On the study of the shape of language: A Topological Analysis of Universal Text Encoder Embedding Spaces

Editors: List of editors' names

Abstract

Encoders have become fundamental to the development of Natural Language Processing tasks and other machine learning domains. However, the incompatibility of their generated embedding representation spaces constitutes a significant challenge for their productive use in real world applications. Methodologies such as *vec2vec* are highly important in this context, raising questions about the possibility of generating a universal embedding representation space. In this article, persistence homology is employed to study the common representation spaces generated by *vec2vec* from various text encoders, combining models with identical and different backbones across multiple datasets. Our methodology utilizes topological data analysis tools, such as persistence diagrams and persistence landscapes, along with distance metrics such as the Wasserstein distance or the l_2 norm to objectively evaluate the similarity of the generated representations. Our findings indicate that while the translated space produced by *vec2vec* is effective in terms of cosine similarity, the translation process is shown to introduce topological artifacts. Besides, a statistically significant correlation was not found between geometric and topological metrics. Furthermore, no statistical evidence was observed to suggest that a common backbone leads to a more robust preservation of topological features. Finally, it was determined that a common representation space with similar topological characteristics is generated by *vec2vec* across different encoders, although the degree of similarity is shown to be dependent on the specific dataset used for training.

Keywords: text encoders, persistent homology, persistence diagram, persistence landscape, Wasserstein distance, l_2 norm

1. Introduction

The Transformer architecture, introduced by Vaswani et al. (2017), and specifically its encoder component, has sparked a real revolution in machine learning solutions. Despite the recent surge and significant momentum of generative models, encoders remain responsible for major advancements in classical supervised and unsupervised learning tasks such as classification, clustering, and anomaly detection. In the context of Natural Language Processing (NLP), a multitude of tasks—including text classification [da Costa et al. (2023)], sentiment analysis [Das and Singh (2023)], entailment [Alharahseheh et al. (2022)], or named entity recognition [Ji et al. (2025)]. among others—have achieved state-of-the-art results thanks to these encoders.

However, the inherent nature of neural network architectures and their training processes results in each encoder generating vector representations, or embeddings, that are incompatible with those from other models. This incompatibility creates significant challenges in both academia and industry. Recently, Jha et al. (2025) introduced *vec2vec*, a methodology designed to translate any embedding into a universal latent space in an unsupervised manner. This approach is grounded in the Platonic Representation Hypothesis

[Huh et al. (2024)]. In fact, the authors propose a reinforced version of this premise, termed the Strong Platonic Representation Hypothesis, which establishes that any neural network trained on a common data modality converges to a shared, universal latent structure. Their experimentation demonstrates that for a given text, representations generated by different encoders –which may be notably distant in their original spaces– become aligned within this common latent space. Nonetheless, beyond the local fidelity assessed by cosine similarity, a truly universal latent structure would suggest the existence of isomorphisms between the different embedding manifolds. To robustly evaluate the Strong Platonic Representation Hypothesis, it is essential to employ tools capable of describing and comparing the intrinsic shape of the data.

For this reason, in this work, Topological Data Analysis (TDA) [Wasserman (2018)] is used to gain a deeper understanding of the nature of this universal latent space. These properties will be analyzed using persistent homology [Edelsbrunner et al. (2008)], a foundational TDA technique that extracts and summarizes the topological structure of data across different spatial scales simultaneously.

We propose a well-established methodology to systematically characterize the topology of the embedding spaces generated by different models and datasets, addressing the following research questions:

- **RQ1:** Does the *vec2vec* translation process introduce characteristic topological artifacts?
- **RQ2:** Does topological fidelity correlate with geometric fidelity?
- **RQ3:** Is topological preservation easier for closer model pairs?
- **RQ4:** Is the unification of topological features of different source models achieved via the latent space?

The rest of the paper is organized as follows: Section 2 presents an overview of representation alignment methods and topology-based approaches for analyzing data structures. Our proposal is described in Section 3. Section 4 presents the experimentation that validates the methodology. Subsequently, Section 5 discusses the results. Finally, Section 6 outlines the study's main conclusions.

