ISSN 3083-7715 (online) *Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)*
ISSN 3083-7650 (0135-1710) (print) *Management Information System and Devices. 2026. No. 2 (189)*

UDC 004.032.26 DOI: https://doi.org/10.30837/0135-1710.2026.189.284

Anna Nikolaichuk, Oleg Kobylin, Ilya Kobylin, Oleksandra Putyatina

A COMPREHENSIVE EVALUATION OF TRANSFORMER MODELS FOR SENTENCE-LEVEL SEMANTIC SIMILARITY IN ENGLISH AND UKRAINIAN

The subject of this study is transformer-based models for semantic textual similarity and approaches to their evaluation in monolingual and cross-lingual settings for English and Ukrainian. **The aim of the work** is to evaluate and compare the effectiveness of transformer models on English, Ukrainian, and English-Ukrainian sentence pairs, using different approaches to assessing their quality and practical applicability. To achieve this aim, the following tasks are addressed: comparing models on monolingual and cross-lingual datasets; analyzing results using Pearson and Spearman correlation coefficients; assessing practical applicability by classifying predictions based on error thresholds; and investigating the impact of language on model accuracy. **Methods.** The study employs transformer models GTE, LaBSE, MiniLM, and MPNet, evaluated on the STS-B dataset, as well as its Ukrainian and English-Ukrainian versions. Cosine similarity is used to compute sentence similarity, while evaluation is performed using Pearson and Spearman correlation coefficients together with classification of predictions based on an error threshold. **Results.** The findings show that the GTE model demonstrates the best overall performance across combined metrics, while MiniLM provides an optimal balance between accuracy and computational efficiency. They also indicate that high correlation scores do not always correspond to a high percentage of correct predictions, revealing limitations of traditional evaluation approaches. In addition, a systematic tendency of models to overestimate similarity in the lower range is observed. It is further found that cross-lingual pairs may achieve higher accuracy under threshold-based evaluation despite lower correlation values, indicating differences in model behavior depending on the evaluation metric. **Conclusions.** The study demonstrates the effectiveness of a combined evaluation approach for semantic similarity models, which enables a more comprehensive assessment of their real–world performance. The results confirm the importance of considering language-specific characteristics and practical task requirements when selecting models and evaluation methods, and highlight the role of comprehensive evaluation strategies.

Keywords: semantic similarity; transformer models; model evaluation; Pearson correlation coefficient; Spearman rank correlation coefficient; English language; Ukrainian language; cross-lingual similarity; practical model performance.

1. Introduction

One of the key challenges in the field of natural language processing (NLP) is the accurate assessment of semantic similarity between two texts, commonly formalized as the semantic textual similarity (STS) task. Effective solutions to this problem provide the foundation for a wide range of applied NLP tasks.

Substantial progress in semantic modeling has been achieved through methods for learning vector representations (embeddings) of linguistic units, most notably Word2Vec [13] and GloVe [15]. However, these methods share an inherent limitation: they primarily operate at the level of individual words and therefore fail to adequately capture context-dependent variations in meaning. This was followed by the development of transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers), which generate significantly higher-quality contextualized representations [7]. This progress, in turn, facilitated the emergence of models specifically designed for text comparison tasks and optimized for more accurate semantic similarity assessment.

The fundamental role of the STS task within NLP is reflected in its frequent use as a component of more complex applied systems. In particular, STS is widely used for automatic essay grading, plagiarism detection, question answering, and filtering text pairs based on semantic similarity. In this context, the accuracy and robustness of STS models directly determine the effectiveness of the corresponding applications.

At the same time, the performance of semantic similarity models may vary considerably depending on the task, dataset properties, and linguistic context. This variability highlights the importance of systematic analysis of existing models and evaluation methodologies, as well as a comprehensive investigation of the factors that influence their performance.

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

2. Analysis of sources and definition of the problem

Recent studies increasingly demonstrate that general-purpose transformer models do not exhibit consistently stable performance across different application scenarios. Their effectiveness depends on the domain, dataset composition, and the chosen evaluation metric [21, 20]. This indicates that even high-quality vector representations, when considered in isolation, do not guarantee stable performance in STS tasks. Therefore, comparative studies across different types of data are necessary to assess the practical suitability of such models.

Separately, the literature addresses the problem of evaluating the quality of model predictions in semantic similarity tasks [1, 11, 16]. Traditionally, particularly within the framework of SemEval-2017 Task 1 [4], Pearson's correlation coefficient is used to compare model predictions with human annotations. This metric enables the assessment of linear correspondence between predicted and ground-truth scores. However, its application has certain limitations, as identical coefficient values may correspond to different error distributions and do not always adequately reflect the practical effectiveness of a model.

As an alternative, Spearman's rank correlation coefficient [24] is often used, as it better captures monotonic and nonlinear relationships between scores. This metric has also been successfully applied in the evaluation of models on STS tasks [18]. However, it is not universal either: high rank consistency does not necessarily ensure effectiveness of the system in applied scenarios where it is critical to make threshold-based decisions, ranking decisions, or selection of text pairs.

For this reason, other evaluation metrics are being investigated, including normalized cumulative gain (nCG), normalized discounted cumulative gain (nDCG), precision, F1-score, and combined approaches [19]. However, existing results indicate that there is no universal metric for evaluating STS models, and the choice of a specific evaluation method should be determined by the properties of the task at hand. A similar conclusion is supported by findings in other domains where system performance monitoring, early error detection, and optimization with respect to speed and accuracy are essential [9, 10, 23].

Another under-researched problem is cross-linguistic semantic similarity, particularly for languages with different grammatical and syntactic structures, such as Ukrainian and English. Most existing studies focus on monolingual datasets or high-resource language pairs. Although recent approaches to constructing multilingual sentence embeddings have demonstrated promising results in zero-shot transfer settings [3], their performance remains limited for typologically distant languages, where a decrease in semantic alignment quality is still observed.

In this context, the Ukrainian language faces additional challenges due to the lack of standardized datasets, which complicates accurate benchmarking and fine-tuning of the models for natural language processing tasks, particularly STS [12, 14]. Although cross-lingual benchmarks have been proposed, such as those in SemEval-2017 Task 1 [4], most comparative studies continue to focus on high-resource languages, thereby limiting the evaluation of models for Ukrainian and constraining cross-lingual comparisons between English and Ukrainian.

