Science and Technology in Science Fiction: A Pre- and Post-WWII Comparison through Word Embedding and Mixed-Method Topic Modelling

Keywords: Science Fiction; Word Embedding; Mixed-Method Topic Modelling; Historical Semantic Shift; Computational Social Science

Extended Abstract

Science Fiction literature (SciFi) research has long focused on the thematic shifts of SciFi works over time[1], but there are still two gaps that remain unresolved. First, qualitative SciFi studies often concentrate on individual works from specific authors or periods, lacking large-scale analysis across extended timeframes, which may overlook marginalized works. Second, there has been no systematic examination of how SciFi imagines and represents science and technology. This study aims to address both issues by comparing SciFi works from before and after World War II (WWII), employing a mixed-methods approach with natural language processing. To avoid bias caused by differences in the length of works, this study uses book summaries with stable lengths for analysis. Given the development of large language models (LLMs), obtaining high-quality book summaries will likely become easier in the future, making this methodological approach valuable. The study aims to identify differences in the representation of science and technology in SciFi works before and after WWII, which may reflect the influence of the arms race and Cold War ideologies on broader technological culture[2].

The dataset ($N = 852$) was sourced from the Gutenberg Project, which aims to collect public-domain works regardless of author, genre, prominence, or ideology. For a large portion of the catalog, the Gutenberg Project provides automatically generated summaries, along with author information and original publication details. We developed a Python web scraper, using "Science Fiction" as the search term. For books missing original publication dates, our script retrieved the corresponding Wikipedia page and applied regular expressions to extract the first known publication year. The dataset was restricted to English-language SciFi novels.

We first identified key thematic dimensions in SciFi literature. Given that purely quantitative topic modeling may lack semantic interpretability, we adopted a mixed-methods approach[3]. We began with inductive content analysis on summaries from 80 randomly selected books, segmenting analysis units by minimum semantic coherence (typically 1-3 sentences per unit). Based on the emergent themes and candidate attribute words, we applied Correlation Explanation (CorEx) topic modeling, which imposes fewer structural assumptions and allows for anchor words[4]. We iteratively tested different seed attribute word combinations and assessed semantic coherence within the original context to determine whether to retain specific themes. For the next part of analysis, we focused on paired themes that represent conceptual opposites as analytical dimensions. This yielded two thematic dimensions with four distinct themes: Individual Story (Pos) vs. Macro Narrative (Neg) and External Exploration (Pos) vs. Internal Conflict (Neg), each associated with a set of attribute words.

We divided the preprocessed dataset into two parts using 1949 as the boundary ($N_{\text{pre}} = 389$, $N_{\text{post}} = 463$). For each subset, the Word2Vec technique was employed to generate word embedding vector spaces, containing 4149 and 4590 words, respectively. Given the relatively

small dataset, the dimensionality of word embeddings was set to 50, the context window size to 3, and the number of negative samples to 5. To ensure comparability between the two spaces, we normalized their basis vectors. For a vector $\mathbf{v}$ in the word embedding space Ω, its normalized form is defined as:

$$\mathbf{v}_{\text{normalized}} = \frac{\mathbf{v} - \text{mean}(\mathbf{v}_i)}{\text{std}(\mathbf{v}_i)} \mid_{\mathbf{v}_i \in \Omega}$$

For each target word ("science", "technology", "scientific", "technological" and "technical"), we computed its similarity s with the four topics in both word embedding spaces. The similarity between a target word and a topic is defined as the average cosine similarity $\cos(\text{target}, \text{attribute}_i)$ between the target word and all attribute words within that topic—the higher the value, the closer the target word is to the given topic. The position l of the target word along a dimension is defined as the difference between its similarity to the positive and negative topics:

$$\begin{aligned} l_{\mathbf{v}} &= s(\mathbf{v}, \text{topic}_{\text{positive}}) - s(\mathbf{v}, \text{topic}_{\text{negative}}) \\ &= \text{mean}(\cos(\mathbf{v}, \text{attribute}_{\text{positive},i})) - \text{mean}(\cos(\mathbf{v}, \text{attribute}_{\text{negative},j})) \end{aligned}$$

We compared whether the position of each target word along the two dimensions changes between the two subsets, defined as effect $= l_{\text{post}} - l_{\text{pre}}$. We also calculated the effect size by computing the z-score of the target word's l relative to the distribution of l across all words in the word embedding space. Finally, we computed the p-value of the effect by estimating the probability of randomly selecting a word from the word embedding space with a more extreme position than the target word[5].