2. Background

The following section reviews different approaches from the scientific literature for the construction of latent spaces in which embeddings generated by encoder models are aligned.

2.1. Representation, semantic and alignment

The study of how neural networks represent acquired knowledge and concepts is a profound field of research. One of the seminal contributions, Singular Vector Canonical Correlation Analysis (SVCCA), was proposed by Raghu et al. (2017). This method focuses on comparing different representations in a manner that is invariant to affine transformations and computationally efficient.

Morcos et al. (2018) introduced an enhancement to SVCCA, known as Canonical Correlation Analysis (CCA). This technique enables the differentiation between meaningful signals and noise, thereby demonstrating that networks which generalize converge to more similar representations than those that merely memorize the training data. The authors also observed that wider networks converge to more similar solutions than narrower ones, and that networks with identical topologies but trained with different learning rates converge to distinct clusters with diverse representations.

With repect to multimodal models, Norelli et al. (2023) proposed a method to create a common space between pre-trained vision models and text encoders without additional training, utilizing single-domain encoders and a small set of image-text pairs. Similarly, in the context of tasks involving both images and text, such as image captioning and the CLIP model [Radford et al. (2021)], Moayeri et al. (2023) introduced the text-to-concept method. This approach aligns features from a fixed pre-trained model into the CLIP latent space via a linear transformation.

From a more general perspective, Yamagiwa et al. (2023) employed Independent Component Analysis (ICA) to uncover inherent semantic structures within word or image embeddings. Their findings suggest that each embedding can be expressed as a composition of a few interpretable axes, and that the semantics defined by these axes remain consistent across different languages, algorithms, and modalities.

2.2. TDA applied to the shape of language

Over the last few years, numerous initiatives have been published concerning the use of Topological Data Analysis (TDA) for the analysis of language models. These studies range from examining how fine-tuning affects the semantic representation of word embeddings [Rathore et al. (2023)] to a topology-based modification of the attention layer [Perez and Reinauer (2022)], and even the application of TDA for model interpretability, simplification, and optimization [Uchendu and Le (2025); Balderas et al. (2025)]. In any case, TDA is increasingly becoming a robust and highly useful tool for improving tasks in natural language processing.

3. Our proposal

This paper presents a methodology based on homology theory for a rigorous and global evaluation of the Strong Platonic Representation Hypothesis. This hypothesis claims not only that certain embeddings can be correctly mapped, but also that the topological structure of the embedding space generated by the source model is preserved through the translations generated by the *vec2vec* method.

Tools such as persistence diagrams (PD), which plot the birth and death times of n-dimensional topological features; or persistence landscapes (PL), which provide a transformation of the discrete information of persistence diagrams into a more functional representation; and specific distance metrics will be applied to measure the differences between the various spaces generated by the encoders. Fundamentally, the objective is to characterize the intrinsic topology of the embedding manifolds. The methodology to be followed is outlined in Algorithm 1.

Algorithm 1: From text and embeddings to topology

1. **Encoder and Translator Training:** The encoders and the translator are trained following the *vec2vec* framework. The procedure adheres to the methodology of the original paper to ensure the fidelity of the results.
2. **Persistent Homology Computation:** Persistent homology is computed for the embeddings of the source models, the latent space, and the translated outputs.
3. **Application of Topological Tools:** TDA tools are utilized for the study of the topological features.
4. **Metric Evaluation:** Distance metrics are evaluated on the resulting persistence diagrams and persistence landscapes to address the posed research questions.

4. Evaluation

For a proper evaluation of the methodology and to answer the proposed research questions, a thorough experimentation has been conducted, which includes different models, datasets, and metrics. This is described in detail below.

