDs of items that do not have an English Wikipedia page associated with them. We also tried to remove non-feature movies by excluding subclasses of classes "short film" and "television series episode". However, this filtering was not perfect and further filtering of Wikidata classes was performed during post processing. After acquiring the list of identifiers, we processed them in small batches of 20 and queried Wikidata for each movie's features including title, release date, duration, genres, directors, actors, English Wikipedia link, and very importantly links to external movie databases TMDb and IMDb. Additionally, box office, box office currency, budget and budget currency values were also queried, but they had very low coverage in the raw dataset and were not used in the final analysis.

Second, we enriched the dataset using TMDb, a community driven database that offers quantitative measures of popularity and user engagement. Wikidata's external identifiers enabled direct mapping to TMDb entries, from which we programmatically retrieved vote counts, vote averages, and popularity metrics for each film. These measures served as proxies for audience engagement and cultural impact, informing downstream film 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 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 the data was collected in a tabular format, the textual plot descriptions required transformation into vector representations via a suitable embedding model for downstream analysis. The plot descriptions extracted from Wikipedia pages exhibit substantial variability in length, ranging from 10 to 20,479 characters, corresponding to approximately 6 to 5,296 tokens in an English tokenizer. All plot descriptions in our corpus are in English, which simplifies the embedding process by eliminating cross lingual considerations. After performing the explicitly described data pipeline steps, the final dataset contained 161,533 data points (movies) with a average coverage of 81% in the categories of actors, directors, genres, and year.

3.2. 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. Then we performed the following data filtration and cleaning steps:

- **Filtering out movies without a Wikipedia plot.**

- **Removal of non feature movies.** We removed samples from our dataset that had a Wikidata class that is an indirect subclass of a class that does not describe a feature movie. Some examples of not feasible Wikidata classes include trailers, television series episodes, short films and radio programs.

- **Filtering out movies with excessively long plots.** We filtered out movies with plots longer than 14,000 characters from our dataset because these plots are labeled by Wikipedia as *excessively long*.

- **Removal of movies with low entropy plots.**

- **Genre filtering.** We filtered out genres that appear only once in the dataset because these genres obviously do not describe a group of movies.

- **Exclusion of explicit content.** We excluded movies whose primary genres fell within explicit or highly exploitative categories, such as Bavarian porn, Nazi exploitation, cannibal film, cartoon pornography, erotic film, sexploitation film, and related genres. These categories were removed in order to focus our analysis on mainstream cinematic narratives, to avoid the distorting effects that fringe, pornographic, or exploitation genres would have on cultural and semantic trends in the wider corpus, and because including them would not have been appropriate or useful for a university project of this scope.

The most critical cleaning 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 compute 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.

3.2.1. EMBEDDING OF THE MOVIE PLOT SUMMARIES

The selection of an appropriate embedding model was guided by the Massive Text Embedding Benchmark (MTEB) leaderboard results[1], which provides comprehensive eval-
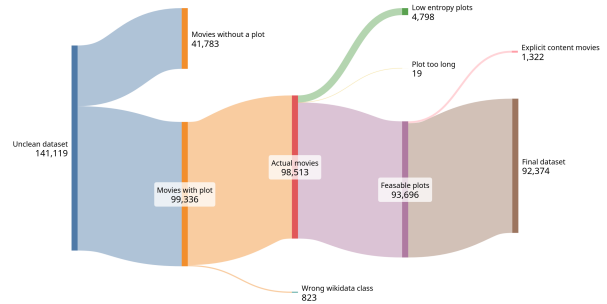

Figure 1. Data cleaning pipeline: number of movies retained after each filtering step.
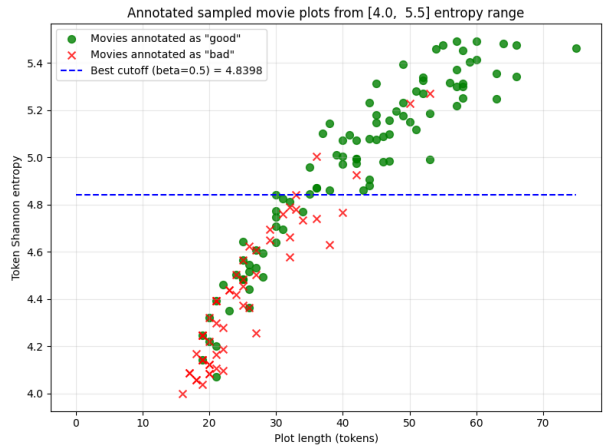