Thus, two interrelated research gaps can be identified. First, there is no universal approach to evaluating the quality of semantic textual similarity (STS) models that adequately reflects their performance across different application scenarios. Second, the behavior of modern transformer-based models in cross-lingual settings, particularly for English–Ukrainian sentence pairs, remains underexplored. These gaps jointly motivate a comparative analysis of STS models on both monolingual and cross-lingual datasets using multiple evaluation approaches.

3. Research objectives and tasks

Given that the performance of STS models is strongly language-dependent, this study compares several models on English, Ukrainian, and English–Ukrainian sentence pairs. This approach enables the assessment of model stability as well as sensitivity to typological differences in grammar and syntax.

Since different correlation-based metrics capture different aspects of model behavior, the evaluation is conducted using both Pearson and Spearman correlation coefficients. This allows for identifying potential differences in performance assessment depending on the evaluation metric and provides a more comprehensive view of the agreement between model predictions and human annotations.

However, correlation analysis alone is insufficient for assessing the practical applicability of the models.

ISSN 3083-7715 (online) Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)
ISSN 3083-7650 (0135-1710) (print) Management Information System and Devices. 2026. No. 2 (189)

Therefore, it is complemented by classifying sentence pairs based on the magnitude of the error between predicted and ground-truth similarity scores. This makes it possible to analyze how the models handle sentence pairs with varying levels of semantic similarity and to identify cases in which they exhibit large deviations or, conversely, achieve high accuracy.

Accordingly, the purpose of this work is to evaluate and compare transformer-based models on monolingual and cross-lingual datasets of English and Ukrainian sentences, taking into account multiple approaches to semantic similarity assessment.

4. Materials and research methods

Selected models and their characteristics

To compare performance on sentence-level semantic similarity tasks, several transformer-based models were selected. These models differ in architectural characteristics, support a wide range of languages (including English and Ukrainian), and have demonstrated strong performance in prior studies. Table 1 presents the embedding dimensionality of each model, the number of supported languages, and their base architectures.

Table 1. *Parameters of the models under comparison*

Model	Embedding dimension	Supported languages	Base model
gte-multilingual-base (GTE)	768	75	GTE
LaBSE	768	110	BERT
paraphrase-multilingual-MiniLM-L12-v2 (MiniLM)	384	50	BERT
paraphrase-multilingual-mpnet-base-v2 (MPNet)	768	50	XLM-RoBERTa

The gte-multilingual-base (GTE) model is based on an encoder-only transformer architecture, supports 75 languages, and processes input sequences of up to 8192 tokens. A key distinguishing feature of GTE is its ability to generate both dense and sparse embeddings, which improves performance in retrieval and re-ranking tasks. As a result, GTE achieves state–of–the–art performance on text benchmarks while maintaining high computational efficiency [25].

The LaBSE model (Language-Agnostic BERT Sentence Embedding) is a multilingual extension of BERT that produces language-independent embeddings and ensures strong cross-lingual alignment [8]. Consequently, it demonstrates strong performance in cross-lingual retrieval tasks and supports over 100 languages. However, this design focus may lead to reduced performance on monolingual semantic similarity tasks.

The paraphrase-multilingual-MiniLM-L12-v2 (MiniLM) model is also based on a multilingual BERT-derived architecture and supports 50 languages. Through knowledge distillation, the model remains compact and computationally efficient while maintaining competitive performance relative to BERT-base [22]. However, the reduced dimensionality of its embeddings may limit accuracy on more complex semantic tasks.

The paraphrase-multilingual-mpnet-base-v2 (MPNet) model is based on XLM-RoBERTa and supports 50 languages. It combines masked and permutation-based language modeling strategies, enabling more effective modeling of token dependencies [6]. This results in improved performance compared to BERT and XLNet across a wide range of tasks. Its main limitation is relatively high computational cost due to the higher embedding dimensionality.

All selected models were fine-tuned using the Sentence-Transformers framework and adapted for semantic similarity tasks.

Semantic textual similarity benchmark datasets

To compare model performance, the STS-B (Semantic Textual Similarity Benchmark) dataset [4] is selected as the evaluation benchmark. It contains sentence pairs from various domains, including news headlines, image captions, and natural language inference tasks. Each pair is annotated with human-assigned similarity scores ranging from 0 (completely dissimilar sentences) to 5 (semantically equivalent sentences).

In the experiments, a normalized version of the English dataset is used, in which similarity scores are rescaled to the interval [0, 1]. The dataset consists of 5749 sentence pairs, each associated with a corresponding similarity score.

A Ukrainian version of the dataset is also created [2], where the original data are translated using state–of–the–art neural machine translation systems to ensure semantic consistency. Although human translation may preserve certain linguistic nuances more accurately in some cases, machine translation is chosen due to

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

its accessibility, efficiency, and high quality for general-domain texts. A comparison of human and machine translation results is left for future work investigating the impact of translation quality on semantic similarity assessment.

Evaluation metrics

Cosine similarity is used to compute the similarity between vector representations of sentences, and the results are evaluated using the Pearson and Spearman correlation coefficients, which measure the correlation between predicted similarity scores and ground-truth labels.

Cosine similarity measures the similarity between two vectors in a high-dimensional space by computing the cosine of the angle between them:

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

where A and B are vector representations of two sentences, with $\|A\|$ and $\|B\|$ denoting their L2-norms.

To enable comparison between cosine similarity values, which lie in the interval [–1, 1], and ground-truth labels, min-max normalization is applied to rescale the values to the interval [0, 1]:

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

where $\min(x)$ and $\max(x)$ are the minimum and maximum cosine similarity values. Thus, the minimum value is mapped to 0, the maximum to 1, and all intermediate values are linearly scaled within this range.

The Pearson correlation coefficient (r) measures the linear relationship between predicted and ground-truth values and is defined as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}, \quad (3)$$

where x_i and y_i are the predicted and true similarity scores, with $\bar{x}$ and $\bar{y}$ denoting their mean values.

To address the limitations of Pearson correlation in capturing non-linear relationships, Spearman's rank correlation coefficient (ρ) is also used, as it evaluates the monotonic relationship between ranked values:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (4)$$

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

Additionally, an error-based evaluation is performed by classifying sentence pairs as correctly or incorrectly predicted based on a predefined threshold (0.3 in this study). Correct predictions correspond to small deviations between predicted and ground-truth scores, whereas incorrect predictions correspond to large deviations. This classification enables the identification of patterns in which models most frequently achieve high or low accuracy and helps reveal systematic biases, such as consistent overestimation or underestimation of semantic similarity.