4.1. Models and Datasets

Following the methodology of the original paper, the Natural Questions (NQ) [Kwiatkowski et al. (2019)] and Fineweb Tiny [Penedo et al. (2023)] datasets were employed for training (using the training set) and the subsequent measurement of topological features (using the validation set). Regarding the models, the experimentation focuses on text encoders with different backbones, thereby enabling the comparison of relatively similar models. Specifically, the models employed are gtr [Ni et al. (2021)] (T5 backbone); gte [Li et al. (2023)], e5 [Wang et al. (2024)], and stella [Zhang et al. (2025)] (BERT backbone); and granite [Granite Embedding Team (2024)] (RoBERTa backbone).

4.2. Metrics

Regarding the metrics, distances are employed to measure the similarity of the manifolds generated by each encoder using the available topological tools. In particular, to compare persistence diagrams, the $2-$Wasserstein distance [Villani (2009)], is utilized, with calculations performed separately for each dimension. In general, for $p \in [1, \infty)$ and Borel probability measures P, Q on $\mathbb{R}^n$ with finite $p-$moments, their $p-$Wasserstein distance is defined as [Ramdas et al. (2015)]:

$$W_p(P, Q) = \left(\inf_{\pi \in \Gamma(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} ||X - Y||^p d\pi \right)^{1/p} \tag{1}$$

where $\Gamma(P, Q)$ is the set of all joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are P, Q. In addition to persistence diagrams, persistence landscapes are also employed. For the comparison of these landscapes, the l_2 norm is utilized. This norm is defined as:

$$\|x, y\|_2 = \left(\sum_{i=1}^{n} |x - y|^2 \right)^{1/2} \tag{2}$$

This provides a robust method for quantifying the similarity or dissimilary between two persistence landscapes and, by extension, between the underlying topological spaces they represent.

4.3. Experiments and Results

To address the proposed research questions, a series of experiments were developed. These experiments are designed to evaluate the topological proximity between the translation generated by the *vec2vec* framework and the target space.

Specifically, in the first experiment, all pairs of encoders, denoted as M_1 and M_2, were trained using the datasets mentioned in the preceding section. Consequently, if $F(M_1)$ represents the translation of the space generated by encoder M_1, and M_2 generates the target space, the evaluation was conducted as follows: The distance between their respective persistence diagrams was measured using the Wasserstein distance in dimensions 0 and 1 (as shown in Table 1). Additionally, the distance between the corresponding persistence landscapes was calculated using the l_2 norm for persistent homology in dimensions 0, 1, and 2 (Table 2).

Table 1: Experimental results for the analysis of the topological structure. Cos Sim (L/I) refers to the cosine similarity between the embeddings in the latent space (L) and the original inputs (I). Higher is better. The term 2-$W(M_1, M_2)$ denotes the second order Wasserstein distance between the persistence diagrams generated from the output embeddings of models M_1 and M_2, evaluated on the corresponding dataset. Similarly, 2-$W(F(M_1), M_2)$ represents the same distance but calculated between the persistence diagrams of the translated embedding and the target space. In both cases, lower is beter. As shown, these distances are applied to the 0-dim and 1-dim persistence diagrams.

$M_1 \to M_2$	Dataset	Cos Sim (L/I)	2-$W(M_1, M_2)$ (0d)	2-$W(M_1, M_2)$ (1d)	2-$W(F(M_1), M_2)$ (0d)	2-$W(F(M_1), M_2)$ (1d)
gte → gtr	NQ	0.82/0.03	0.07	**0.41**	**0.06**	1.65
	Fineweb Tiny	0.52/0.04	**0.06**	**0.35**	0.08	1.01
gte → e5	NQ	0.86/0.68	**0.05**	**0.25**	0.08	1.07
	Fineweb Tiny	0.86/0.67	0.08	**0.29**	**0.05**	1.24
gte → stella	NQ	0.93/0.56	**0.06**	**0.38**	0.1	0.99
	Fineweb Tiny	0.92/0.57	0.09	0.4	0.05	0.82
gte → granite	NQ	0.7/0.01	0.11	0.72	**0.08**	**0.56**
	Fineweb Tiny	0.64/0.01	0.05	**0.47**	0.05	0.64
stella → granite	NQ	0.61/0.007	0.08	**0.42**	**0.03**	1.03
	Fineweb Tiny	0.55/0.007	0.05	**0.32**	0.05	1
gtr → granite	NQ	0.37/-0.02	**0.05**	0.49	0.11	**0.32**
	Fineweb Tiny	0.33/-0.03	0.06	0.36	0.06	0.36
gtr → stella	NQ	0.33/0.002	0.1	**0.32**	**0.05**	0.75
	Fineweb Tiny	0.45/0.005	**0.06**	**0.29**	0.09	0.68

