UDC 519.816:004.85

SEMANTIC TEXTUAL SIMILARITY MODELS EVALUATION FOR TASKS REQUIRING BINARY THRESHOLD DECISIONS

Nikolaichuk A.I.
e-mail: anna.nikolaichuk@nure.ua
Kharkiv National University of Radio Electronics, Department of SysEng
Kharkiv, Ukraine

This work evaluates Semantic Textual Similarity (STS) models for tasks requiring binary threshold decisions. The experiment used English and Ukrainian STS datasets, categorizing sentence pairs as successful or challenging based on a similarity score difference threshold. GTE demonstrated higher resistance to typos in English, while MPNet and MiniLM struggled more with lexical and morphological variations in Ukrainian. The analysis highlights the importance of considering task-specific and dataset-related challenges when selecting models for optimal performance.

Semantic Textual Similarity (STS) models are used in various natural language processing tasks, each requiring specific characteristics. In plagiarism detection, only a binary decision (above or below threshold) matters [2]. In question-answer pairing, models should prioritize relevant answers, while in text summarization, they must detect dissimilar sentences for diversity. As correlation metrics provide only general accuracy, task-specific evaluation is often necessary.

To address this, the experiment categorizes sentence pairs as successful or challenging based on a 0.3 error threshold, where successful pairs align closely with human judgment. This approach helps identify error patterns and provide insights into linguistic challenges in STS performance.

The Semantic Textual Similarity Benchmark (STSB) is a dataset with 5749 English sentence pairs and human-assigned similarity scores. For this experiment, the version with normalized scores [4] was used to compare gold labels with model predictions. The Ukrainian dataset (UK-UK) was created by machine translation of the English dataset (EN-EN), using the same similarity scores.

To choose models for this experiment, six sentence-transformer models were evaluated using Pearson correlation coefficient. The highest correlation scores of 0.86, 0.85 and 0.83 were achieved by gte-multilingual-base (GTE – General Text Embedding), paraphrase-multilingual-mpnet-base-v2 (MPNet) and paraphrase-multilingual-MiniLM-L12-v2 (MiniLM) models, respectively.

The GTE Multilingual Base is an encoder-only transformer that can generate both dense and sparse vectors, outperforming similar models while requiring less compute for inference. The MPNet model maps sentences into 768-dimensional dense vectors for tasks like clustering and semantic search, supporting multilingual sentence embeddings with efficient mean pooling. The MiniLM is a lightweight model that maps text into 384-dimensional dense vectors with strong multilingual performance, being more efficient than larger transformers.

A metric for semantic similarity used is cosine similarity, which measures the cosine of the angle between two vectors in a high-dimensional space as:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (1)$$

where A and B – the embedding vectors of two sentences, $A \cdot B$ – dot product of the vectors, $\|A\|$ and $\|B\|$ – the magnitudes of the vectors.

As gold labels are normalized to 0-1 scale, the cosine similarity values, ranging from -1 to 1 was normalized using Min-Max normalization with formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \quad (2)$$

where x' is normalized similarity score, $\min(x)$ is the lowest cosine similarity value, and $\max(x)$ is the highest.

Initially, the cosine similarity values (-1 to 1) were mapped to a 0-1 scale. However, the results were unsatisfactory, with prediction scores starting around 0.4-0.6, and only 51% of predictions on average were successful. Based on these findings, the minimum and maximum values for scaling were taken from the cosine similarity computation for each model's embeddings. This approach helped "stretch" the scores to start from 0, which raised the success rate to 77%. Graphs of the categorization results for each model are shown in Figure 1.