5. Research results

Correlation-based performance comparison

At the first stage of the study, model performance was evaluated using correlation-based metrics to assess how well the models align with benchmark sentence-level semantic similarity scores. Pearson and Spearman correlation coefficients were computed for different language pairs: EN-EN (monolingual English), UK-UK (monolingual Ukrainian), and EN-UK (cross–lingual English–Ukrainian). The results are presented in Table 2, with values expressed as percentages ($r \times 100$ and $\rho \times 100$).

Table 2. *Pearson and Spearman correlation coefficients for each model*

Model	EN-EN pairs		UK-UK pairs		EN-UK pairs		Averages	
	r	ρ	r	ρ	r	ρ	r	ρ
GTE	87.4	87.4	85.1	84.1	80.6	78.1	**84.4**	**83.2**
LaBSE	74.9	73.6	74.4	73.1	71.7	69.9	73.7	72.2
MiniLM	85.1	83.9	81.6	80.1	78.1	74.6	81.6	79.5
MPNet	86.3	85.9	83.1	82.1	80.6	77.8	**83.3**	**81.9**

Based on the average correlation values computed across different datasets, the GTE and MPNet models demonstrated the best performance, achieving scores exceeding 80% for both coefficients. This indicates their

ISSN 3083-7715 (online) Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)
ISSN 3083-7650 (0135-1710) (print) Management Information System and Devices. 2026. No. 2 (189)

ability to generate embeddings that effectively capture both semantic content and syntactic properties in both monolingual and cross-lingual sentence pairs.

The MiniLM model ranked third, lagging by up to 3% in Pearson correlation and up to 4% in Spearman correlation. At the same time, due to its compact architecture, it provides an optimal balance between accuracy and computational cost in practical applications.

The LaBSE model demonstrated the lowest performance, with a gap of 12-13% for both coefficients compared to the best-performing models. This reflects the so-called "curse of multilinguality", in which the inclusion of a large number of languages leads to reduced performance under fixed model capacity [5]. In other words, broader language coverage is achieved at the cost of less specialized optimization for individual languages. At the same time, due to its language-agnostic embeddings, LaBSE exhibits relatively stable performance across different language pairs (varying within 1–7%). If stability across both monolingual and cross–lingual settings is prioritized, LaBSE may be preferable, as other models exhibit substantially greater variability, ranging from 3% to 12% depending on the language combination.

Overall, Spearman correlation was lower than Pearson correlation, indicating stronger linear agreement between predictions and ground-truth labels, while rank order is not fully preserved. Nevertheless, the overall conclusions regarding model performance remain consistent, suggesting that both metrics lead to similar conclusions depending on the evaluation context.

A comparison of model performance across monolingual and cross-lingual settings shows that monolingual pairs yield higher scores for both correlation metrics and exhibit smaller differences between them. This result is expected, as cross-lingual tasks are more challenging due to grammatical and semantic differences, as well as cultural context, which require more specialized model training. At the same time, general-purpose models still achieve strong results in such settings, although the performance gap remains noticeable.

Based on these evaluations, an initial ranking of the models was established, enabling the selection of an optimal model depending on priorities such as accuracy, computational efficiency, or stability across languages. However, correlation-based results do not always guarantee similar performance in real-world applications. Therefore, to obtain a more detailed assessment of practical applicability, an additional analysis is conducted by classifying semantic similarity predictions for sentence pairs.

Detailed analysis of semantic similarity predictions for sentence pairs. The problem of score scaling

To further investigate model behavior, predictions are compared with human-assigned labels. At the initial stage, min-max scaling is applied for data preprocessing, with the minimum and maximum values defined according to the theoretical range of cosine similarity. Figure 1 presents the distributions of ground-truth scores and model predictions obtained from vector representations of the three models with the highest correlation performance (GTE, MPNet, and MiniLM) for English and Ukrainian sentence pairs.

The results indicate a substantial difference between the distributions of ground-truth scores and model predictions. For all three models, predicted values are shifted toward higher similarity scores, with insufficient differentiation in the lower and mid-range values. Most predictions are concentrated within a narrow interval from 0.7 to 1.0, while low scores (0–0.4) are nearly absent. In contrast, ground-truth labels are more evenly distributed across the entire range, from low to high similarity.

This behavior is consistent across all models and language pairs and complicates both overall performance evaluation and the analysis of language effects, as the narrow distribution of predictions limits the ability to identify subtle differences in model accuracy.

Based on these observations, an adjusted scaling approach is proposed in which the actual minimum and maximum prediction values for a given model and dataset are used instead of the theoretical bounds of cosine similarity. This transformation distributes predicted scores more evenly across the entire scale, providing a clearer representation of differences between sentence pairs with low, medium, and high semantic similarity.

Figure 2 illustrates the results after applying the adjusted scaling.

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

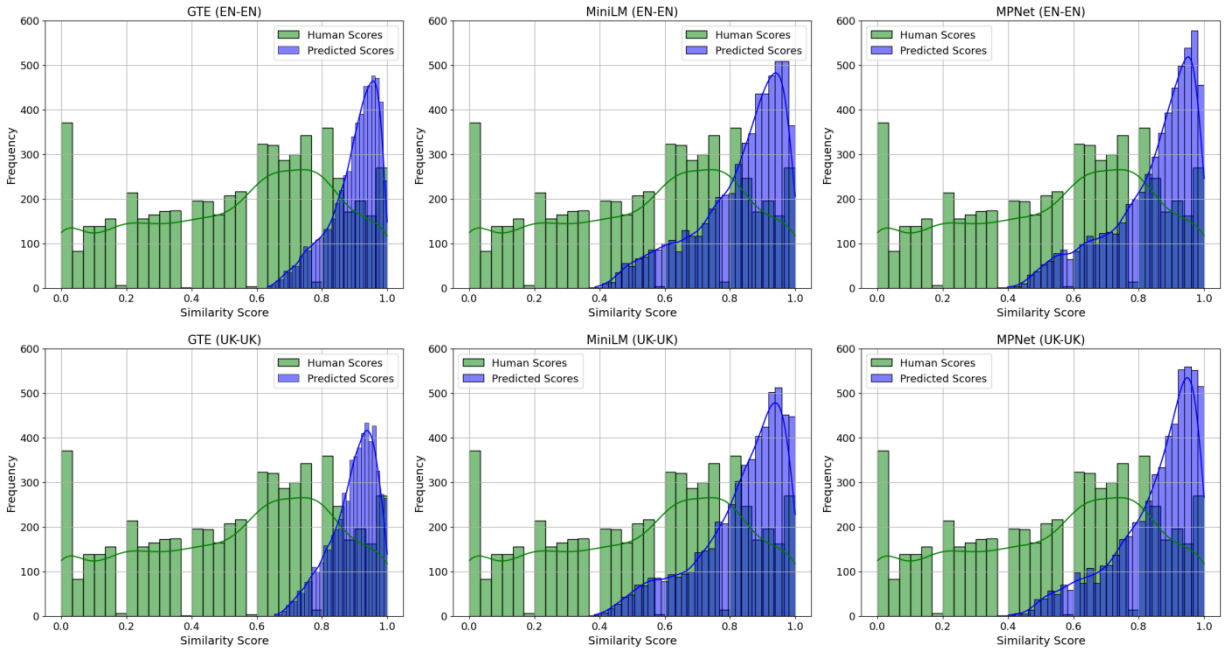

Fig. 1. Comparison of the distributions of predicted and gold semantic similarity scores for monolingual sentence pairs

