Topology as a lens for semantic organization in transformer embeddings

Igor Goshev[1], Petar Sekuloski[2], Ivan Chorbev[3], Dimitar Kitanovski[4] and Vesna Dimitrievska Ristovska[5]
Faculty of Computer Science and Engineering, “Ss. Cyril and Methodius” University, Skopje, Macedonia
[1]igor.goshev@finki.ukim.mk, [2]petar.sekuloski@finki.ukim.mk, [3]ivan.chorbev@finki.ukim.mk, [4]dimitar.kitanovski@finki.ukim.mk
[5]vesna.dimitrievska.ristovska@finki.ukim.mk

***Abstract**: This paper examines the geometric structure of sentence embeddings through the lens of persistent homology. The goal is to determine whether semantic similarity produces distinctive topological patterns in a controlled embedding environment. To isolate semantic effects, a single sentence template was combined with different target words, forming two point clouds in a transformer embedding space: one derived from semantically similar words and one from dissimilar words. A Vietoris–Rips filtration was applied to both clouds, and the resulting persistence diagrams were summarized by average lifetime, entropy of birth–death intervals, and the area under the Betti curve. The results show a coherent difference across topological dimensions: similar words generate stable connected components with lower variability, while dissimilar words produce a richer set of cycle features that persist across a broader range of scales. These findings indicate that persistent homology can capture multi-scale structural differences in embedding spaces that are not visible through standard distance-based comparisons. Although the experiment is intentionally simple, it highlights the potential of topological methods for studying how semantic structure is distributed across levels of a neural embedding space.*
***Keywords**: TOPOLOGICAL DATA ANALYSIS, PERSISTENT HOMOLOGY, SENTENCE EMBEDDINGS, SEMANTIC SIMILARITY, SYNONYMS AND ANTONYMS, NATURAL LANGUAGE PROCESSING;*

1. Introduction

Word meanings are rarely isolated. Some words share almost the same meaning, others express the exact opposite, and many are simply unrelated. Modern language models try to capture these relationships by representing words or sentences as points in a high-dimensional space called an embedding space. If two meanings are close, their vectors should also be close; if they differ strongly, they should drift apart. While this idea is intuitive, the actual structure of such embedding spaces is still not fully understood.

Topological Data Analysis (TDA) offers a way to look at the global shape of a dataset rather than only pairwise distances or statistical averages. Persistent homology, one of the core tools in TDA, can detect patterns such as clusters, loops or fragmentation that may not be obvious from direct numerical comparisons. Its main advantage is stability: small changes in data usually do not destroy the overall topological signal.

There have already been attempts to apply TDA to text data, mostly in the context of large corpora or document classification. These studies show that topological descriptors can sometimes reveal hidden structures or improve machine learning tasks. However, much less is known about how basic semantic relationships, like synonymy or antonym, appear from a topological perspective. In other words, does the *shape* of an embedding set change depending on the type of meaning relation?

The aim of this paper is to explore this question in a simple and controlled way. Instead of working with a large or noisy dataset, we focus on a small subset of word pairs taken from the SimLex-999 benchmark. The dataset allows us to separate pairs with high similarity scores from those with very low similarity. For each word, we generate short template-based sentences, embed them using a transformer-based model, and compute persistent homology on the resulting point clouds.

Our goal is not to build a classifier or prove superiority of one embedding method over another. Rather, we want to see whether the topological signatures of the two groups differ in a consistent way. As the results will show, high-similarity pairs tend to form more compact structures with shorter persistence, while low-similarity pairs exhibit more spread and higher complexity. Although preliminary, these findings suggest that topology can provide a different viewpoint on how meaning is encoded in modern embedding models.

2. Related work

Topological Data Analysis has become a useful tool for studying complex, high-dimensional datasets, where traditional geometric or statistical methods struggle to capture global structure. A key milestone in the development of persistent homology as a computational framework was the work by Zomorodian and Carlsson, who formalized how topological features can be detected and tracked across different spatial scales [1]. Later, Adams et al. introduced persistence images, offering a stable and vector-based representation of persistence diagrams suitable for machine learning workflows [2]. These foundations make it possible to analyze embedding spaces not only by distance, but also by their structural and topological properties.

