Data Literacy 2025 Project Report

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

Abstract

This project addresses the fundamental question of how cultural meaning evolves over time by quantitatively modeling seventy-five years of cinematic history through 200,000 film synopses embedded in a single static semantic space using the BGE-M3 model. Temporal change is measured by tracking the movement of genre centroids within this space—analyzing their velocity, acceleration, and curvature to distinguish continuous evolution from structural paradigm shifts. The framework provides a reproducible and data-driven foundation for cultural analytics, testing whether established linguistic laws of semantic drift extend to the domain of cinema.

1. Introduction

Motivate the problem, situation or topic you decided to work on. Describe why it matters (is it of societal, economic, scientific value?). Outline the rest of the paper (use references, e.g. to Section 2: What kind of data you are working with, how you analyse it, and what kind of conclusion you reached. The point of the introduction is to make the reader want to read the rest of the paper.)

2. Data and Methods

Data collection process, bla bla bla, the pipline explained, etc.

After the data was collected in a tabular format, the plots of the data needed to be embedded

Table 1. Per-decade feature coverage (%) of key metadata fields in the movie dataset.

Decade	Actors+Director	Genre	Plot	Vote Count
1950s	86.56	63.67	83.03	90.19
1960s	83.82	61.01	77.31	85.26
1970s	86.55	62.47	79.57	86.58
1980s	84.68	59.41	79.80	85.41
1990s	82.22	58.33	82.00	84.49
2000s	77.34	60.51	84.19	83.73
2010s	70.25	60.55	84.94	85.60
2020s	70.66	64.85	77.59	89.63
Average	80.26	61.35	81.05	86.36

Table 2. Summary of results for each chunking method and embedding configuration. The best value in each column is shown in bold.

Method	Length normcorr	Isotropy 1st PC	Isotropy abtt2	Mean Between Dist.
MeanPooling	0.643	12.23	**3.43**	**0.59**
CLS Token	-0.01	5.14	2.22	0.57
ChunkFirst E. 512/256	-0.38	5.83	2.26	0.56
ChunkFirst E. 1024/512	-0.32	5.28	2.22	0.57
ChunkFirst E. 2048/1024	-0.09	5.14	2.22	0.57
LateChunking 512/256	**0.84**	**12.58**	**3.43**	**0.59**
LateChunking 1024/512	0.75	12.32	**3.43**	**0.59**
LateChunking 2048/1024	0.66	12.23	**3.43**	**0.59**
LateChunking 2048/512	0.66	12.23	**3.43**	**0.59**
LateChunking 512/0	**0.84**	12.24	3.42	**0.59**
LateChunking 1024/0	0.75	12.16	3.42	**0.59**
LateChunking 2048/0	0.66	12.23	**3.43**	**0.59**

3. Results

In this section outline your results. At this point, you are just stating the outcome of your analysis. You can highlight important aspects ("we observe a significantly higher value of x over y"), but leave interpretation and opinion to the next section. This section absoultely *must* include at least

[*]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). Style template based on the ICML style files 2025.

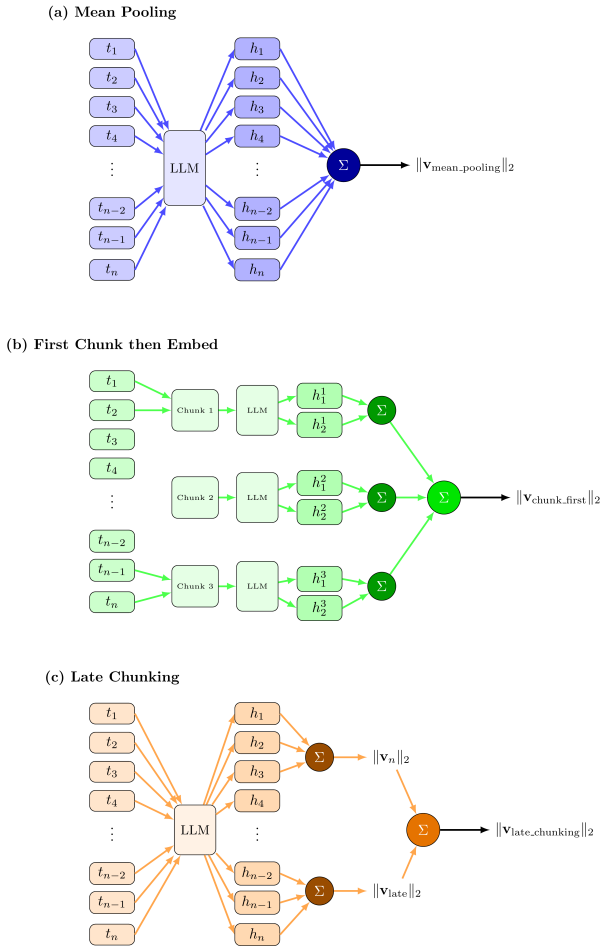

Figure 1. Comparison of different chunking methods: (a) Mean Pooling, (b) First Chunk then Embed, and (c) Late Chunking.

two figures.

4. Discussion & Conclusion

Use this section to briefly summarize the entire text. Highlight limitations and problems, but also make clear statements where they are possible and supported by the analysis.

Contribution Statement

Explain here, in one sentence per person, what each group member contributed. For example, you could write: Max Mustermann collected and prepared data. Gabi Musterfrau and John Doe performed the data analysis. Jane Doe produced visualizations. All authors will jointly wrote the text of the report. Note that you, as a group, a collectively responsible for the report. Your contributions should be roughly equal in amount and difficulty.

Notes

Your entire report has a **hard page limit of 4 pages** excluding references and the contribution statement. (I.e. any pages beyond page 4 must only contain the contribution statement and references). Appendices are *not* possible. But you can put additional material, like interactive visualizations or videos, on a githunb repo (use links in your pdf to refer to them). Each report has to contain **at least three plots or visualizations**, and **cite at least two references**. More details about how to prepare the report, inclucing how to produce plots, cite correctly, and how to ideally structure your github repo, will be discussed in the lecture, where a rubric for the evaluation will also be provided.