These results will allow for answers to be provided for research questions RQ1 and RQ3. Furthermore, by applying Spearman's rank correlation test to the distribution of cosine similarity values and Wasserstein distance values, it will be possible to answer RQ2.

Table 2: Table 1 continued. l_2 norm applied to the persistence landscapes of either the original encoders or the translation from M_1 to M_2 ($F(M_1)$) with respect to the target (M_2). Lower values are better. As shown, these distances are applied to the 0-dim, 1-dim and 2-dim persistence landscapes.

$M_1 \to M_2$	**Dataset**	**Paired models**	l_2 **norm 0-dim PL**	l_2 **norm 1-dim PL**	l_2 **norm 2-dim PL**
gte → gtr	NQ	M_1, M_2	3.96	**2.67**	**1.36**
		$F(M_1), M_2$	**0.89**	3.7	1.39
gte → gtr	Fineweb Tiny	M_1, M_2	**3.29**	**2**	**1.45**
		$F(M_1), M_2$	7.83	5.42	1.83
gte → e5	NQ	M_1, M_2	**2.65**	**3.24**	**1.31**
		$F(M_1), M_2$	8.28	6.07	2.33
gte → e5	Fineweb Tiny	M_1, M_2	**3.62**	**1.87**	**1.49**
		$F(M_1), M_2$	12.07	5.03	2.57
gte → stella	NQ	M_1, M_2	**2.72**	4.55	2.58
		$F(M_1), M_2$	5.71	**3.13**	**1.75**
gte → stella	Fineweb Tiny	M_1, M_2	**0.9**	5.2	1.94
		$F(M_1), M_2$	3.08	**3.97**	**1.88**
gte → granite	NQ	M_1, M_2	13.64	**2.22**	**1.4**
		$F(M_1), M_2$	**1.5**	5.48	2.74
gte → granite	Fineweb Tiny	M_1, M_2	**5.03**	4.03	1.45
		$F(M_1), M_2$	6.1	**3.18**	**1.1**
gtr → granite	NQ	M_1, M_2	9.83	**2.22**	**1.44**
		$F(M_1), M_2$	**0.12**	5.13	2.87
gtr → granite	Fineweb Tiny	M_1, M_2	**5.03**	4.03	1.45
		$F(M_1), M_2$	6.1	**3.18**	**1.1**
gtr → stella	NQ	M_1, M_2	**7.81**	3.2	2.27
		$F(M_1), M_2$	17.26	**2.46**	**1.93**
gtr → stella	Fineweb Tiny	M_1, M_2	**0.04**	**4.07**	**2.01**
		$F(M_1), M_2$	11.07	6.24	2.18
stella → granite	NQ	M_1, M_2	**1.92**	**5.78**	**2.69**
		$F(M_1), M_2$	4.13	7.13	3.05
stella → granite	Fineweb Tiny	M_1, M_2	4.99	**3.95**	**1.73**
		$F(M_1), M_2$	**0.84**	5.08	2.05

Finally, to answer RQ4, the Wasserstein distance between the persistence diagrams of the latent spaces generated by each encoder is calculated to measure their similarity. The results can be found in Table 3.

Table 3: Results regarding the latent space experiment. The embeddings of both models ($M_1 \leftrightarrow M_2$) are constructed in the common latent space, and the second-order Wasserstein distance is applied to the persistence diagram in dimensions 0 and 1.