The idea of applying persistent homology to natural language data is relatively recent. One of the first steps in this direction was taken by Fontana et all., who explored whether word embeddings contain detectable topological patterns and showed that such patterns can be used as features in text classification tasks [3]. While their work demonstrated feasibility, it focused on downstream prediction rather than understanding how specific semantic relationships appear in embedding space.

To evaluate semantic similarity more explicitly, benchmark datasets such as SimLex-999 were introduced, providing human-judged similarity scores that allow systematic comparison of word meaning [4]. Alongside this, new embedding models such as Sentence-BERT have made it possible to generate meaningful sentence-level representations while preserving relational structure between words and phrases [5]. More recently, research has shifted toward studying the global behavior of embedding spaces themselves. For example, De Silva et al. showed that persistent homology can reveal structural differences between embedding spaces, such as non-isometric distortions across languages or models [6].

Although these studies point toward the usefulness of topology in language representation research, there is still limited work on controlled experiments that focus specifically on semantic relations. The present paper addresses this gap by analyzing whether persistent homology can reveal measurable differences between embeddings of highly similar versus weakly related word pairs.

3. Methodology

3.1 Dataset and Sentence Construction

When introducing a dataset into a project, it is of crucial importance to have it well organized, especially when working with topology, where everything can be viewed and presented in different ways. In our case, we take word pairs and introduce them into context-independent sentences, which will later be embedded and topologically analyzed.

As the basis of this experiment, we decided to use the SimLex-999 dataset, which is composed of 999 pairs of words. The pairs can consist of abstract adjectives, nouns, or verbs, where each pair is graded according to how similar the words are. Words with related meanings (synonyms) are graded higher, while words with opposite meanings are graded lower.

Besides SimLex-999, there are other datasets as well, such as WordSim-353 and MEN, which, in contrast to the dataset used in this work, grade the meaning only between two nouns and their association. These differences can lead to divergent results, which opens new possibilities for the future, but for now we decided to continue with SimLex-999, as it provides more appropriate data for our needs.

For this experiment, we separated the 999-word pairs into two subsets. First, the dataset was sorted by the similarity score and then split in half:

- 0.23 – 4.67: group with lower semantic similarity
- 4.68 – 9.80: group with higher semantic similarity

This separation allowed us to perform a controlled analysis and comparison between these groups and their semantic behavior.

Since isolated words cannot provide sufficient information about their usage or contextual nature, every word was integrated into context-independent sentences. These types of sentences are of a good value because they retain their semantic structure regardless of which word is inserted into them.

The decision to use these types of sentences allows us to examine the dependency between words with high semantic similarity and those with low similarity, as well as to observe their behavior in these controlled contexts.

Each of the 1,998 words was inserted into ten different template sentences, resulting in a total of 19,980 sentences. This enabled us to observe the behavior not only of the individual words but also of the words within full sentence contexts, which is a significant factor for the subsequent topological and embedding analysis discussed in the next section.

3.2 Sentence Embeddings and Point-Cloud Construction

After constructing the full set of generated sentences, the next step was to convert them into numerical representations that could be meaningfully compared. For this purpose, the Sentence-BERT AllMiniLM-L6-v2 model was used as the embedding engine. This model is widely used because it produces semantically meaningful sentence-level embeddings without requiring task-specific fine-tuning.

In practice, the model takes each sentence and maps it to a fixed-dimensional vector:

$$x = E(s) \in \mathbb{R}^{384}$$

Each embedding can be interpreted as a coordinate in a 384-dimensional semantic space. Because differences in magnitude are not informative for this analysis, all vectors were normalized to unit length using:

$$\hat{x} = \frac{x}{||x||_2}$$

Once normalized, the cosine similarity between two embeddings simplifies to a distance measure based purely on the angle between them. In this form, Euclidean distance and cosine similarity are interchangeable, and distance can be expressed as:

$$d(\hat{x}_i, \hat{x}_j) = 1 - \langle \hat{x}_i, \hat{x}_j \rangle$$

With the embeddings computed and normalized, the next step was to apply PCA, reducing their dimensionality from 384 to 150. This reduction decreased the computational cost, enabling faster execution and lower energy consumption.

