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

We analyze semantic evolution in cinema by embedding movie plot summaries from 1930 to 2024 into a unified semantic space. Using distance distributions, novelty scores, and statistical tests, we quantify how genres and thematic clusters shift over time. Our analysis reveals periods of semantic stability and reorganization, providing quantitative measures of cultural change in narrative structures across nearly a century of cinema.

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 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 narrative evolve over time.

[*]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>.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201).

2. Data and Methods

2.1. 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.

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. 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, erotic film, and related genres. These categories were removed in order to focus our analysis on mainstream cinematic narratives 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.

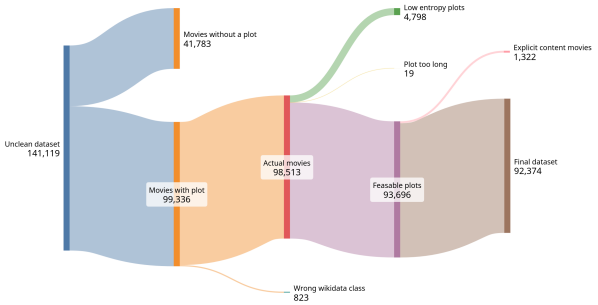

Figure 1. Data cleaning pipeline: number of movies retained after each filtering step.

2.2.1. 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].

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

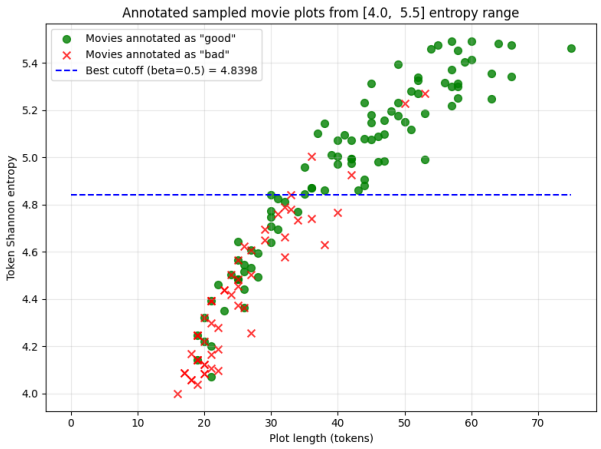

Figure 2. Results of manual labeling of 150 plots in the borderline entropy region with the chosen entropy threshold

BGE-M3 is compact (0.5B parameters), supports 8,192-token context length, and can embed entire movie plots in a single vector.

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 minimal. We evaluated several document embedding methods (chunking and pooling) and metrics as described in (Devlin et al., 2018; Gong et al., 2019; Raffel et al., 2023), and selected CLS Token as our approach.

2.2.2. 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.

2.2.3. 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.

2.3. 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.

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?

2.4. Methodology

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

2.4.1. 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

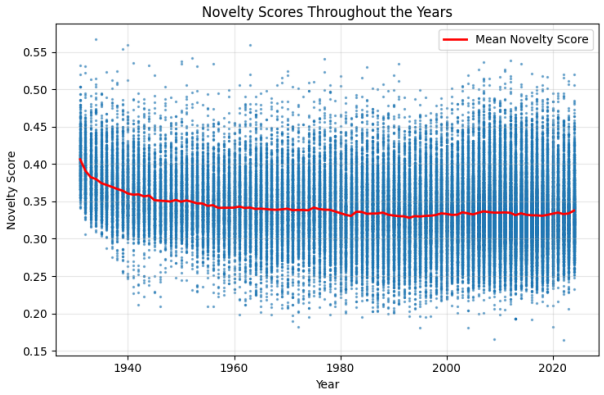

Figure 3. Novelty scores of movies over time. The blue line represents the average novelty score per year, while individual movie scores are shown as scattered points.

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

Genre drift and acceleration: drift (Equation 2) 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.

$$\mathbf{d}_g^{(\tau)} = \bar{\mathbf{e}}_g^{(\tau+\Delta t)} - \bar{\mathbf{e}}_g^{(\tau)} \tag{2}$$

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:

$$\hat{v}_g^{(\tau)} = \frac{v_g^{(\tau)}}{\sigma_{\text{pooled}} \cdot \sqrt{\frac{1}{n_g^{(\tau)}} + \frac{1}{n_g^{(\tau+\Delta t)}}}} \tag{3}$$

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.