$M_1 \leftrightarrow M_2$	**Dataset**	**2-W (0-dim)**	**2-W (1-dim)**
gte ↔ gtr	NQ	0.03	0.53
	Fineweb Tiny	0.19	1.21
gte ↔ e5	NQ	0.04	0.39
	Fineweb Tiny	0.06	0.3
gte ↔ stella	NQ	0.06	0.52
	Fineweb Tiny	0.12	0.73
gte ↔ granite	NQ	0.1	0.26
	Fineweb Tiny	0.4	0.26
gtr ↔ granite	NQ	0.07	0.74
	Fineweb Tiny	0.17	0.95
gtr ↔ stella	NQ	0.25	0.98
	Fineweb Tiny	0.07	0.27
stella ↔ granite	NQ	0.08	0.25
	Fineweb Tiny	0.1	0.41

5. Discussion

In this section, the principal results are discussed, and the research questions proposed in this work are answered.

5.1. RQ1: Does *vec2vec* translation process introduce characteristic topological artifacts?

The choice of the neural network architecture in the *vec2vec* framework and the training configuration may introduce inherent characteristics, in the form of bias, into the generated space. Moreover, adversarial training is challenging to stabilize, particularly in the context of Generative Adversarial Networks (GANs). Consequently, novel topological features, such as clusters ($0-$dimension persistent homology), loops ($1-$dimension persistent homology), or voids ($2-$dimension persistent homology) may emerge in the translated space that are absent in the target space.

To verify this phenomenon, in addition to the metrics proposed in Table 1, a specific model pair (gte → gtr) and the NQ dataset were selected. The persistence landscapes for the spaces generated by the original models and for the resulting translated space were then represented graphically (Figures 1, 2 and 3).

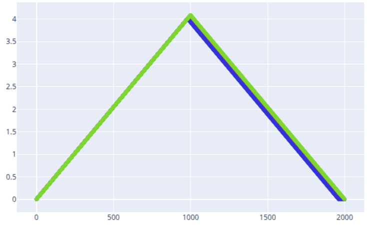

Figure 1: 0-dim PL. gte (green), gtr (orange, hidden by green) and translation from gte to gtr (blue)

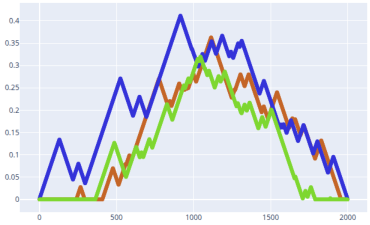

Figure 2: 1-dim PL. gte (green), gtr (orange) and translation from gte to gtr (blue)

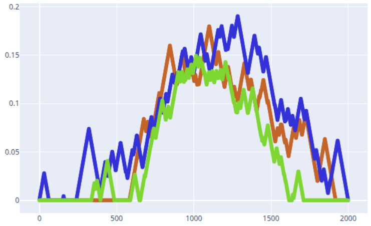

Figure 3: 2-dim PL. gte (green), gtr (orange) and translation from gte to gtr (blue)

As evidenced by Figure 1 and the associated metrics, the dimension 0 persistence landscapes are highly similar. However, for dimensions 1 and 2, it is observed that although the blue curve (representing the translation) resembles the orange curve (representing the target space), prominent new peaks are introduced in the translation. These peaks are absent from the topological feature space generated by gtr, indicating the appearance of artifacts (new loops and voids). This is also evidenced by the fact that the norm of the persistence landscapes for the translation and the target space is even greater than the distance between the target space and the space generated by gte in dimensions 1 and 2 (3.7 vs. 2.67 and 1.39 vs. 1.36, respectively, as shown in Table 2).

5.2. RQ2: Does topological fidelity correlate with geometric fidelity?

Intuitively, it could be assumed that a space translated by *vec2vec* must exhibit high similarity to the target space, not only in terms of cosine similarity but also according to metrics that reflect the preservation of global topology. In other words, a strong correlation should exist between the distribution of cosine similarity values and the Wasserstein distances (in both dimension 0 and dimension 1) generated by the model pairs.