As we previously discussed, from there they were grouped according to their source word's similarity classification from SimLex-999. This created two distinct and balanced point sets:

$$X_{HS} = \{\hat{x}_1^{HS}, \dots, \hat{x}_{9990}^{HS}\} \;\; X_{LS} = \{\hat{x}_1^{LS}, \dots, \hat{x}_{9990}^{LS}\}$$

Each sentence embedding corresponds to one point in the space, meaning each group forms a point cloud. Conceptually, these clouds represent how the model organizes meaning internally.

3.3 Persistent Homology Analysis

To move beyond pairwise distances and explore the shape of the embedding space, persistent homology was applied. Unlike standard clustering or dimensionality reduction methods, persistent homology does not rely on assumptions about convexity, density, or linear separability. Instead, it examines how connectivity patterns evolve across distance scales and records geometric structure that remains stable across those scales.

The analysis begins by constructing a Vietoris–Rips filtration, which builds a family of simplicial complexes based on a threshold parameter ε. For a given value of ε, a simplicial complex is defined as:

$$VR_{\{\varepsilon\}(X)} = \{\sigma \subseteq X \mid d(x_i, x_j) \leq \varepsilon, \forall x_i, x_j \in \sigma\}$$

At ε = 0, every point is isolated. As ε increases, points connect into edges, triangles, and higher-order simplices, eventually forming a fully connected structure. Persistent homology tracks when these structures appear (their birth) and when they vanish (their death).

For each topological dimension k, this process produces a set of birth-death intervals:

$$D_k = \{(b_i, d_i), i = 1, \dots, N_k\}.$$

In this work, only the first two dimensions are examined:

k = 0: connected components, representing how quickly points merge into clusters.

k = 1: loops or cycle structures, which suggest curved or hollow regions in the cloud.

To compare the two point clouds quantitatively, several commonly used persistence-based descriptors were computed. The first was the average lifetime of features:

$$\bar{L}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} (b_i - d_i)$$

which serves as a measure of how long topological features persist before collapsing.

A second descriptor was persistence entropy, which captures how evenly or unevenly feature lifetimes are distributed:

$$p_i = \frac{d_i - b_i}{\sum_{j=1}^{N_k} (d_j - b_j)} H_k = -\sum_{i=1}^{N_k} p_i \log(p_i).$$

Finally, Betti curves were generated to provide a scale-dependent view of topological structure:

$$\beta_k(\varepsilon) = |\{(b_i, d_i) \in D_k | b_i \leq \varepsilon < d_i\}|)$$

4. Results

Persistent homology was computed separately for the high-similarity and low-similarity groups of sentence embeddings. For each filtration, birth–death intervals were extracted and summarized using the average lifetime, the persistence entropy H_k, and the area under the Betti curve AUC_{β_k}. All numerical descriptors are reported in Table 1.

Table 1 *The results for average lifetime, entropy and Betti AUC*

	$\bar{L}_k$	$entropy_k$	AUC_{β_k}
High similarity, $k = 0$	0.547886	8.874602	1901.88
Low similarity, $k = 0$	0.357307	9.129177	1579.12
High similarity, $k = 1$	0.099668	7.013554	417.92
Low similarity, $k = 1$	0.185433	9.057350	497.34

For dimension $k = 0$, which reflects the merging of connected components, the high-similarity group exhibits a larger average lifetime ($\bar{L}_0$=0.547886) compared to the low-similarity group ($\bar{L}_0$=0.357307). This indicates that components in the high-similarity embedding cloud persist over a wider range of scales before merging. The corresponding entropies show the opposite trend: the low-similarity cloud exhibits a higher entropy value ($entropy_0$=9.129177) than the high-similarity cloud ($entropy_0$=8.874602), suggesting that lifetimes in the low-similarity group are more dispersed and heterogeneous. The Betti AUC values follow the same pattern (AUC_{β_0}=1901.88 vs. AUC_{β_k} =1579.12).