In Table A1, the results for five target words are reported separately, along with their average effect. All effects are negative, and the average effect is significant at the 0.1 level. This indicates that post-WWII SciFi represents science and technology more in the context of macro narratives (such as civilizations, politics, and authority) and internal conflicts (such as war, conflict, and military). In contrast, pre-WWII SciFi constructs science and technology more in relation to individual stories and emotions (such as adventure, longing, and bravery), as well as external exploration (such as discovery, creatures, and the cosmos). We repeatedly adjusted the model's hyperparameters, attribute word combinations, and dataset partitioning rules. Additionally, for books with missing original publication years, we imputed the data using the average of the author's birth and death years to expand the sample size. The results remained robust.

This study makes contributions both methodologically and theoretically. Methodologically, it demonstrates that topic modeling combined with word embedding techniques is an effective approach for extracting information from book summaries and comparing semantic differences across periods or genres, allowing for a comprehensive examination of large-scale literary works with relatively low computational and time costs while mitigating biases caused by variations in text length or authors' productivity[6]. Future research could leverage fine-tuned or self-trained LLMs to generate book summaries, enabling more precise and targeted information extraction. Theoretically, this study provides quantitative evidence on how representations of science and technology evolve over time, illustrating that SciFi is not merely a product of imagination but also a reflection of social and political contexts. Specifically, we find that post-WWII works embed science and technology within narratives of politics and warfare, whereas earlier works emphasize individual heroism and cosmic exploration. This insight offers a heuristic analytical framework for the study of SciFi.

References

[1] Seed, D. (2013). *Science Fiction: A Very Short Introduction*. Oxford Academic. https://doi.org/10.1093/actrade/9780199557455.001.0001, accessed 11 Feb.

[2] Gerlach, N., & Hamilton, S. N. (2003). *Introduction: a history of social science fiction*. Science Fiction Studies, 161-173.

[3] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). *Reading tea leaves: How humans interpret topic models*. Advances in neural information processing systems, 22.

[4] Gallagher, R. J., Reing, K., Kale, D., & Ver Steeg, G. (2017). *Anchored correlation explanation: Topic modeling with minimal domain knowledge*. Transactions of the Association for Computational Linguistics, 5, 529-542.

[5] Yang, E., & Roberts, M. E. (2021, March). *Censorship of online encyclopedias: Implications for NLP models*. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 537-548).

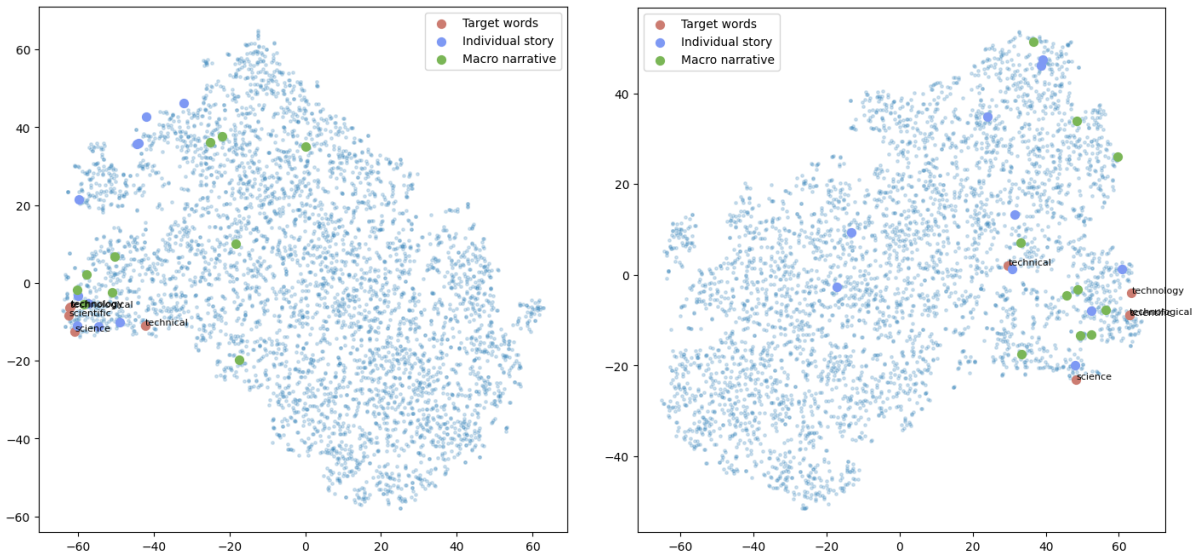
[6] Gerlach, M., & Font-Clos, F. (2020). *A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics*. Entropy, 22(1), 126.