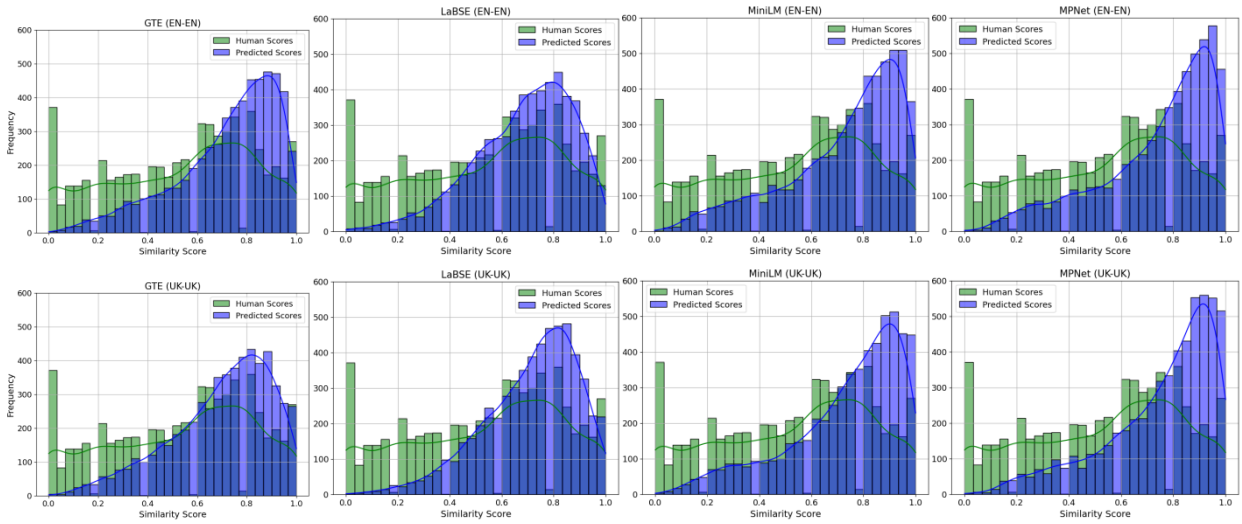

Fig. 2. Comparison of the distributions of predicted and gold semantic similarity scores for monolingual sentence pairs after scaling correction

The plots demonstrate improved alignment between model predictions and ground-truth labels, particularly in the lower and middle ranges, where substantially more predictions are observed than with the initial scaling. This transformation does not alter the relative ordering of sentence pairs by similarity, but only modifies the mapping of predictions onto the target scale. Thus, the adjusted scaling serves as a calibration mechanism that improves the interpretability of the results without affecting the actual accuracy of the models.

Overall, the analysis of distributions highlights the need to adapt scaling to the observed range of prediction values, which is an important step in assessing the practical applicability of the models.

Comparison of model suitability for practical use based on classification of semantic similarity predictions

Figure 3 presents the results of the classification of model predictions for all evaluated models. In the plots, green corresponds to correctly predicted

ISSN 3083-7715 (online) *Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)*
ISSN 3083-7650 (0135-1710) (print) *Management Information System and Devices. 2026. No. 2 (189)*

similarity scores, i.e., predictions that fall within a threshold of 0.3 from the corresponding ground-truth labels. Red indicates predictions with substantial deviations, including both overestimation and underestimation, where the model assigns higher or lower similarity scores than those provided by human annotators. The dashed diagonal line represents the ideal case in which predicted scores exactly match the ground-truth labels. For high-quality models, most predictions are expected to lie close to this line.