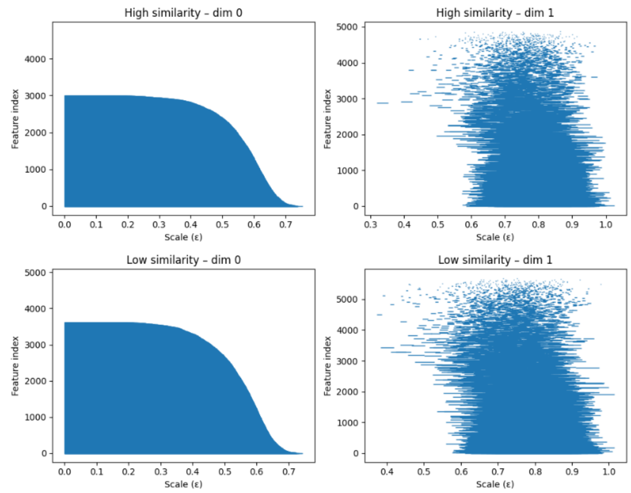

Figure 1 *Persistence Barcodes for High-similarity and Low-similarity group*

For dimension $k = 1$, which captures loop-like configurations, the differences are more pronounced. The low-similarity group shows a larger average lifetime ($\bar{L}_1$==0.185433) than the high-similarity group ($\bar{L}_1$==0.099668), meaning that cycles in the low-similarity cloud persist across a broader part of the filtration. The entropy values are again higher for the low-similarity group ($entropy_1$=9.057350 vs. $entropy_1$==7.013554). This is reflected in the Betti AUC values (AUC_{β_1}=497.34 vs. AUC_{β_1}=417.92). The persistence diagrams are shown in *Figure 1.*

Taken together, these observations show a consistent pattern across dimensions. Semantic similarity produces longer-lived components with lower variability, while lower similarity yields shorter component lifetimes but longer and more heterogeneous cycles. Even though sentences were generated using a fixed template, persistent homology detects differences in the geometry of the embedding space.

5. Discussion

The results suggest that semantic similarity shapes the topology of the embedding space in two complementary ways. At small filtration scales, the embeddings of sentences built around closely related words tend to remain separated for longer and collapse into connected components in a more uniform fashion. This behavior is reflected in the higher average lifetime $\bar{L}_0$, lower persistence entropy, and a larger AUC_{β_0} for the high-similarity group. Since the sentence template was identical across all samples, the only source of variation was the target word inserted into the sequence. The stability of these small-scale features therefore points to a genuine effect of semantic coherence: similar concepts induce tighter local neighborhoods in the model's latent space, and their merging across scales follows a more stable trajectory.

At larger scales, the situation reverses. Embeddings derived from sentences containing dissimilar words produce richer and more sustained loop-like configurations, captured by longer average lifetimes $\bar{L}_1$, higher entropy, and a larger AUC_{β_1}. These signatures indicate that the low-similarity point cloud spans a broader portion of the embedding manifold, creating intermediate gaps that take more time to fill during the filtration. In this case, the absence of semantic coherence does not simply produce random noise; instead, it seems to encourage exploration of directions that unfold into higher-order geometric structure. The difference between the two scenarios aligns with an intuitive interpretation of how deep language models organize meaning: clusters form readily around coherent concepts in local neighborhoods, while global manifold structure emerges from contrasts between unrelated semantic axes.

From a methodological perspective, the coexistence of these two effects is significant. Classical embedding analyses often rely on clustering, dimensionality reduction, or cosine similarity comparisons. While such tools can reliably capture local coherence and relative distances, they provide limited information about how structure evolves across scales. Persistent homology offers a complementary view: the transition from isolated components to higher-order simplices is tracked throughout the filtration, making it possible to evaluate not only whether two concepts are close but also ow their neighborhoods merge and what features arise along the way. In this experiment, the topological descriptors captured differences that cannot be reduced to a single similarity score, showing that topology can reveal both stability within semantic clusters and variation driven by conceptual diversity.