Appendix

Table 1: Pre- and Post-WWII Comparison of Target Words

	Individual vs. Macro			Exploration vs. Conflicts		
	effect	effect size	p-value	effect	effect size	p-value
Science	-0.568	-1.859	0.041**	-0.398	-1.593	0.059*
Technology	-0.240	-0.875	0.205	-0.321	-1.290	0.119
Scientific	-0.521	-1.718	-0.058*	-0.460	-1.831	0.026**
Technological	-0.331	-1.149	0.137	-0.432	-1.724	0.038**
Technical	-0.689	-2.221	0.023**	-0.480	-1.908	0.016**
Overall	-0.470	-1.565	0.076*	-0.418	-1.669	0.048**

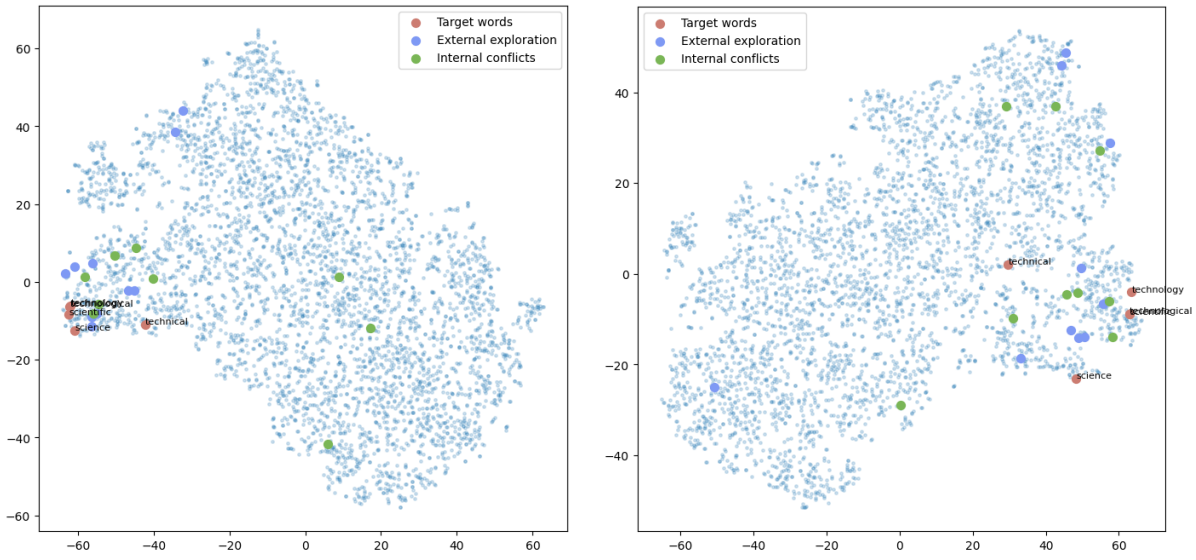
Figure 1: Comparison of Word Embedding Spaces (Individual vs. Macro) using t-SNE



(a) Pre-WWII

(b) Post-WWII

Figure 2: Comparison of Word Embedding Spaces (Exploration vs. Conflicts) using t-SNE



(a) Pre-WWII

(b) Post-WWII