Similarity Between Original and Augmented Data

Matheus Yasuo Ribeiro Utino[1], Elton H. Matsushima[2], Aline Paes[3], Paulo Mann[4]

[1] Institute of Mathematics and Computer Science, University of São Paulo
[2] Department of Psychology, Fluminense Federal University
[3] Institute of Computing, Fluminense Federal University
[4] Institute of Computing, Federal University of Rio de Janeiro
matheusutino@usp.br, eh.matsushima@gmail.com, alinepaes@ic.uff.br, paulomannjr@gmail.com

1 Similarity Metrics Between Embeddings

To evaluate the similarity between the original and augmented embeddings, we use three main metrics: the Average Pairwise Cosine Similarity, the Centroid Cosine Similarity, and the Hausdorff Distance.

Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{N_1}\}$ be the set of original embeddings, where each $\mathbf{e}_i$ represents the embedding of the i-th post or data sample. Let $\{\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_{N_2}\}$ denote the set of embeddings generated through data augmentation, derived from the original set as a whole.

1.1 Average Pairwise Cosine Similarity (Avg Pairwise CosSim)

The Average Pairwise Cosine Similarity measures the average similarity between all pairs of embeddings within a set. For embeddings $\mathbf{e}_i$ and $\mathbf{e}_j$ from the original set, the cosine similarity is given by:

$$\text{CosSim}(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|} \tag{1}$$

Where $\mathbf{e}_i \cdot \mathbf{e}_j$ denotes the dot product, and $\|\mathbf{e}_i\|$ and $\|\mathbf{e}_j\|$ are the vector norms. The average is computed over all unique pairs in the set:

$$\text{Avg Pairwise CosSim} = \frac{1}{N_1(N_1 - 1)} \sum_{i \neq j} \text{CosSim}(\mathbf{e}_i, \mathbf{e}_j) \tag{2}$$

Lower values of Avg Pairwise CosSim indicate greater diversity among embeddings (i.e., more spread in the vector space), while higher values suggest a more homogeneous set.

1.2 Centroid Cosine Similarity (Centroid CosSim)

The Centroid Cosine Similarity compares the centroids of the original and augmented embedding sets. The centroids C_1 and C_2 are defined as:

$$C_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{e}_i, \quad C_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}\mathbf{e}'_i \tag{3}$$

Their similarity is given by the cosine of the angle between the centroid vectors:

$$\text{CosSim}(C_1, C_2) = \frac{C_1 \cdot C_2}{\|C_1\|\|C_2\|} \tag{4}$$

A high Centroid CosSim implies that the original and augmented sets occupy nearby regions in the embedding space, while a low value indicates dissimilarity between the distributions.

1.3 Hausdorff Distance (Hausdorff Dist)

The Hausdorff Distance quantifies the maximum discrepancy between the two sets. Formally, for the original set $A = \{\mathbf{e}_1, \ldots, \mathbf{e}_{N_1}\}$ and the augmented set $B = \{\mathbf{e}'_1, \ldots, \mathbf{e}'_{N_2}\}$, the Hausdorff distance is defined as:

$$d_H(A, B) = \max\left(\max_{a\in A}\min_{b\in B}\|a-b\|, \max_{b\in B}\min_{a\in A}\|a-b\|\right) \tag{5}$$

Where $\|a - b\|$ is the distance between points a and b. This metric is useful for capturing the most significant differences between the sets of embeddings, providing a measure of discrepancy between the distributions.

Lower values of Hausdorff Distance indicate that the sets of embeddings are more similar to each other, as the points in one set are closer to the points in the other set. Higher values suggest greater dissimilarity, meaning the sets are more spread out in the embedding space.