Figure 2. Results of manual labeling of 150 plots in the borderline entropy region with the chosen entropy threshold

uations of embedding models across diverse retrieval and semantic similarity tasks. Based on these benchmarks, we selected the BGE-M3 (Beijing Academy of Artificial Intelligence Multilingual, Multifunctional, Multi granularity) model (Chen et al., 2024). The BGE-M3 model achieved competitive performance (28th place on the MTEB leaderboard) while maintaining a relatively compact architecture with 0.5 billion parameters. Critically, the model supports a context length of 8,192 tokens, which enables embedding entire movie plot descriptions into a single vector representation without requiring chunking.

A key methodological choice was to use a single, static embedding model for all time periods, rather than training separate or temporally aligned models. This ensures all film plots are represented in a unified latent space, avoiding complex post hoc alignment and minimizing artifacts. The Wikipedia based plot summaries are not contemporaneous texts from those eras, instead, they are modern English descriptions collectively maintained and updated since Wikipedia's founding in 2001. Thus, any linguistic variation or semantic drift in the summaries themselves is

---

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

minimal.

Embedding variable length documents presents challenges for transformer based models due to fixed context windows and representational biases. Because plot summaries in our dataset span from a few sentences to thousands of words, it is crucial to select chunking and pooling strategies that minimize length bias while retaining semantic content. Relying solely on the [CLS] token for global representation can introduce substantial bias: it is sensitive to input length, often overemphasizes the first 128 tokens (Devlin et al., 2018; Gong et al., 2019), and underperforms mean pooling on long documents (Raffel et al., 2023). With over 75% of our plots exceeding 512 tokens, a more robust aggregation method is required.

We evaluated four document embedding approaches: Mean Pooling (averaging token embeddings), CLS Token (using pretrained [CLS] token), Early Chunk-then-Embed (splitting documents before embedding), and Late Embed-then-Chunk (embedding full documents then aggregating). Each method offers different tradeoffs between bias, variance, and semantic preservation. We systematically compared these approaches on 5,000 movie plots using four metrics: length bias (correlation between document length and embedding norm), isotropy (variance in first principal component), genre classification accuracy, and class separation (silhouette scores and separation ratios). Results showed substantial variation: length-norm correlation ranged from 0.366 to 0.822, with CLS Token achieving near zero correlation (0.004). Isotropy varied from 3.32% to 11.92%, while genre classification accuracy showed minimal differences (0.326 to 0.349).

Given that no single method dominated across all dimensions, we selected CLS Token based on its superior length bias mitigation (correlation of 0.004), strong isotropy properties (3.32% first-PC variance), high separability (separation ratio of 0.958), computational efficiency, and standard usage in transformer based embeddings. While Mean Pooling and Late Chunking preserved more semantic detail, their strong length correlations (0.629 to 0.822) introduced systematic biases that could confound temporal analyses.

3.2.2. SANITY CHECK OF THE EMBEDDINGS

To validate that the embeddings capture meaningful semantic relationships, we performed a sanity check using well known movie franchises. We computed the mean embedding vector for movies belonging to three prominent franchises: Star Wars, Harry Potter, and James Bond. We then calculated two types of cosine distances: the inner group distance (between movies within the same franchise) and the group to random distance (between the franchise mean and a random sample of movies from the dataset). This analysis revealed that movies within the same franchise cluster

together in the embedding space, with inner group distances significantly lower than distances to random movies. Specifically, for the Harry Potter franchise, the inner group distance was 0.210 and the group to random distance was 0.572; for Star Wars, these values were 0.349 and 0.560 respectively; and for James Bond movies, they were 0.260 and 0.527. In comparison, the global average cosine distance between two randomly selected movies was 0.520, confirming that the embeddings successfully capture semantic similarity and that movies within the same franchise are indeed closer together in the embedding space than to unrelated movies.

