Application of Harmonic Token Projection (HTP) on STS data sentences: Benchmarking deterministic embeddings

Débora de Faria Ferreira Gomes, **Tcharlies Schmitz**,
Data Science - PX.Center
Joinville, SC, Brazil
E-mail: debora.gomes@px.center, tcharlies.schmitz@px.br

Talia Correia Schulz
Postgraduate Program in Numerical Methods and Engineering – UFPR
Curitiba, PR, Brazil
E-mail: talia.correia@ufpr.br

ABSTRACT

This work presents the *Harmonic Token Projection* (HTP), a reversible and deterministic framework for generating text embeddings without training, vocabularies, or stochastic parameters. Unlike neural embeddings that depend on statistical co-occurrence or optimization, such as *Word2Vec* [2], *GloVe* [3], and transformer-based architectures like *BERT* [1] and *Sentence-BERT* [4], HTP encodes each token analytically as a harmonic trajectory derived from its Unicode integer representation, establishing a bijective and interpretable mapping between discrete symbols and continuous vector space. The harmonic formulation provides phase-coherent projections that preserve both structure and reversibility, enabling semantic similarity estimation from purely geometric alignment.

The method was evaluated on the *Semantic Textual Similarity Benchmark* (STS-B), comparing its analytical correlation with human judgments to conventional embeddings. Despite its simplicity and the absence of any training, HTP achieved a Spearman correlation of $\rho = 0.68$ and Pearson $r = 0.67$, approaching the performance of transformer-based models at a fraction of their computational cost. The model encodes thousands of sentences per second with sub-millisecond latency on CPU, using less than one megabyte of memory.

Method	**Training**	**Spearman (ρ)**	**Pearson (r)**	**RAM (MB)**
BERT (base) [1]	Supervised	0.68	0.70	3000
GloVe [3]	Supervised	0.65	0.66	450
Sentence-BERT [4]	Supervised	0.77	0.78	2100
Word2Vec [2]	Supervised	0.61	0.63	400
HTP (proposed)	**None**	**0.68**	**0.67**	<1

Table 1: Comparative results on the *STS-Benchmark* dataset. Results for Word2Vec [2] and GloVe [3] correspond to reported baselines on the STS-B dataset using pre-trained models. Transformer-based results for BERT [1] and Sentence-BERT [4] were taken from their respective publications and benchmark leaderboards. HTP results were computed analytically on CPU without training or corpus statistics, with empirical RAM usage measured directly during inference.

The results demonstrate that HTP captures a large fraction of linguistic similarity through deterministic geometry alone, without statistical learning. Its analytical design ensures full reversibility, mathematical transparency, and reproducibility across environments. The harmonic

formulation bridges number theory, signal processing, and language modeling, suggesting that part of what is interpreted as "semantic similarity" can arise from the intrinsic geometry of symbolic systems.

Keywords: *Deterministic embedding, Harmonic Token Projection, Reversible encoding, Semantic similarity, Analytical geometry.*

References

[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *NAACL-HLT*, 2019.

[2] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*, 2013.

[3] J. Pennington, R. Socher and C. D. Manning, GloVe: Global vectors for word representation, *EMNLP*, pp. 1532–1543, 2014.

[4] N. Reimers and I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT networks, *EMNLP*, 2019.