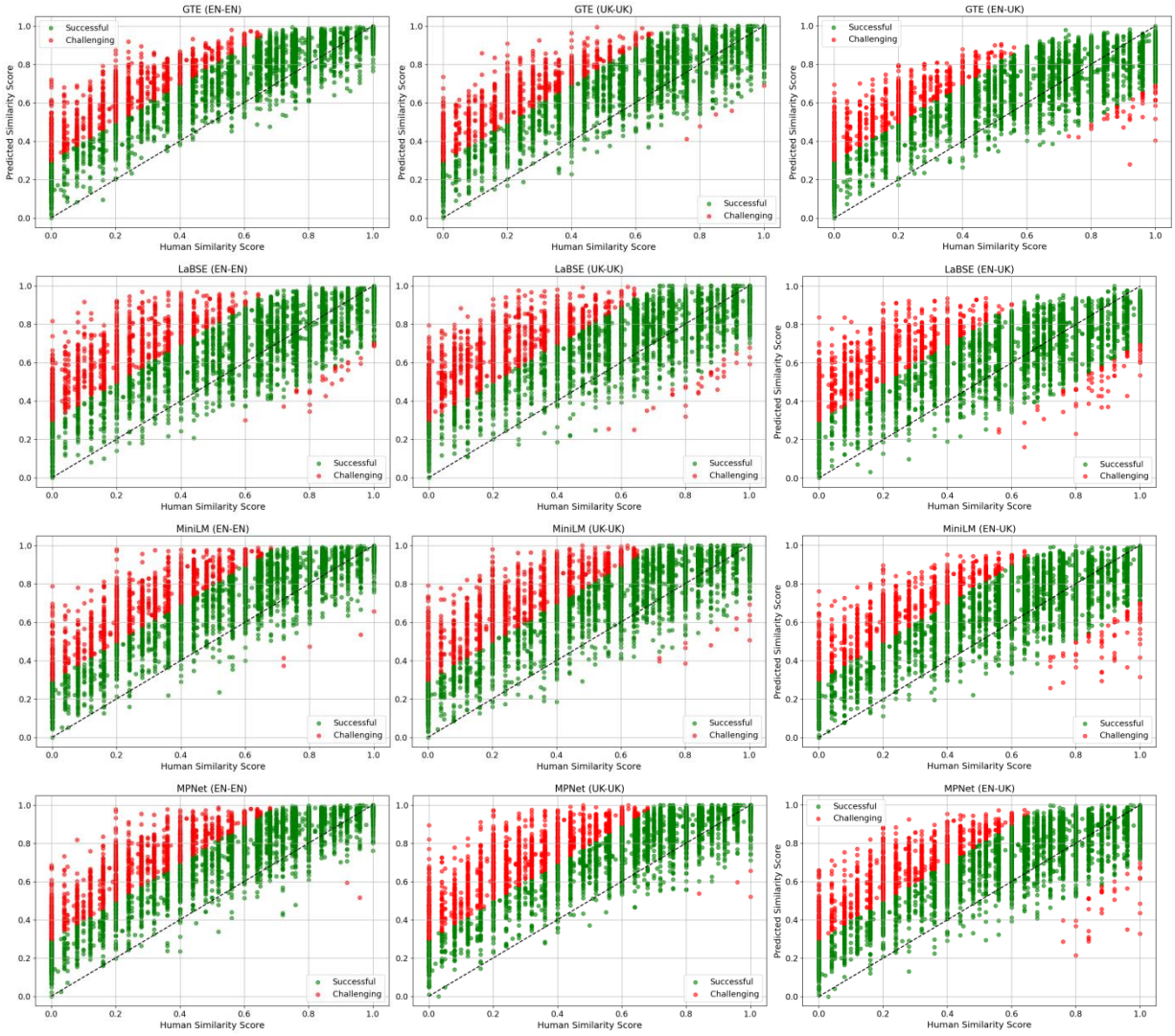

Fig. 3. Comparison of predicted and gold semantic similarity scores for sentence pairs

To complement the graphical analysis, Table 3 reports the percentage of correctly predicted similarity scores for each model, separately for monolingual (EN-EN and UK-UK) and cross-lingual (EN-UK) sentence pairs.

Table 3. *Percentage of correctly predicted sentence pair similarity for each model*

Model	EN-EN pairs	UK-UK pairs	EN-UK pairs	Averages
GTE	**78.4**	**80.1**	**82.3**	**80.3**
LaBSE	76.6	72.7	77.6	75.6
MiniLM	**77.7**	**75.3**	**80.4**	**77.8**
MPNet	75.9	71.4	78.3	75.2

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

Performance on monolingual sentence pairs

As shown in Table 3, the ranking of models based on the percentage of correct predictions differs slightly from the ranking obtained using correlation-based metrics. The GTE and MiniLM models maintain high performance, while LaBSE and MPNet exhibit lower average values. Notably, despite achieving high correlation coefficients, MPNet shows the lowest percentage of correct predictions. This indicates that high correlation does not necessarily imply better agreement between predicted similarity scores and ground-truth values: a model may preserve the correct ordering of sentence pairs while making larger errors in the actual numerical predictions.

The analysis of the distribution plots (Figure 2) and the comparison of predictions with ground-truth values (Figure 3) show that, for all models, the largest number of correct predictions occurs in the medium and high similarity ranges, as reflected by peaks in the histograms within the interval 0.6–0.9. At the same time, the highest number of incorrect predictions is observed for sentence pairs with low human-assigned similarity scores. In this range, incorrect predictions are predominantly located above the diagonal, indicating systematic overestimation of similarity. Thus, the main difficulty for the models lies in their limited ability to distinguish between dissimilar sentence pairs.

The GTE model demonstrates the best agreement with human annotations. Although its prediction distribution is shifted toward higher values, it exhibits a relatively smooth shape and partially overlaps with the ground-truth distribution in the range 0.4–0.8. A substantial proportion of predictions lies close to the diagonal, and overestimation in the low-similarity range is less pronounced than in the other models. Cases of underestimation for highly similar sentence pairs are rare for Ukrainian and even rarer for English. Therefore, GTE not only preserves the relative ranking of sentence pairs, as reflected by its high correlation values, but also achieves the closest alignment between predicted scores and human annotations.

The MiniLM model exhibits a similar distribution pattern; however, its predictions are more concentrated in the upper part of the scale (0.8–1.0). In the medium similarity range, they align well with the ground-truth values, whereas in the high-similarity range, underestimation occurs more frequently, particularly for Ukrainian sentence pairs. In the low-similarity range, a tendency toward overestimation is also observed, although it is less pronounced than in the other models. Overall, these results confirm the competitiveness of MiniLM, especially given its compact architecture.

The LaBSE model is characterized by a less smooth distribution, with peak values in the range 0.7–0.9 and a near absence of low scores. Consequently, the largest number of errors occurs for sentence pairs with low similarity, where overestimation is predominant. In addition, cases of underestimation in the high-similarity range are more frequent than in the previously discussed models. Although these deviations are generally not large, as indicated by the relatively close proximity of predictions to the diagonal, the model overall provides stable but less accurate predictions compared to GTE and MiniLM in this evaluation.

The distribution of MPNet predictions is among the most compressed across all models, with a strong concentration in the narrow interval 0.8–1.0. The most pronounced pattern is the overestimation of similarity for sentence pairs with low ground-truth scores, where predicted similarity scores are substantially higher. At the same time, in the medium and high similarity ranges, the model demonstrates relatively good agreement with human annotations, and cases of underestimation are comparatively rare. Thus, while MPNet effectively preserves the relative ranking of sentence pairs (as reflected in its high correlation scores), systematic overestimation in the low-similarity range results in a lower percentage of correct predictions.

Effect of language on prediction quality

Separately, it is important to consider the effect of language on model prediction performance. According to the results presented in Table 3, English sentence pairs generally exhibit a higher percentage of correct predictions compared to Ukrainian pairs. The largest differences, ranging from 5–6%, are observed for models with lower average performance, such as LaBSE and MPNet, whereas for GTE and MiniLM the difference between languages is minimal (2–3%). This is due to the more complex morphology and greater syntactic variability of Ukrainian, which make accurate prediction of exact similarity scores more challenging and amplify the limitations of models with lower overall performance.

Interestingly, high correlation values for English sentence pairs do not always correspond to higher accuracy in terms of exact similarity scores. For example, in the case of GTE, the relative ranking of sentence pairs is better preserved for English, whereas the percentage of predictions within the specified threshold is slightly higher for Ukrainian.