3.2.3. GENRE TAXONOMY CONSTRUCTION

The raw dataset contained 975 unique genre labels, many of which represented semantically equivalent or highly similar categories. For instance, labels such as "science fiction film" and "science fiction comedy" capture essentially the same thematic content despite their different surface forms. To address this redundancy and create a more coherent genre taxonomy, we implemented a multi stage genre organization pipeline. First, we removed all genre labels that appeared only once in the dataset, as these singleton genres do not represent meaningful groupings of movies. This filtering step reduced the number of unique genres from 975 to 463.

Next, we sought to consolidate semantically similar genres by leveraging their textual descriptions. For each remaining genre, we retrieved its corresponding Wikipedia article using the Wikidata QID identifier. Of the 463 genres, 359 had associated Wikipedia articles containing genre descriptions. We then embedded these genre descriptions using the BGE-M3 model, creating vector representations that capture the semantic content of each genre's definition. To identify groups of semantically similar genres, we applied k-means clustering with the Lloyd algorithm to the embedded genre descriptions. We selected $k = 20$ clusters based on the balance between granularity and interpretability, resulting in 20 distinct genre categories. Each cluster was manually labeled by examining the genres it contained and identifying the common thematic elements that unified them. For all further analysis, we will use this taxonomy.

3.3. Novelty analysis

This part in introduction maybe: Common public sentiment is that the film industry is "running out of ideas" resulting in movies that are becoming less creative and more similar to each other over time.

Then later: To investigate the claim that movies are becoming less novel over time, we developed a metric for novelty 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:

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

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 compiute these scores, we use the Faiss library (Douze et al., 2024). Movies were sorted chronologically and processed in yearly batches. For a given batch we queried the Faiss index containing all prior movies to find distances to the nearest neighbor for each movie in the current batch. After that, the current batch was added to the index for subsequent queries.

Results: 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 3.
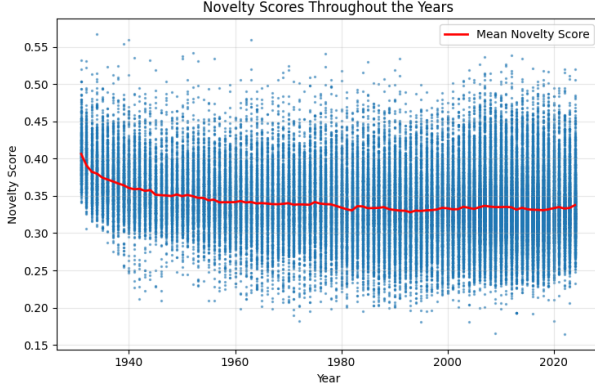

Figure 3. Novelty scores of movies over time. The blue line represents the average novelty score per year, while individual movie scores are shown as scattered points.

The resulting plot indicates that the average yearly novelty remained relatively constant from 1950s onwards. TODO: Find better way to plot this, maybe novelty of all movies vs novelty of Oscar nominees?

3.4. Methodology

After the data was collected and cleaned, the first step was to embed the movie plot summaries into a semantic space.

3.4.1. 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. Such distributions encode structural properties of the embedding space and reveal whether semantic relationships exhibit systematic patterns across temporal periods or thematic categories.

3.4.2. KOLMOGOROV-SMIRNOV TEST

To rigorously compare distance distributions across different subsets of movies, for instance films from different decades or belonging to distinct genres or distinct epsilon balls arround anchor movies, we employ the Kolmogorov-Smirnov test, a non parametric statistical method for assessing whether two empirical distributions arise from the same underlying continuous distribution (Massey, 1951). The two sample KS test compares the empirical cumulative distribution functions (ECDFs) of two samples by computing the maximum vertical distance between them.

Formally, given two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, their empirical cumulative distribution functions are defined as:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x}, \quad G_m(y) = \frac{1}{m}\sum_{j=1}^{m} 1_{Y_j \leq y} \quad (2)$$

where 1 denotes the indicator function. The KS test statistic is defined as the supremum of absolute differences between these ECDFs:

$$D_{n,m} = \sup_x |F_n(x)G_m(x)| \quad (3)$$

Under the null hypothesis that both samples are drawn from the same continuous distribution, the distribution of $D_{n,m}$ is known and can be used to compute p values for hypothesis testing. The test is particularly suited for our application because it makes no assumptions about the underlying distributional form.