2.4.2. Kolmogorov-Smirnov test

To rigorously compare distance distributions across different subsets of movies, for instance films from different decades or belonging to distinct genres, we employ the Kolmogorov-Smirnov test (Massey, 1951), a non parametric statistical method for assessing whether two empirical distributions arise from the same underlying continuous distribution.

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. 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 anchor movies representing a specific thematic category (e.g., spy films), all movies within the epsilon ball exhibit high plot similarity to the anchors, defining a local semantic neighborhood. We compare the distance distributions of movies within this epsilon ball to those from a control group (constructed using the mean embedding of all movies) to quantify whether the local semantic structure differs from the global distributional properties. To analyze temporal evolution 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 exhibits a different temporal distribution than expected under temporal uniformity. Interpretation of observed temporal shifts is performed by examining historical context and culturally significant events within the relevant time periods.

3. Results

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

3.1. 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. The distribution of these distances is approximately normal, with a mean cosine distance

of $\mu = 0.5195$ and a standard deviation of $\sigma = 0.0624$, measured over multiple samples. This summarizes the typical dissimilarity between movie plots and serves as a reference point for subsequent analyses.

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.

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. After normalization, the cosine distance of each time group epresentative embedding with respect to the previous one remained stochastic. Inter genre analysis yielded the same result.

3.2. 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. All 3 metrics were computed on centered yearly embeddings (i.e. yearly centroid was subtracted from each movie embedding before computing the norms) (Yamagiwa & Shimodaira, 2024). We could not interpret much from the results of this analysis of movie spread over the years. Mean L2 norm and Frobenius norm stayed relatively constant at 0.7 and 12.4 respectively. Spectral norm had a slight increase from 2.1 to 2.8. We interpreted this as the overall spread of movies remaining relatively constant over the years with outliers becoming more polarizing. It was hard to determine what these polarizing axis were, as they changed yearly and were a combination of multiple dimensions.

3.3. 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 construct cumulative distribution functions of release years for movies within the epsilon ball and compare them to the control group. Figure 4 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.

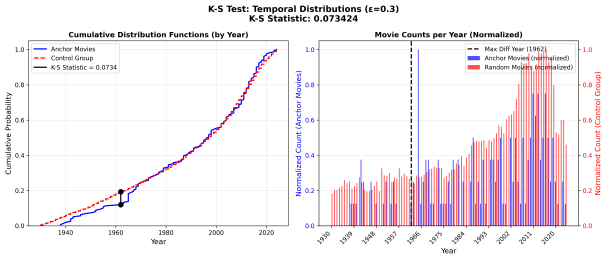

Figure 4. KS test on temporal distributions for James Bond epsilon ball versus control group. Left: Cumulative distribution functions of release years showing divergence beginning in the 1960s. Right: Normalized histograms of movie counts per year, revealing distinct temporal patterns in spy film production compared to the broader corpus.

To further validate the methodology, we applied the same framework to movies focused on Middle East conflicts, using anchor movies such as *Black Hawk Down*, *The Hurt Locker*, *Zero Dark Thirty*, and *American Sniper*. Figure 5 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. This pattern suggests that Middle East conflict films represent a temporally concentrated genre that emerged in response to specific historical events.

4. 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

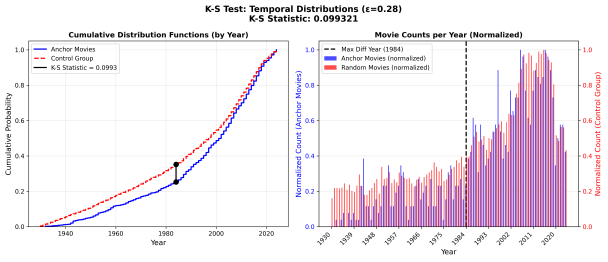

Figure 5. KS test on temporal distributions for Middle East conflict films epsilon ball ($\epsilon = 0.28$) versus control group. The temporal divergence is more pronounced than in the spy film case, with the largest difference occurring before the Gulf War period, followed by rapid convergence as production of conflict-themed films increased.

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.

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.

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.