Overall, the results indicate that models make more errors on Ukrainian sentence pairs. In particular,

ISSN 3083-7715 (online) *Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)*
ISSN 3083-7650 (0135-1710) (print) *Management Information System and Devices. 2026. No. 2 (189)*

underestimation is more common for highly similar pairs, whereas overestimation is less language-dependent, as it is also prevalent for English pairs. These findings highlight the need to account for language-specific characteristics when evaluating the practical applicability of semantic similarity models.

Performance on cross-language sentence pairs

To evaluate how model behavior changes under more challenging conditions, predictions for English–Ukrainian sentence pairs were analyzed. Figure 4 presents the distribution of predictions for these cross-lingual pairs, complementing the comparison of predicted and ground-truth scores shown in Figure 3.

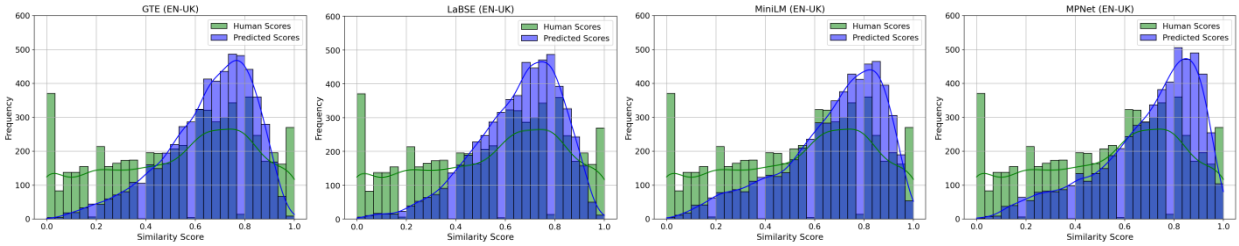

Fig. 4. Comparison of the distributions of predicted and gold semantic similarity scores for cross-lingual sentence pairs

According to the results in Table 3, all models achieve a higher percentage of correct predictions for cross-lingual pairs than for monolingual ones. GTE once again achieves the highest values, with an average increase of approximately 4% compared to English and Ukrainian pairs. A similar pattern is observed for MiniLM, with an improvement of about 5%. LaBSE and MPNet also show slightly better results on cross-lingual data, though the increase is more modest, at around 4%.

The plots in Figure 3 indicate that the overall pattern of predictions for cross-lingual pairs is similar to that observed for monolingual data. Most correct predictions occur in the medium and high similarity ranges, while the most challenging cases correspond to sentence pairs with low ground-truth scores. However, underestimation is more frequent for highly similar pairs. At the same time, performance in the low-similarity range improves: overestimation occurs less frequently, and such predictions lie closer to the diagonal.

The histograms in Figure 4 further confirm the higher percentage of correct predictions. Although predicted scores remain skewed toward higher values, most predictions are concentrated near the peak of the ground-truth distribution (0.6–0.9). This suggests that the improved performance on cross-lingual pairs does not imply the absence of distortions; rather, it indicates that models handle sentence pairs with medium and high semantic similarity more effectively, reflecting patterns observed in human annotations.

Thus, contrary to the expectation that comparing sentences across different languages is more difficult due to the need to construct a shared semantic representation for English and Ukrainian, the classification results indicate higher accuracy. At the same time, this finding contrasts with the correlation analysis, where model performance on cross-lingual pairs is lower than on monolingual data. In other words, for English–Ukrainian pairs, models are less effective at preserving the overall ranking of similarity, while their predicted scores more often fall within the acceptable threshold. This discrepancy highlights the importance of analyzing not only aggregate statistical metrics but also the nature of prediction errors.

6. Discussion of results

The results demonstrate that the evaluation of semantic similarity models cannot be limited to correlation-based metrics alone. Although Pearson and Spearman coefficients provide a general measure of agreement between model predictions and human annotations, they do not always reflect the practical applicability of models in real-world systems, where decisions are often made based on predefined thresholds.

Prediction classification thus serves as an important complement to correlation analysis: even with high correlation values, a model may be less suitable for practical use due to systematic overestimation or underestimation of similarity. This effect is particularly evident in the low-similarity range, where model behavior is less predictable because similarity scores are frequently overestimated for both monolingual and cross-lingual pairs. Therefore, evaluation based solely

Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189) *ISSN 3083-7715 (online)*
Management Information System and Devices. 2026. No. 2 (189) *ISSN 3083-7650 (0135-1710) (print)*

on correlation-based metrics fails to capture important aspects of model behavior in practical scenarios.

Given these findings, model selection should be guided by the specific application context. A clear example of this is GTE, which achieves the best results in terms of both correlation metrics and the percentage of correct predictions and is therefore the most versatile model, suitable for systems requiring high overall accuracy and reliability.

A comparable pattern is observed for the MiniLM model, which demonstrates slightly lower accuracy while maintaining a consistently high level of practical applicability. This makes it a strong candidate as a trade-off between performance and computational efficiency. It is particularly well suited for large-scale production systems, where processing speed, latency, and scalability are critical.

The results for MPNet illustrate why evaluation based solely on correlation metrics is insufficient. Although this model achieves one of the highest correlation scores, the percentage of correct predictions is lower. At the same time, this characteristic makes MPNet well suited for tasks where relative ranking is the primary objective, such as information retrieval or re-ranking, but less suitable for systems that rely on strict thresholds, such as plagiarism detection.

Another important consideration is model stability across languages. LaBSE, despite lower average performance, demonstrates relatively stable behavior across different languages and their combinations. This property is particularly important for systems in which consistent performance across languages is more critical than achieving maximum accuracy for a single language, such as cross-lingual retrieval or multilingual applications.

The analysis of cross-lingual pairs reveals a distinct pattern of model behavior: according to correlation metrics, such pairs appear more challenging, whereas threshold-based classification shows improved results. In practical systems, it is often more important to ensure sufficiently accurate separation between similar and dissimilar sentence pairs than to perfectly preserve the global ranking of all examples. Thus, in cross-lingual scenarios, models may perform comparably, or in some cases even more effectively, when evaluated in terms of threshold-based accuracy.

7. Conclusions

This study demonstrates the continued relevance of the semantic textual similarity (STS) task for modern natural language processing systems, particularly in multilingual settings, where no universal evaluation approach currently exists. Previous research shows that evaluating transformer-based models requires a comprehensive framework that goes beyond correlation-based metrics alone. A combined evaluation approach is therefore employed, incorporating both correlation-based metrics and classification-based measures of prediction accuracy. Such an approach provides a more complete assessment of model performance across different linguistic scenarios, capturing not only the agreement between predicted and human-assigned scores but also the practical applicability of models in real-world tasks.