A key design choice in applying this framework is the selection of a reference point from which distances are computed. One natural approach is to use the mean vector of a baseline subset of movies as a reference embedding, then compute the distribution of distances from this reference point to all movies in the corpus. This enables quantitative analysis of how semantic representations are spatially organized relative to fixed reference points in the embedding space.

In the context of temporal semantic analysis, the KS test enables systematic comparison of distance distributions across decades. By computing distances from fixed reference points (such as mean embeddings of genre clusters) to movies from different decades, we can assess whether the spatial organization of semantic representations evolves over time. If the semantic structure of cinema remains stable over time, distance distributions should remain statistically similar. Conversely, significant differences in these distributions, as detected by the KS test, would indicate structural

reorganization of the semantic space, suggesting periods where narrative conventions undergo fundamental shifts.

3.4.3. EPSILON BALL CONSTRUCTION

To operationalize this framework, we construct epsilon balls around selected anchor movies by collecting all movies within a specified cosine distance threshold, typically $\epsilon \in [0.24, 0.30]$. Given a set of anchor movies representing a specific thematic category (e.g., spy films), all movies within the epsilon ball exhibit high plot similarity to the anchors, effectively defining a local semantic neighborhood. By comparing the distance distributions of movies within this epsilon ball to those from a control group (constructed using the mean embedding of all movies), we can quantify whether the local semantic structure differs from the global distributional properties of the corpus.

However, distance distributions alone may not fully capture temporal evolution, particularly when anchor movies belong to a series or franchise where sequels naturally cluster together. To address this limitation and analyze the temporal dimension explicitly, we construct cumulative distribution functions (CDFs) of release years for movies within the epsilon ball and compare them to the corresponding CDFs from the control group. A temporal shift in movie plots manifests as a divergence between these CDFs, indicating that the semantic neighborhood defined by the anchor movies exhibits a different temporal distribution than expected under a null model of temporal uniformity. While this approach identifies distributional differences, it does not directly test specific causal hypotheses. Interpretation of observed temporal shifts is therefore performed by examining historical context and culturally significant events within the relevant time periods.

3.5. Genre analysis

As movie genres provide a meaningful taxonomy with potential temporal evolution patterns, we examine semantic drift across different time periods. 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 embeddings are computed: the **centroid** (arithmetic mean) $\bar{\mathbf{e}}_g^{(\tau)}$ and the **medoid** (cosine distance minimizer embedding) $\tilde{\mathbf{e}}_g^{(\tau)}$.

With an arbitrary number of years Δt per group, the period index τ is calculated by flooring the movie year to the nearest multiple of Δt:

$$\tau = \left\lfloor \frac{\text{year}}{\Delta t} \right\rfloor \cdot \Delta t \tag{4}$$

We computed the following metrics to analyse the drift dynamics across the groups:

Genre drift and acceleration: drift (Equation 5) measures displacement between representative embeddings of consecutive periods, capturing how a genre's semantic center evolves over time. Acceleration quantifies the change in drift between consecutive periods.

$$\mathbf{d}_g^{(\tau)} = \bar{\mathbf{e}}_g^{(\tau+\Delta t)} - \bar{\mathbf{e}}_g^{(\tau)} \tag{5}$$

Inter genre distance: determines cosine distance between representatives of each pair of genres for each year, enabling pairwise comparison between specific genres.

Due to group size differences between time periods, two alternative normalization approaches have been employed: (1) downsampling, ensuring equal sampling error across groups, and (2) z-score normalization, which accounts for the standard error of the difference between group means:

$$\hat{v}_g^{(\tau)} = \frac{v_g^{(\tau)}}{\sigma_{\text{pooled}} \cdot \sqrt{\frac{1}{n_g^{(\tau)}} + \frac{1}{n_g^{(\tau+\Delta t)}}}} \tag{6}$$

where σ_{pooled} is the pooled within group standard deviation of cosine distances, and $n_g^{(\tau)}$ is the number of movies in genre g at time τ.