2 Results

Table 1 shows BERT-based models (BERTM and BERTimbau) exhibit consistent characteristics with respect to embedding similarity metrics. BERTM yields an average pairwise cosine similarity (Avg Pairwise CosSim) of 0.316, while BERTimbau demonstrates a slightly higher value of 0.331, indicating a degree of homogeneity in the generated embeddings. This consistency is further reflected in the Centroid CosSim values, with scores of 0.813 for BERTM and 0.832 for BERTimbau, suggesting that the embeddings are closely clustered around their respective centroids. The Hausdorff Distance, which quantifies the maximum discrepancy between points, is slightly higher for BERTM at approximately 1.152, suggesting a moderate level of diversity in the resulting embeddings. Overall, both models display similar behavior in terms of embedding variation.

In contrast, Large Language Models (LLMs)—specifically Dolphin 3, Gemini, and Mistral—exhibit more varied behavior, particularly when embeddings are generated using input data that includes the Beck's Depression Inventory

(BDI-II). For instance, Dolphin 3 with BDI-II shows a notably lower Avg Pairwise CosSim of 0.305, indicating greater diversity among the embeddings. This is accompanied by a higher Hausdorff Distance, reflecting increased disparity between points. When BDI-II is excluded, the Avg Pairwise CosSim for Dolphin 3 increases to 0.337, indicating greater homogeneity, with the Hausdorff Distance correspondingly decreasing to 1.152.

A similar trend is observed for Gemini. With BDI-II, the Avg Pairwise CosSim decreases to 0.334 and the Hausdorff Distance increases to 1.162, indicating a more dispersed embedding space. Without BDI-II, these metrics improve: Avg Pairwise CosSim rises to 0.357 and the Hausdorff Distance decreases to 1.137, suggesting greater representational consistency. Mistral follows the same pattern; with BDI-II, it produces an Avg Pairwise CosSim of 0.336 and a Hausdorff Distance of 1.187, while without BDI-II, these values shift to 0.362—the highest Avg Pairwise CosSim among the LLMs evaluated—and 1.112, respectively, indicating enhanced consistency.

When comparing the LLMs with and without BDI-II, it becomes evident that the inclusion of BDI-II tends to increase representational variability, as shown by lower Avg Pairwise CosSim values and higher Hausdorff Distances. Conversely, the exclusion of BDI-II promotes more uniform embeddings, reflected in higher Avg Pairwise CosSim and lower Hausdorff Distance values.

In comparison to the LLMs, BERT-based models generate more consistent embeddings, as evidenced by higher Centroid CosSim values, indicating tighter clustering around the mean representation. LLMs, particularly when incorporating BDI-II, yield more dispersed embeddings with lower Avg Pairwise CosSim and higher Hausdorff Distances. These findings suggest that while LLMs exhibit greater flexibility in capturing subtle data variations, they also introduce increased variability in the resulting embeddings.

Table 1: Comparison of Original vs. Augmented Embedding Similarity Metrics by Model and BDI Status. We rely on OpenAI's text-embedding-3-large embeddings for computing similarity metrics. The Copy method was not included, as it represents an exact copy of the original data.

Model	**BDI**	**Avg Pairwise CosSim**	**Centroid CosSim**	**Hausdorff Dist**
BERTM	—	0.316 ± 0.011	0.813 ± 0.003	1.152 ± 0.003
BERTimbau	—	0.331 ± 0.001	0.832 ± 0.001	1.144 ± 0.001
Dolphin 3	BDI	0.305 ± 0.005	0.652 ± 0.010	1.226 ± 0.005
Dolphin 3	No BDI	0.337 ± 0.002	0.732 ± 0.002	1.152 ± 0.002
Gemini	BDI	0.334 ± 0.001	0.762 ± 0.002	1.162 ± 0.003
Gemini	No BDI	0.357 ± 0.002	0.842 ± 0.002	1.137 ± 0.002
Mistral	BDI	0.336 ± 0.001	0.787 ± 0.002	1.187 ± 0.003
Mistral	No BDI	0.362 ± 0.002	0.870 ± 0.002	1.112 ± 0.002