Using the values from Table 1, the Spearman correlation test was applied between these data distributions. In both cases, the resulting $p-$values (0.7 and 0.086, respectively) do not provide sufficient statistical evidence to confirm a correlation, and therefore, such a relationship between the two distributions cannot be asserted.

Furthermore, it can be observed in Table 1 that for certain model pairs, such as gtr → granite, the Wasserstein distance values (dimension 1) are significantly smaller than in other cases (meaning the global shape was preserved well), despite this pair having the lowest cosine similarity (meaning the individual vectors were not well-aligned). In light of these results, it can be concluded that while some relationship may exist between geometric and topological fidelity, there is no statistically significant correlation.

5.3. RQ3: Is topological preservation easier for "closer" model pairs?

The translation generated by *vec2vec* stems from encoders that may have completely different backbones (e.g., gte → gtr, gtr → stella, BERT vs. T5; gtr → granite, T5 vs RoBERTa), similar ones (e.g., gte → granite, stella → granite, BERT vs. RoBERTa), or identical ones (e.g., gte → e5, gte → stella, BERT vs. BERT). It might be expected this translation to be simpler, meaning it would exhibit smaller distance values for both l_2 and Wasserstein metrics across all dimensions of persistent homology compared to a pair of models with different backbones. To test this hypothesis, the obtained Wasserstein distance (2-$W(PD(F(M_1)), PD(M2))$) and l_2 norm values ($l_2(PL(F(M_1)), PL(M_2))$) were divided into two groups: one for same-encoder pairs and one for cross-encoder pairs. This resulted in two distributions for each metric.

A t-test was then applied to these distributions. The null hypothesis (H_0) posited that the mean of the distributions (of the distances) were equal, while the alternative hypothesis (H_A) stated that the means were different.

The test results yielded $p-$values greater than 0.05. Consequently, there was no statistically significant evidence to reject the null hypothesis.

Therefore, from a statistical point of view and based on the results of this experiment, it cannot be asserted that models sharing the same backbone –that is, models with identical architectures– generate a translated space that is topologically more similar to the target space than models with different backbones.

5.4. RQ4: Is the unification of topological features of different source models achieved via the latent space?

Finally, the last research question was posed, concerning the universal latent space that can unify the topology generated by different models. By fixing a set of documents, their representation was generated in a common latent space. The Wasserstein distance of the persistence diagrams of these spaces was measured to compare the presence or absence of topological features at the manifold level. As can be seen in Table 3, the distance between the different diagrams is small. It is apparent that for the NQ dataset, the level of agreement in topological features is greater (0.079 and 0.459 Wasserstein distance for dimension 0 and 1, respectively) than for the Fineweb Tiny dataset (0.139 and 0.516, respectively). However, in both cases, a common space (not universal, as it depends on the datasets and training setup) is achieved. This reinforces the idea conveyed by *vec2vec*'s results regarding cosine similarity.

6. Conclusion

Encoders are fundamental models in the development of artificial intelligence and its applications, having led to significant advancements in NLP tasks. However, the incompatibility of embedding representation spaces makes their deployment and use in production complex for both industry and academia. Methodologies such as the one proposed in *vec2vec* are of great importance and raise questions regarding the possibility of generating a universal embedding representation space, as outlined in the Strong Platonic Representation Hypothesis.

In this work, topological data analysis is employed to study the common representation spaces generated with *vec2vec* from different text encoders, combining models with the same and different backbones and various datasets. To this end, an objective methodology is defined that utilizes topological analysis tools, such as persistence diagrams and persistence landscapes, as well as specific distance metrics to evaluate the similarity of the representations generated by these tools.

The results of this investigation indicate that while the translated space generated by *vec2vec* has been shown to be useful and robust from the perspective of cosine similarity, the translation process introduces topological artifacts, in other words, new topological features that were not present in the original models (RQ1). Furthermore, it is confirmed that there is no statistically significant correlation (RQ2) between geometric metrics (cosine similarity) and topological ones (Wasserstein distance for persistence diagrams and l2 norm for persistence landscapes). It is also observed that there is no statistical evidence that models with the same starting backbone, when *vec2vec* is applied, achieve a more pronounced preservation of topological features than for pairs of encoders with different backbones (RQ3). Finally, the experiments carried out show that *vec2vec* generates a common representation space across different encoders with quite similar topological features, as the distance between their persistence diagrams is small. Nonetheless, it should not be considered as an universal representation, taking to account that achieving a greater or lesser degree of similarity depends on the specific dataset used for training.