None of the genre based analyses yielded statistically significant results, suggesting that genres may be too broad as analytical categories and any underlying patterns are likely obscured by noise. Figure 4 illustrates temporal drift for the three most popular genres, computed over 5-year periods. Each period was downsampled to 145 movies, with 95% confidence intervals estimated via bootstrapping (1,000 samples).
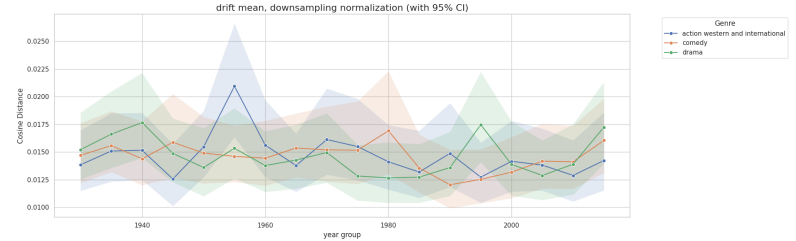

Figure 4. Genre drift from 1930 to 2025 of three most popular genres over 5-year periods

4. Results

In this section, we present and interpret the main empirical findings of our analysis on the embedding space of movie plot summaries. Our results address the spatial structure of the embedding space, overall trends in movie similarity, and how key summary statistics illustrate broader cultural and semantic patterns.

4.1. General Spatial Analysis

We begin with an overview of the global structure of the embedding space by examining the pairwise cosine distances between movie embeddings. Figure 5 displays the distribution of these cosine distances, as well as a fitted normal distribution to summarize their spread. The histogram aggregates results from three independent samples of 5000 movie pairs each, demonstrating that the relationships among movie plots in this high-dimensional space are approximately normally distributed. The mean cosine distance ($\mu = 0.5195$) and standard deviation ($\sigma = 0.0624$) characterize the typical degree of dissimilarity between movie plots. This provides a reference for understanding subsequent analyses—for example, when judging whether certain genres or time-periods are more tightly clustered or more widely dispersed than the dataset as a whole.
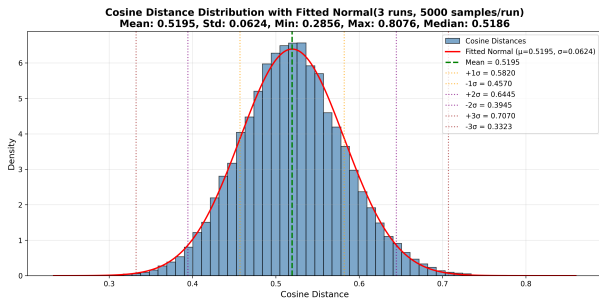

Figure 5. Cosine distance distribution with fitted normal distribution. The distribution shows the pairwise cosine distances between movie embeddings, with mean $\mu = 0.5195$ and standard deviation $\sigma = 0.0624$. The histogram represents cosine distances from 3 runs with 5000 samples each, demonstrating the approximately normal structure of semantic relationships in the embedding space.

To assess the extent to which genre labels correspond to distinct regions in the embedding space, we analyzed separation metrics across 19 genres. The overall intra-genre distance (mean cosine distance between movies within the same genre) was 0.5042, while the overall inter-genre distance (mean cosine distance between movies from different genres) was 0.5268. This yields a separation ratio of 1.0448 and a separation gap of 0.0226. The proximity of these values, with inter-genre distances only marginally exceeding intra-genre distances, indicates substantial overlap between genre clusters in the semantic space. This interpretation is further supported by a silhouette score of -0.0334, where negative values indicate that genres are not well separated and exhibit significant intermingling. These findings suggest that while embeddings capture semantic similarity, genre boundaries in this high dimensional space are relatively porous, reflecting the hybrid and overlapping nature of cinematic categorization.

