UDC 004.934

A PEARSON-SPEARMAN APPROACH FOR EVALUATING SEMANTIC SIMILARITY TASKS

Nikolaichuk A.I.
e-mail: anna.nikolaichuk@nure.ua
Науковий керівник – к.т.н., ас. Кобилін І. О.
Kharkiv National University of Radio Electronics, Department of SysEng
Kharkiv, Ukraine

This work evaluates the performance of Semantic Textual Similarity (STS) models across different language pairs and model characteristics. The comparison is conducted using both Pearson and Spearman correlation coefficients to mitigate the limitations of each. The results show that the newer models, GTE and MPNet, achieved the best performance, followed by MiniLM, which has a smaller embedding dimensionality. The findings indicate that model performance is influenced by a combination of factors, rather than a single one. Additionally, the study highlights the challenges of cross-lingual similarity assessment.

One of the key problems in Natural Language Processing (NLP) is determining how similar two pieces of text are in meaning – a Semantic Textual Similarity (STS) task. It is crucial for information retrieval, text summarization, machine translation and question-answering systems. However, the performance of STS models varies depending on the specific task and dataset.

A critical issue is choice of metrics for evaluating the quality of similarity predictions compared to human judgments. Pearson correlation is commonly used, but [1] highlights that it can be misleading, especially for ranking or classification tasks.

For evaluation, the Semantic Textual Similarity Benchmark (STSB) is used – a dataset containing 5749 English (EN-EN) sentence pairs with corresponding normalized similarity scores [2] provided by human judges. For Ukrainian (UK-UK), the dataset was translated using advanced machine translation. To assess cross-lingual performance, mixed English-Ukrainian datasets (UK-EN, EN-UK) were created.

Table 1 summarizes the parameters of the evaluated models.

Table 1 – Parameters of the models evaluated

Model Name	Embedding Dimensionality	Language Support	Base Model Architecture
LaBSE	768	110	BERT
MiniLM	384	50	
StaticSim	1024	51	
DistilUSE	512	50	DistilBERT
GTE	768	75	GTE
MPNet	768	50	XLM-RoBERTa

The STS of generated embeddings is calculated using cosine similarity, and the results are evaluated using correlation coefficients. Pearson correlation (r) measures the linear relationship between the predicted scores and gold labels using the following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

where x_i and y_i are individual predicted and true scores, $\bar{x}$ and $\bar{y}$ are the mean values of the corresponding scores.

To address the limitations of Pearson correlation, such as its sensitivity to outliers and linearity restriction, Spearman correlation (ρ) is suggested by [1]. This metric ranks values, reducing outlier sensitivity, capturing non-linear relationships and handling non-normally distributed data. It is computed as:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)},$$

where d_i is the difference between the ranks of predicted and true scores, n is the total number of sentence pairs.

Results of evaluation of model predictions using correlation coefficients are presented in Table 2 as $r \times 100$ for Pearson correlation and $\rho \times 100$ for Spearman.

Table 2 – Correlation scores for intra- and cross-lingual sentence pairs

Model Name	**EN-EN**		**UK-UK**		**UK-EN**		**EN-UK**		**Average**	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
LaBSE	74.9	73.6	74.4	73.1	72.0	70.3	71.7	69.9	73.3	71.7
MiniLM	85.1	83.9	81.6	80.1	78.8	75.1	78.1	74.6	**80.9**	**78.4**
StaticSim	82.1	79.6	72.7	70.1	63.6	60.5	63.4	60.2	70.5	67.6
DistilUSE	81.0	78.7	76.7	74.4	73.5	70.0	73.0	69.5	76.1	73.1
GTE	87.4	87.4	85.1	84.1	80.6	77.8	80.6	78.1	**83.4**	**81.9**
MPNet	86.3	85.9	83.1	82.1	80.7	78.0	80.6	77.8	**82.7**	**81.0**

Overall, average correlation values show that the GTE and MPNet models captured semantic meaning and syntactic nuances most effectively. MiniLM, with similar results, offers the added advantage of a smaller embedding size, making it more suitable for environments with limited computational power.

Results show patterns linking model performance to specific characteristics. Models with embedding dimensionality (which determines the ability to capture deeper semantic details) of 768 performed best (MPnet and GTE). However, MiniLM, with a dimensionality of 384 (the lowest among evaluated models), also performed well. This supports findings in [3] that some model optimization techniques can reduce dimensionality while minimizing performance loss (dimensionality can be reduced to as low as 128). Interestingly, StaticSim, with

the highest dimensionality of 1024, performed poorly, likely due to its use of precomputed static embeddings, which are less flexible with unseen examples.

Base model architecture also influences performance, as the best-performing models have newer architectures – XLM-RoBERTa (MPNet) and GTE. Older BERT-based models (LaBSE, StaticSim) struggled the most and were even outperformed by BERT-derived, reduced-size models (MiniLM, DistilUSE).

In general, models performed better on intra-lingual sentence pairs. Generating embeddings of the same sentence in different languages is challenging, as models cannot always generate symmetrical representations. Overall, the results for cross-lingual sentence pairs reflect the previously established trend, with MPNet, GTE, and MiniLM outperforming other models.

The evaluation of sensitivity to language order, using two datasets with the same sentences where the languages were switched, showed minimal correlation differences between the datasets. MPNet and GTE performed best with low sensitivity to language order, likely due to balanced training data and better encoding. Older BERT-based and smaller models (MiniLM, DistilUSE, LaBSE) showed greater variability. StaticSim displayed moderate asymmetry, suggesting that static embeddings do not necessarily hinder sentence order handling.

In conclusion, this evaluation of STS models shows that performance varies depending on the language pair and model characteristics. While GTE and MPNet demonstrated the highest correlation scores across all tasks, MiniLM also emerged as a strong competitor, due in part to its smaller embedding size, highlighting the effectiveness of optimized architectures. The results suggest that embedding dimensionality alone does not guarantee superior performance. Instead, a combination of model architecture, training strategy and optimization techniques plays a crucial role.

The evaluation also underscores the challenge of cross-lingual similarity assessment, where models generally performed worse than in intra-lingual settings. Notably, the impact of language order on results was minimal.

In this experiment, both Pearson and Spearman correlations determined the models' performance similarly. This suggests that the relationships between predicted and true similarity scores for the datasets used were mostly linear and that models handled ranking tasks effectively.

References:

1. Reimers N., Beyer P., Gurevych I. Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity. *COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka. 2016. P. 87–96.

2. STSB: Semantic Textual Similarity Benchmark. *Hugging Face*. URL: https://huggingface.co/datasets/sentence-transformers/stsb (date of access: 03.03.2025).

3. Wang H., Zhang H., Yu D. On the Dimensionality of Sentence Embeddings. 2023. 11 p. (Preprint).