Acknowledgments

This research has been partially supported by Proyecto PID2023-151336OB-I00 financiado por MICIU/AEI /10.13039/501100011033 y por FEDER, UE.

References

Yara Alharahseheh, Rasha Obeidat, Mahmoud Al-Ayoub, and Maram Gharaibeh. A survey on textual entailment: Benchmarks, approaches and applications. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 328–336. IEEE, 2022.

Luis Balderas, Miguel Lastra, and José M. Benítez. A green ai methodology based on persistent homology for compressing bert. *Applied Sciences*, 15(1), 2025. ISSN 2076-3417. doi: 10.3390/app15010390. URL https://www.mdpi.com/2076-3417/15/1/390.

Liliane Soares da Costa, Italo L. Oliveira, and Renato Fileto. Text classification using embeddings: a survey. *Knowledge and Information Systems*, 65(7):2761–2803, Jul 2023. ISSN 0219-3116. doi: 10.1007/s10115-023-01856-z. URL https://doi.org/10.1007/s10115-023-01856-z.

Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Comput. Surv.*, 55(13s), July 2023. ISSN 0360-0300. doi: 10.1145/3586075. URL https://doi.org/10.1145/3586075.

Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453(26):257–282, 2008.

IBM Granite Embedding Team. Granite embedding models, December 2024. URL https://github.com/ibm-granite/granite-embedding-models/.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024. URL https://arxiv.org/abs/2405.07987.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings, 2025. URL https://arxiv.org/abs/2505.12540.

Lixia Ji, Yiping Dang, Yunlong Du, Wenzhao Gao, and Han Zhang. Nested named entity recognition: A survey of latest research. *Expert Systems*, 42(7):e70052, 2025.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023. URL https://arxiv.org/abs/2308.03281.

Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment, 2023. URL https://arxiv.org/abs/2305.06386.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation, 2018. URL https://arxiv.org/abs/1806.05759.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers, 2021. URL https://arxiv.org/abs/2112.07899.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training, 2023. URL https://arxiv.org/abs/2210.01738.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. URL https://arxiv.org/abs/2306.01116.

Ilan Perez and Raphael Reinauer. The topological bert: Transforming attention into topology for natural language processing, 2022. URL https://arxiv.org/abs/2206.15195.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, 2017. URL https://arxiv.org/abs/1706.05806.

Aaditya Ramdas, Nicolas Garcia, and Marco Cuturi. On wasserstein two sample testing and related families of nonparametric tests, 2015. URL https://arxiv.org/abs/1509.02237.

Archit Rathore, Yichu Zhou, Vivek Srikumar, and Bei Wang. Topobert: Exploring the topology of fine-tuned word representations. *Information Visualization*, 22(3):186–208, 2023.

Adaku Uchendu and Thai Le. Unveiling topological structures from language: A comprehensive survey of topological data analysis applications in nlp, 2025. URL https://arxiv.org/abs/2411.10298.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Cedric Villani. Optimal transport. old and new. *Springer-Verlag*, 338, 2009.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pretraining, 2024. URL https://arxiv.org/abs/2212.03533.

Larry Wasserman. Topological data analysis. *Annual review of statistics and its application*, 5(2018):501–532, 2018.

Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. Discovering universal geometry in embeddings with ica, 2023. URL https://arxiv.org/abs/2305.13175.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models, 2025. URL https://arxiv.org/abs/2412.19048.

Appendix A. *vec2vec* framework

vec2vec [Jha et al. (2025)] is an unsupervised method for translating text embeddings from one vector space to another. It does not require paired data or known encoders. Its primary purpose is to enable information extraction from vector sets in contexts where one only has access to the embeddings, without the original text or the model that generated them.