4.2. Kolmogorov-Smirnov test

We begin by examining a concrete example using James Bond films as anchor movies to assess whether their distance distributions differ from those of randomly selected movies. Figure 6 displays the cumulative distribution function (left panel) and histogram (right panel) of cosine distances from the anchor movies and from the mean embedding vector to all other movies in the dataset. The initial steep rise in the anchor movie CDF corresponds to other Bond films and thematically related spy movies, which exhibit minimal distances. However, these constitute a small fraction of the corpus, resulting in a CDF that remains close to zero initially before rising more gradually. In contrast, the mean vector curve lies consistently to the left of the anchor curve, indicating that the global mean embedding is more similar to the majority of movies than the highly specific Bond anchor movies. This is expected: the mean embedding represents an average over all narrative types, whereas the Bond anchor is semantically constrained to a narrow subgenre, resulting in greater distances to most films. The histogram in the right panel confirms this pattern, showing that the anchor movies exhibit a broader and more right-skewed distribution of distances compared to the mean vector.
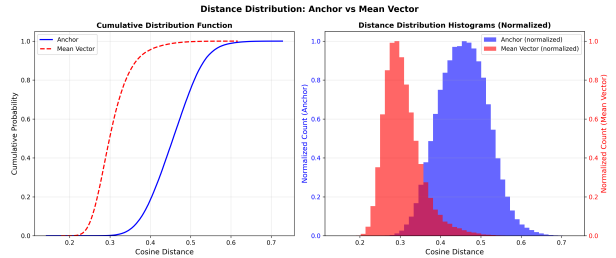

Figure 6. Distance distribution comparison for James Bond anchor movies. Left: Cumulative distribution functions of cosine distances from anchor movies (Bond films) and mean embedding vector to all movies in the dataset. Right: Corresponding histogram showing the distribution of distances. The anchor movies show greater distances to most films due to their thematic specificity.

To examine the temporal dimension, we construct cumulative distribution functions of release years for movies within the epsilon ball and compare them to the control group. Figure 9 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. Both histograms are normalized by their respective maximum yearly counts to facilitate direct comparison of temporal shapes. 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.
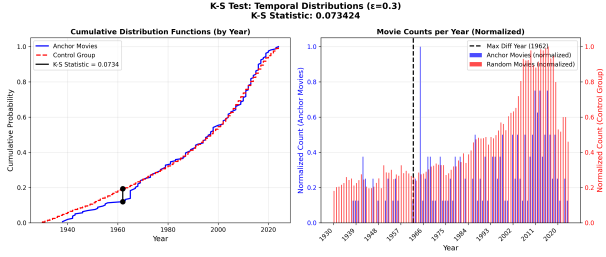

Figure 7. KS test on temporal distributions for James Bond epsilon ball versus control group. Left: Cumulative distribution functions of release years showing divergence beginning in the 1960s. Right: Normalized histograms of movie counts per year, revealing distinct temporal patterns in spy film production compared to the broader corpus.

To further validate the methodology and explore distinct thematic categories, we applied the same framework to movies with a focus on Middle East conflicts, particularly films centered on the Gulf War and American military operations in Iraq and Afghanistan. Using anchor movies such as *Black Hawk Down*, *The Hurt Locker*, *Zero Dark Thirty*, *American Sniper*, *Lone Survivor*, and *13 Hours: The Secret Soldier of Benghazi*, we constructed an epsilon ball with radius $\epsilon = 0.28$. Figure 10 displays the temporal distribution analysis for this thematic category. The temporal shift is even more pronounced than in the spy film case, with the largest divergence occurring prior to the Gulf War period. Following this point, the frequency of movies semantically similar to the anchor movies increases rapidly, and their temporal distribution converges more quickly toward the control group distribution. This pattern suggests that Middle East conflict films represent a temporally concentrated genre that emerged in response to specific historical events, with production rates that accelerated dramatically following the onset of these conflicts.

Spread analysis There were 3 metrics used to analyze the spread of movies each year: (1) Mean L2 norm, (2) Frobenius and (3) Spectral norm of each movie to its yearly centroid as defined in section 3.5.

(1) Mean L2 Norm is the average L2 distance from each movie embedding to its yearly centroid, interpretable as Intra Year Dispersion. L2 norm in our data can vary from 0 to 2, since the embeddings are already normalized to magnitude 1.