Building on this evaluation framework, the comparison of four models on English, Ukrainian, and English-Ukrainian sentence pairs revealed significant differences in performance. The GTE model demonstrated the highest overall effectiveness, achieving balanced results across the combined evaluation framework. MiniLM proved to be a competitive alternative, offering an optimal balance between accuracy and computational efficiency. In contrast, MPNet and LaBSE showed that high correlation scores or cross-lingual stability alone do not necessarily guarantee superior performance in applied scenarios.

A more detailed analysis of predictions revealed important patterns in model behavior. In particular, a systematic bias toward higher similarity scores was observed, and correlation-based results did not always align with classification-based evaluations. These findings highlight the importance of a combined evaluation approach that integrates multiple metrics to provide a more reliable assessment of model performance.

These findings have both theoretical and practical significance. They contribute to a deeper understanding of transformer model behavior under different linguistic conditions and emphasize the need to adapt evaluation methodologies for cross-lingual scenarios. From a practical perspective, they enable the informed application of STS models across a range of tasks, such as information retrieval, classification, automated text evaluation, and other applied scenarios.

Overall, this study highlights the continued relevance of the problem addressed and demonstrates the effectiveness of comprehensive evaluation approaches for transformer models. Such approaches enable the selection of models that can simultaneously achieve high accuracy in multilingual environments and meet the requirements of specific application scenarios.

ISSN 3083-7715 (online) Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)
ISSN 3083-7650 (0135-1710) (print) Management Information System and Devices. 2026. No. 2 (189)

Finally, a promising direction for future research is the development of new evaluation metrics for semantic similarity models that go beyond correlation coefficients and account for practical applicability. In particular, combining correlation-based and ranking-based metrics, along with a more detailed analysis of error distributions, may provide deeper insights, especially considering the observed tendency of models to overestimate similarity in the low-similarity range. Another important direction is the adaptation of models to domain-specific applications, such as legal or medical texts, where additional fine-tuning can improve accuracy, robustness, and the ability to handle highly specialized data. These directions can contribute to the development of practical recommendations for model selection, evaluation strategies, and deployment in real-world systems.

The results of this research were obtained under the international research project INITIATE under the grant No. 101136775-HORIZON-WIDERA-2023-ACCESS-03.

Conflict of Interest

The authors declare that they have no conflicts of interest, including financial, personal, copyright, or any other conflicts that could influence the research or the results published in this article.

Funding

The study was conducted without financial support.

Data Availability

The manuscript contains data in the form of electronic supplementary material.

Use of Artificial Intelligence

During the preparation of this article, the GPT-5 mini model was used for technical proofreading of the text, including checking spelling, punctuation, and style. All changes suggested by this tool were critically reviewed, edited, and verified by the authors. The main content, scientific ideas, and conclusions belong exclusively to the authors and have not been influenced by any third parties.

References

1. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W. (2013), "SEM 2013 shared task: Semantic Textual Similarity", *Joint Conference on Lexical and Computational Semantics*, Vol. 1, pp. 32–43.
2. anikol12 (2026), "STSB-UK", *Huggingface.co*. Available at: https://huggingface.co/datasets/anikol12/STSB-UK (Accessed 1 Apr. 2026).
3. Artetxe, M., Schwenk, H. (2019), "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond", *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 597–610. DOI: https://doi.org/10.1162/tacl_a_00288
4. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L. (2017), "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation", *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*. DOI: https://doi.org/10.18653/v1/S17-2001
5. Chang, T.A., Arnett, C., Tu, Z., Bergen, B. (2024), "When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096. DOI: https://doi.org/10.18653/v1/2024.emnlp-main.236
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020), "Unsupervised Cross-lingual Representation Learning at Scale", *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. DOI: https://doi.org/10.18653/v1/2020.acl-main.747
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 4171–4186. DOI: https://doi.org/10.18653/v1/n19-1423
8. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W. (2022), "Language-agnostic BERT Sentence Embedding", *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. DOI: https://doi.org/10.18653/v1/2022.acl-long.62
9. Gorokhovatskyi, V., Tvoroshenko, I., Kobylin, O., Vlasenko, N. (2023), "Search for Visual Objects by Request in the Form of a Cluster Representation for the Structural Image Description", *Advances in Electrical and Electronic Engineering*, Vol. 21, No. 1. DOI: https://doi.org/10.15598/aeee.v21i1.4661

10. Kobylin, I., Nikolaichuk, A. (2024), "Monitoring and Diagnosing Faults in Online Mode Using Time Series Data", *Information Processing Systems*, No. 3(178), pp. 27–32. DOI: https://doi.org/10.30748/soi.2024.178.03
11. Kour, G., Ackerman, S., Farchi, E., Raz, O., Carmeli, B., Tavor, A.A. (2022), "Measuring the Measuring Tools: An Automatic Evaluation of Semantic Metrics for Text Corpora", *Proceedings of the Workshop on General Evaluation of NLP Models*. DOI: https://doi.org/10.18653/v1/2022.gem-1.35
12. Maksymenko, D., Turuta, O. (2024), "Interpretable Conversation Routing via the Latent Embeddings Approach", *Computation*, Vol. 12, No. 12, p. 237. DOI: https://doi.org/10.3390/computation12120237
13. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013), "Efficient Estimation of Word Representations in Vector Space", *arXiv*. DOI: https://doi.org/10.48550/arXiv.1301.3781
14. Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., Tytarenko, S. (2022), "Ukrainian News Corpus as Text Classification Benchmark", *Communications in Computer and Information Science*, pp. 550–559. DOI: https://doi.org/10.1007/978-3-031-14841-5_37
15. Pennington, J., Socher, R., Manning, C. (2014), "Glove: Global Vectors for Word Representation", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. DOI: https://doi.org/10.3115/v1/d14-1162
16. Poliak, A. (2020), "A Survey on Recognizing Textual Entailment as an NLP Evaluation", *Proceedings of the Workshop on Evaluation for NLP*. DOI: https://doi.org/10.18653/v1/2020.eval4nlp-1.10
17. Reimers, N., Gurevych, I. (2019), "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. DOI: https://doi.org/10.18653/v1/d19-1410
18. Reimers, N., Gurevych, I. (2020), "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.365
19. Reimers, N., Beyer, P., Gurevych, I. (2016), "Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity", *Proceedings of the International Conference on Computational Linguistics*, pp. 87–96.
20. Mehri, S., Eric, M., Hakkani-Tur, D. (2020), "DialoGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue", *arXiv*. DOI: https://doi.org/10.48550/arXiv.2009.13570
21. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. (2018), "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", *Proceedings of the Workshop on Evaluation of NLP Systems*. DOI: https://doi.org/10.18653/v1/W18-5446
22. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M. (2020), "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers", *arXiv*. DOI: https://doi.org/10.48550/arXiv.2002.10957
23. Yakovleva, O., Kovtunenko, A., Liubchenko, V., Honcharenko, V., Kobylin, O. (2023), "Face Detection for Video Surveillance-Based Security System", *Proceedings of the International Conference on Computational Linguistics and Intelligent Systems*, Vol. III, pp. 69–86.
24. Zesch, T. (2010), "Study of Semantic Relatedness of Words Using Collaboratively Constructed Semantic Resources", *Dissertation*, p. 130. Available at: https://d-nb.info/1001286782 (Accessed 1 Apr. 2026).
25. Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., Zhang, M. (2024), "mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1393–1412. DOI: https://doi.org/10.18653/v1/2024.emnlp-industry.103