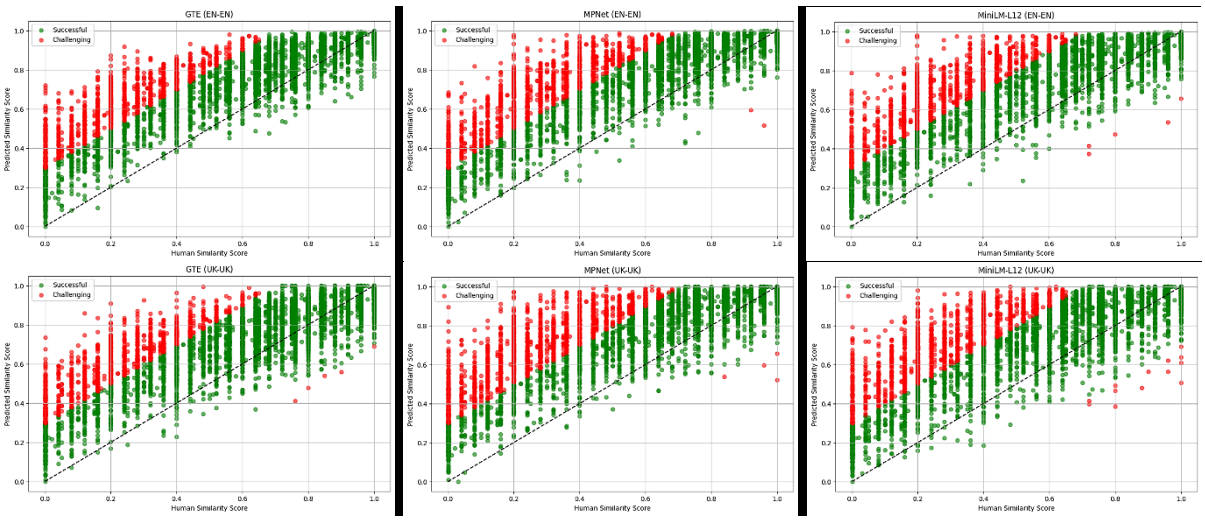

Figure 1 – Categorization of STS predictions based on human similarity scores

The prevalence of red in the upper-left corner of the graphs shows that the models tend to overestimate similarity. This may be due to feature suppression in mean pooling, a common aggregation method in contrastive learning-based sentence embeddings, as discussed in [3]. Since contrastive learning generates

positive samples through data augmentation, the models focus more on lexical overlap than semantic understanding, leading to higher similarity scores for sentences with similar wording but different meanings.

Compared to overestimation, underestimation cases are less frequent. For the English dataset, GTE showed none, MPNet only 2, and MiniLM showed more. MPNet and MiniLM both struggled with similar patterns, particularly with typos. Examples include *"I wood have asked her if she is up for a hiike."* vs. *"I wood have at least asked if a hike was out of the question."*. As noted in [1], transformers like the ones evaluated in this experiment struggle with noisy words not present in the vocabulary. These words are split into sub-words, causing token distribution shifts during embedding generation, which leads to performance drops.

For the Ukrainian dataset, underestimation is more frequent, likely due to its low-resource nature and complex inflectional system, which challenge sentence embeddings. GTE struggled with longer sentences and complex syntactic structures, particularly in news-related content (*"За її словами, Друс досі утримується у в'язниці..."*). The model fails to capture deep semantic similarities when word order varies. MPNet tends to underestimate similarity in paraphrased sentences with lexical variation but the same intent (*"Я міг би запитати її, чи не хоче вона піти в похід."* vs. *"Я хоча б запитав, чи є похід зовсім неможливим."*). The model also struggles with free word order. MiniLM exhibits the highest underestimation rates, particularly for short factual statements and morphologically rich words (*"Чорний птах сидить на землі"* vs. *"Чорний дрізд сидить на землі"*). The model struggles with synonymy and minor morphological differences, which are more prominent in Ukrainian.

In conclusion, selecting an appropriate STS model requires careful consideration of the dataset's characteristics and linguistic challenges. For datasets with typos or abbreviations, a less sensitive model like GTE is preferable. Ukrainian STS remains particularly challenging, as different models struggle with different aspects: GTE with complex structures, MPNet with lexical variation and MiniLM with morphological variations. Additionally, scaling methods may significantly impact results and should be chosen carefully. Thus, achieving optimal STS performance depends on balancing various influencing factors.

References:

1. LEA: Improving Sentence Similarity Robustness to Typos Using Lexical Attention Bias / M. Almagro et al. 2023. 11 p. (Preprint).

2. 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.

3. SNCSE: Contrastive Learning for Unsupervised Sentence Embedding with Soft Negative Samples / H. Wang et al. 2022. 7 p. (Preprint).

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