(2) Frobenius norm measures total variance (Devroye et al., 2018) of the centered yearly embeddings, since its formulation is $\mathbb{E}[\|X - \mathbb{E}[X]\|^2]$ and can be interpreted as Total
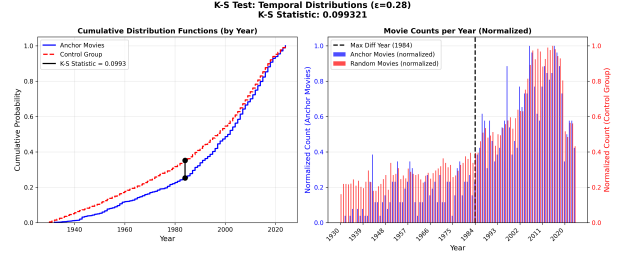

Figure 8. KS test on temporal distributions for Middle East conflict films epsilon ball ($\epsilon = 0.28$) versus control group. The temporal divergence is more pronounced than in the spy film case, with the largest difference occurring before the Gulf War period, followed by rapid convergence as production of conflict-themed films increased.

Yearly Variance.

(3) Spectral norm measures the maximum variance direction of the centered yearly embeddings, since its formulation is $\sup_{\|v\|=1} \|(X - \mathbb{E}[X])v\|^2$ and can be interpreted as the standard deviation of the "Polarizing" axis of the year.
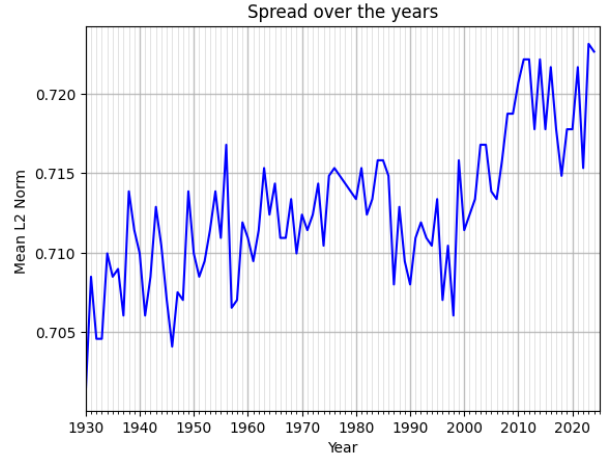

Figure 9. Intra year dispersion

The combination of all 3 metrics gives us a holistic view of how the spread of movies change over the years. From Figure 11, mean is relatively stable but Figure 12 shows both Frobenius and Spectral norm increasing. The increase must be due to some outliers and especially on the "Polarizing" axis. Overall, movies are still very stably spaced out with the exception of a few movies every year. We now analyze which of these movies are on the extreme ends of the polarizing axix (principal component 1).

Table 1 We look at the top 5 and bottom 5 US and Germany movies aligned with the first PC, where we chose arbitrary years. In 1952, the 2 polarizing ends seems to be about action western vs drama comedy, in 2003 it changed
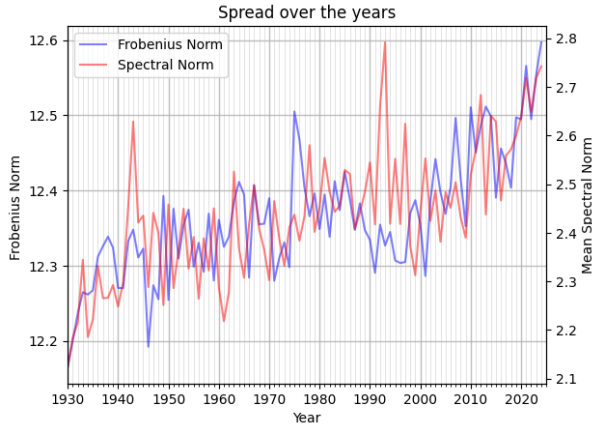
Figure 10. Spectral and Frobenius Norm of yearly movies against yearly centroid

to be about horror/documentaries vs romance and drama, and finally in 2024, it was clear that one end was heavily about horror and thriller while the other end was about documentary/biography.