Received (Надійшла) 06.03.2026
Accepted for publication (Прийнята до друку) 09.04.2026
Publication date (Дата публікації) 29.05.2026

Відомості про авторів / About the Authors

Ніколайчук Анна Ігорівна – Харківський національний університет радіоелектроніки, здобувач другого рівня вищої освіти, Харків, Україна;
Anna Nikolaichuk – Kharkiv National University of Radio Electronics, Master's student, Kharkiv, Ukraine;
e-mail: anna.nikolaichuk@nure.ua
ORCID ID: https://orcid.org/0009-0001-8643-332X

ISSN 3083-7715 (online) *Автоматизовані системи управління та прилади автоматики. 2026. № 2 (189)*
ISSN 3083-7650 (0135-1710) (print) *Management Information System and Devices. 2026. No. 2 (189)*

Кобилін Олег Анатолійович – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, завідувач кафедри інформатики, Харків, Україна;
Oleg Kobylin – PhD (Technical Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Head of the Department of Informatics, Kharkiv, Ukraine;
e-mail: oleg.kobylin@nure.ua
ORCID ID: https://orcid.org/0000-0003-0834-0475

Кобилін Ілля Олегович – кандидат технічних наук, Харківський національний університет радіоелектроніки, старший викладач кафедри інформатики, Харків, Україна;
Ilya Kobylin – PhD (Technical Sciences), Kharkiv National University of Radio Electronics, Senior Professor of the Department of Informatics, Kharkiv, Ukraine;
e-mail: ilya.kobylin@nure.ua
ORCID ID: https://orcid.org/0000-0002-4552-9616

Путятіна Олександра Євгенівна – кандидат технічних наук, Харківський національний університет радіоелектроніки, старший викладач кафедри інформатики, Харків, Україна;
Oleksandra Putyatina – PhD (Technical Sciences), Kharkiv National University of Radio Electronics, Senior Professor of the Department of Informatics, Kharkiv, Ukraine;
e-mail: oleksandra.putiatina@nure.ua
ORCID ID:https://orcid.org/0000-0003-4853-7125

КОМПЛЕКСНЕ ОЦІНЮВАННЯ ТРАНСФОРМЕРНИХ МОДЕЛЕЙ ДЛЯ ЗАДАЧІ СЕМАНТИЧНОЇ ПОДІБНОСТІ РЕЧЕНЬ ДЛЯ АНГЛІЙСЬКОЇ ТА УКРАЇНСЬКОЇ МОВ

Предметом дослідження є трансформерні моделі семантичної подібності речень і підходи до їх оцінювання в одномовних і міжмовних сценаріях для англійської та української мов. **Мета роботи** – оцінити й порівняти ефективність трансформерних моделей на англомовних, українськомовних і англо-українських парах речень з огляду на різні підходи до оцінювання їх якості та практичної придатності. Зважаючи на окреслену мету, необхідно було виконати такі **завдання**: порівняти моделі на одномовних і міжмовних наборах даних; проаналізувати результати за коефіцієнтами Пірсона та Спірмена; оцінити практичну придатність моделей способом класифікації прогнозів за величиною похибки; дослідити вплив мовного чинника на точність моделей. **Методи**. У дослідженні використано трансформерні моделі GTE, LaBSE, MiniLM та MPNet і набір даних STS-B, його українськомовну й англо-українську версії. Для обчислення подібності застосовано косинусну міру, а результати оцінено за допомогою коефіцієнтів кореляції Пірсона та Спірмена й класифікації прогнозів за порогом похибки. **Результати дослідження**. Визначено, що модель GTE демонструє найкращу загальну ефективність за сукупністю метрик, а MiniLM забезпечує оптимальний баланс між точністю та обчислювальними витратами. З'ясовано, що високі значення кореляції не завжди відповідають високій частці коректних прогнозів, що вказує на обмеженість традиційних підходів оцінювання. Встановлено систематичну тенденцію моделей до переоцінювання подібності в низькому діапазоні, а також виявлено, що міжмовні пари можуть демонструвати вищу точність за пороговими оцінками, незважаючи на нижчі значення кореляції, що свідчить про різну поведінку моделей залежно від типу метрики. **Висновки**. Обґрунтовано доцільність використання комбінованого підходу для оцінювання моделей семантичної подібності, який дає змогу більш повно відтворити їх реальну ефективність. Досягнуті результати підтверджують необхідність врахування мовної специфіки та практичних вимог задач у виборі моделей і підходів до їх оцінювання та наголошують на важливості переходу до комплексних підходів оцінювання.

Ключові слова: семантична подібність; трансформерні моделі; оцінювання моделей; коефіцієнт кореляції Пірсона; коефіцієнт кореляції рангу Спірмена; англійська мова; українська мова; міжмовна подібність; практична ефективність моделей.

Бібліографічні описи / Bibliographic descriptions

Ніколайчук А. І., Кобилін О. А., Кобилін І. О., Путятіна О. Є. Комплексне оцінювання трансформерних моделей для задачі семантичної подібності речень для англійської та української мов. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 284–296. DOI: https://doi.org/10.30837/0135-1710.2026.189.284

Nikolaichuk, A., Kobylin, O., Kobylin, I., Putyatina, O. (2026), "A comprehensive evaluation of transformer models for sentence-level semantic similarity in English and Ukrainian", *Management Information System and Devices*, No. 2 (189), P. 284–296. DOI: https://doi.org/10.30837/0135-1710.2026.189.284