The essence of *vec2vec* lies in the Strong Platonic Representation Hypothesis, an enhanced version of the Platonic Representation Hypothesis [Huh et al. (2024)]. This hypothesis posits that deep learning models trained with the same modality and objective, but with different architectures and training data, converge toward a common latent structure, giving rise to a universal representation space. This universal structure is then used to perform “translations” between different embedding spaces.

A.1. Architecture

The architecture of *vec2vec* is modular, inspired by other translation of image vector representations state-of-art methods. In consists of the following components:

- Input Adapters (A_1, A_2): Modules specific to each embedding space, responsible for transforming these embeddings into a shared latent representation space.
- Common Core Network (T): A model that processes the adapted representations within the latent space.
- Output Adapters (B_1, B_2): Modules that decode the latent representations to return the embeddings to their original encoder-specific spaces.

Given that text vector representations do not have a spatial bias like those in images, multilayer perceptrons with residual connections, layer normalization, and SiLU non-linear activation functions are used instead of CNNs.

A.2. Translation process

The translation process is carried out as follows: Let M_1 and M_2 be two text encoders (unknown and known, respectively). The objective is to translate a vector $u_i = M_1(d_i)$ into a representation $F(u_i)$ that approximates the original embedding $v_i = M_2(d_i)$, where d_i is an innaccesible document. The translation functions are defined as a composition of the previously introduced components:

$$F_1 = B_2 \circ T \circ A_1$$

$$F_2 = B_1 \circ T \circ A_2$$

An example of translation between text encoders can be found in Figure 4.

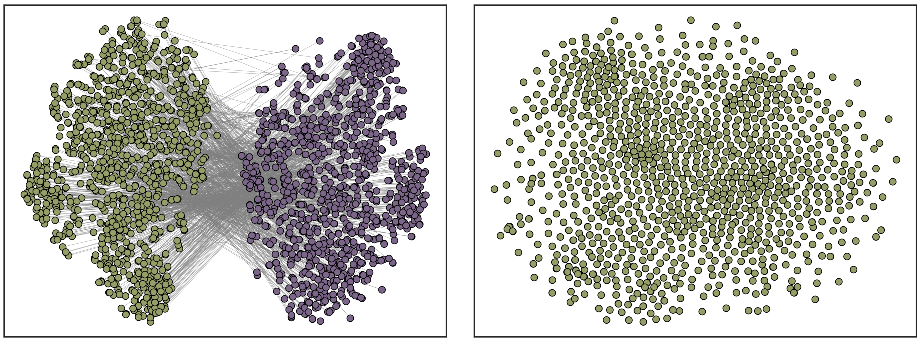

Figure 4: Vector representations generated by *vec2vec*. On the left, we find the embeddings generated for the same dataset by two different encoders (gte and stella). Two connected nodes correspond to the representation of the same document by the different encoders. As can be seen, they are incompatible. After applying the *vec2vec* methodology, alignment is achieved while respecting semantics.

A.3. Optimization and loss functions

The fine-tuning process for vec2vec is formulated as an optimization problem that uses a combination of loss functions to ensure that translations are both accurate and semantically coherent:

- Adversarial Loss: Inspired by GANs (Generative Adversarial Networks), this loss function ensures that the generated embeddings match the distributions of the original embeddings.
- Generator Losses: Three additional constraints are applied:
    - Reconstruction Loss: This loss encourages an embedding, after being mapped to the latent space and then back to its original space, to closely resemble its initial representation.

- Cycle Consistency Loss: Acting as an unsupervised substitute for paired alignment, this loss ensures that translating an embedding to the other space and then translating it back to the initial space results in the least possible information loss.
- Vector Space Preservation Loss: This ensures that the similarity relationships between pairs of embeddings remain consistent after translation.

The combination of these loss functions guides the model to learn a correspondence that not only aligns the embedding distributions but also preserves the geometric and semantic structure of the underlying data.