5. Literature and Methods Review

5.1. Semantic Representation and Cultural Patterns

Our use of word embeddings to represent narrative content draws significant inspiration from the work of *Xu et al. (2019)*. In their study of the "Cinderella Complex" the authors demonstrated that high-dimensional vector spaces can capture latent social biases and emotional dependencies within movie synopses. While *Xu et al.* focused on the internal dynamics of characters and gender stereotypes, our research shifts the lens toward the **inter-textual relationships** between films. By embedding Wikipedia plots, we treat the narrative as a holistic semantic unit, allowing us to map the entire "cinematic universe" into a geometric space where thematic similarity is measured by vector proximity.

5.2. Quantifying Novelty and Innovation

To define and measure "novelty" we adopt the theoretical framework proposed by *Sreenivasan (2013)*, who defines cultural innovation as the emergence of atypical combinations of elements. However, our methodology diverges from Sreenivasan's reliance on keyword frequencies and probabilistic modeling. Instead, we utilize **Geometric Distance Analysis**. In our model, novelty is operationalized as the distance of a movie vector from its k-nearest neighbors in the embedding space. This allows for a more nuanced detection of innovation: a film is considered "novel" if its semantic coordinates lie in a sparse region of the vector space, indicating a narrative structure that is mathematically

distant from established conventions.

5.3. Diachronic Evolution and Genre Trajectories

The temporal dimension of our analysis is informed by the diachronic architectures discussed by *Hamilton et al. (2016)*. Their research into how word meanings shift over time provides a blueprint for our analysis of genre evolution. By calculating the **centroid** of specific genres across different decades, we track their "semantic drift." This approach allows us to test hypotheses regarding cultural homogenization versus diversification, a concept explored in the musical domain by *Di Marco et al. (2025)*. While *Di Marco et al.* utilized network science to observe the simplification of musical structures, we apply these concepts to text, observing whether movie genres are converging toward a standard "formula" or expanding into new, unexplored territories of the embedding world.

6. Discussion & Conclusion

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.

Devroye, L., Mehrabian, A., and Reddad, T. The total variation distance between high-dimensional Gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

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.

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

Matthews, P. and Glitre, K. Genre analysis of movies using a topic model of plot summaries. *Journal of the Association for Information Science and Technology*, 72(12):1511–1527, 2021. doi: 10.1002/asi.24525.

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.

Table 1. Top and Bottom Performing Films by Country and Year

Year	Rank	US Films	German Films
1952	Top 1	Monsoon	Lockende Sterne
	Top 2	Strange Fascination	We're Dancing on the Rainbow
	Top 3	No Time for Flowers	The Colourful Dream
	Top 4	Just Across the Street	Rosen blühen auf dem Heidegrab
	Top 5	Everything I Have Is Yours	Das kann jedem passieren
	Bottom 5	Bend of the River	Toxi
	2	Denver and Rio Grande	The Condemned Village
	3	Red Skies of MontanaH	My Name is Niki
	4	The Savage	The Thief of Bagdad
	5	The Cimarron Kid	All Clues Lead to Berlin
2003	Top 1	Aileen: Life and Death of a Serial Killer	Beyond the Limits
	Top 2	Whole	Nikos the Impaler
	Top 3	Ghosts of the Abyss	Baltic Storm
	Top 4	DC 9/11: Time of Crisis	Debris documentar
	Top 5	Marion's Triumph	Wrong Turn
	Bottom 5	Sinbad: Legend of the Seven Seas	The Suit
	2	The One	Noi the Albino
	3	The Jungle Book 2	The Story of the Weeping Camel
	4	Cosmopolitan	A Little Bit of Freedom
	5	Flavors	Spring, Summer, Fall, Winter... and Spring
2024	Top 1	Terrifier 3	The Devil's Bath
	Top 2	It's What's Inside	Bird
	Top 3	Beetlejuice Beetlejuice	A Sacrifice
	Top 4	Smile 2	Cuckoo
	Top 5	The Thundermans Return	Santosh
	Bottom 5	The Firing Squad	Every You Every Me
	2	The True Story of Tamara de Lempicka and The Art of Survival	Spy vs. Spy
	3	Harvest	The Empire
	4	Waltzing with Brando	Rabia
	5	Putin	Harvest