The patterns observed in the diagrams and summary metrics are also consistent with the geometric behavior of deep language models such as BERT. These models embed linguistic information on a non-linear manifold, where semantic directions do not align to simple axes but arise from a training signal distributed across multiple layers. As a result, a group of sentences containing similar words may form a compact region supported by shared contextual structure, while diverse concepts generate a broader arm of the manifold that wraps around pre-existing clusters. The persistence of 1-dimensional features in the latter case suggests that the global geometry has curvature and gaps that require multiple filtration thresholds to close. This perspective emphasizes the relevance of topological tools for understanding latent spaces shaped by implicit tasks and unsupervised objectives.

Finally, it should be emphasized that these findings arise from a controlled and deliberately simple setting. The sentences were generated using an artificial template, and only one word varied across samples, making this a highly constrained experiment. This setup allowed the effect of semantic similarity to be isolated, but it does not capture the full richness of natural language, nor does it explore how topology behaves under different syntactic constructions, sentence lengths, or contextual structures. The present study illustrates a single experiment designed to highlight the presence of topological signatures rather than to provide comprehensive statistical conclusions.

6. Conclusion

This work explored whether topological features derived from sentence embeddings reflect differences in semantic similarity between words. By constructing controlled sentence templates and applying persistent homology, we obtained a geometric view of the embedding space that goes beyond pairwise distances. The results show that the two groups of sentences produce distinct patterns. Their persistence descriptors and Betti curves, indicating that semantic similarity influences the structure of the point cloud at multiple scales. Although the experiment is intentionally simplified, the observations suggest that topological summaries may serve as a complementary way to characterize embedding behavior.

7. Future work

This study highlights a clear topological signal associated with semantic similarity, but several directions remain open. A natural extension is to move beyond synthetic sentences and apply the same analysis to real text corpora, where variation in sentence structure, context, and discourse patterns may produce richer and less predictable geometric behavior. In particular, persistent homology could be applied to document-level embeddings, contextualized sentence embeddings, or token-level trajectories across layers of a transformer model.

A second line of investigation concerns the embedding models themselves. The present experiment used a single encoder architecture, and it is not yet known whether the observed patterns generalize across different model families, such as BERT variants, sentence-transformers, large decoder-based models, or multilingual encoders. Comparing the topological signatures across architectures could clarify whether cycles and component persistence arise from general properties of modern language representations or from model-specific inductive biases.

Third, the current work employed a Vietoris–Rips filtration on the Euclidean metric. Alternative filtrations such as, witness complexes, or density-based filtrations may highlight different aspects of the underlying geometry and help distinguish between noise-driven features and persistent structures. In addition, the use of persistence landscapes or kernel-based methods may enable statistical testing on collections of diagrams rather than relying on single-instance descriptors.

Finally, a larger-scale study could explore connections between topological descriptors and downstream tasks. If persistent features correlate with clustering quality, classification performance, or interpretability measures, then persistent homology could play a practical role in model diagnostics. More broadly, the topological approach may offer an avenue toward explaining how semantic information is organized in embedding spaces, complementing existing explainable AI techniques that focus on feature importance or attention patterns.

8. Acknowledgement

This work was partially financed by the Faculty of Computer Science and Engineering at Ss. Cyril and Methodius University in Skopje.

References

[1] A. Zomorodian, G. Carlsson, Computing persistent homology, *Discrete & Computational Geometry*, vol. 33, pp. 249–274, (2005).

[2] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, et al., Persistence images: A stable vector representation of persistent homology, *Journal of Machine Learning Research*, vol. 18(8), pp. 1–35, (2017).

[3] F. Fontana, G. L. Torrisi, A. Tulipani, A novel method of extracting topological features from word embeddings, arXiv preprint arXiv:2003.13074, (2020).

[4] F. Hill, R. Reichart, A. Korhonen, SimLex-999: Evaluating semantic models with (genuine) similarity estimation, *Computational Linguistics*, vol. 41(4), pp. 665–695, (2015).

[5] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, Proceedings of EMNLP, pp. 3982–3992, (2019).

[6] M. De Silva, J. P. Jackson, B. Harreld, J. Mordukhovich, The shape of word embeddings: Quantifying non-isometry with topological data analysis, *Findings of EMNLP*, (2024).