A Graph-Based Model for Semantic Textual Similarity Measurement

Van-Tan Bui[1], Quang-Minh Nguyen[2], and Van-Vinh Nguyen[2]

[1] University of Economic and Technical Industries, Vietnam
[2] University of Engineering and Technology, Vietnam National University, Vietnam

Abstract. Measuring semantic similarity between sentence pairs is a fundamental problem in natural language processing with applications in various domains, including machine translation, speech recognition, automatic question answering, and text summarization. Despite its significance, accurately assessing semantic similarity remains a challenging task, particularly for underrepresented languages such as Vietnamese. Existing methods have yet to fully leverage the unique linguistic characteristics of Vietnamese for semantic similarity measurement. To address this limitation, we propose GBNet-STS (Graph-Based Network for Semantic Textual Similarity), a novel framework for measuring the semantic similarity of Vietnamese sentence pairs. GBNet-STS integrates lexical-grammatical similarity scores and distributional semantic similarity scores within a multi-layered graph-based model. By capturing different semantic perspectives through multiple interconnected layers, our approach provides a more comprehensive and robust similarity estimation. Experimental results demonstrate that GBNet-STS outperforms traditional methods, achieving state-of-the-art performance in Vietnamese semantic similarity tasks.

Keywords: Sentence similarity · word embeddings · semantic similarity.

1 Introduction

The measurement of semantic similarity between two sentences, often referred to as sentence similarity, is a fundamental problem in natural language processing. This task has widespread applications in various NLP domains, including information retrieval [8], natural language understanding [17], machine translation [26], speech recognition [19], question-answering systems [4], and text summarization [3]. Research on sentence similarity dates back to Luhn's pioneering work in 1957 [16], and has since evolved through numerous approaches, ranging from lexical similarity methods to deep learning-based sentence embeddings.

Traditionally, structure-based methods have been used to compute sentence similarity by analyzing sentence components and word relations. Wang et al. [25] introduced a technique for segmenting and reorganizing lexical semantics to improve similarity measurement. Similarly, Lee et al. [12] proposed a method based on category and semantic network extraction, while Ferreira et al. [9] leveraged word order and grammatical structure for similarity assessment.

With the advent of deep learning, sentence embedding models have become a dominant approach. Pre-trained models such as BERT, SBERT, and RoBERTa encode sentences into high-dimensional vector representations, effectively capturing semantic meaning in a compact form. However, these embeddings introduce information loss due to dimensionality reduction, making it difficult to capture fine-grained token-level relationships between words [23]. As a result, embedding-based similarity methods often struggle with handling paraphrases, negations, and syntactic variations, leading to suboptimal performance in certain linguistic scenarios.

To address these issues, we propose GBNet-STS (Graph-Based Network for Semantic Textual Similarity), a novel framework for measuring sentence similarity. Our approach integrates multiple semantic layers within a graph-based model, where each layer represents a distinct similarity measurement technique. We leverage the Pre-trained PhoBERT model [20], which is recognized as a state-of-the-art language model for Vietnamese, to enhance the representation of sentence semantics. Additionally, we incorporate multiple similarity metrics, including Cosine Similarity, Longest Common Subsequence (LCS), and Jaccard Similarity, to capture diverse aspects of semantic relationships between sentences.

By utilizing a multi-layered semantic graph, our approach provides a more comprehensive and interpretable similarity measurement. The experimental results demonstrate that GBNet-STS achieves high performance on Vietnamese datasets, outperforming traditional methods that rely solely on either lexical-based measures or deep learning-based embeddings. Our findings emphasize the importance of integrating token-level, sentence-level, and distributional representations to achieve state-of-the-art performance in Vietnamese semantic similarity tasks.

The structure of this paper is as follows: Section II reviews relevant studies in the field. Section III introduces a novel sentence similarity measurement method utilizing a multilayer graph-based approach. Section IV describes the construction of a Vietnamese sentence similarity dataset. Section V presents the conducted experiments and analysis. Section VI concludes with key findings and suggests directions for future research.

2 RELATED WORK

Semantic similarity has long been a fundamental aspect of Natural Language Processing (NLP), serving as a cornerstone for numerous applications across related domains. Semantic Textual Similarity (STS) provides a framework for evaluating the similarity between sentences or documents by analyzing their explicit and implicit semantic relationships. These relationships are often derived from associations within the broader context of a corpus. By leveraging these semantic connections, STS enables a more nuanced understanding of the similarity between textual units. Over time, advancements in this area have led to

the development of various state-of-the-art models, offering robust solutions for semantic similarity measurement.

2.1 phoBERT

Recent advancements in natural language processing (NLP) have been greatly influenced by the emergence of pre-trained language models such as BERT [7] and its variants. While these models have demonstrated remarkable success, their application to languages other than English often requires language-specific training. In the context of Vietnamese language processing, existing models have been limited by the scarcity of large-scale, high-quality training datasets, with many relying solely on the Vietnamese Wikipedia corpus. Furthermore, the syllable-based nature of many models for Vietnamese does not align well with the word-level requirements of many downstream tasks. To address these shortcomings, Nguyen and Nguyen [21] introduced **PhoBERT**, the first publicly available large-scale monolingual language model specifically pre-trained for Vietnamese. PhoBERT, available in both base and large configurations, is built upon the BERT architecture but leverages the robust pre-training procedure introduced by RoBERTa [15]. It is trained on a 20GB word-segmented Vietnamese corpus comprised of a combination of Wikipedia and news data. The experimental results detailed by Nguyen and Nguyen [21] demonstrate that PhoBERT outperforms the current state-of-the-art multilingual model XLM-R [5] and attains new SOTA results on a wide range of Vietnamese specific NLP tasks including part-of-speech tagging, dependency parsing, named-entity recognition, and natural language inference. The public release of PhoBERT is intended to advance future research in Vietnamese NLP and has already become a strong baseline for further experimentation. In this work, our multi-layered approach leverages the advanced capabilities of PhoBERT for embedding the text elements before extracting the semantic relationship.

2.2 Conditional Semantic Textual Similarity (C-STS)

Recent research has introduced Conditional Semantic Textual Similarity (C-STS) as an extension of traditional STS, where the similarity between sentence pairs is assessed under specific conditions [6]. Unlike conventional STS, C-STS considers different contextual perspectives, which enables more fine-grained similarity evaluations. This approach has been particularly useful in applications such as fine-grained retrieval and large language model text attribution.

One of the pioneering methods for C-STS is QuMSE, which combines Quad loss and Mean Squared Error (MSE) loss for optimization. However, QuMSE has been found to suffer from an over-estimation problem, where positive sample similarity scores are excessively high while negative sample similarity scores are too low.

To address this issue, Conditional Contrastive Learning (CCL) has been proposed as an alternative approach [14]. CCL introduces two key innovations:

- Weighted Adaptive Contrastive Loss (W-ACL): This loss function adaptively selects the optimization direction for positive and negative samples based on their similarity scores, mitigating the over-estimation problem.
- Balanced Contrastive Loss (BCL): This loss function ensures a balance between hard negative samples and false negative samples, improving the robustness of the model.

Experiments on the C-STS dataset show that CCL significantly outperforms QuMSE and even achieves better results than state-of-the-art large language models such as GPT-4. The effectiveness of contrastive learning in C-STS suggests its potential applicability in broader STS tasks, including those involving graph-based models.

2.3 Semantic Similarity Graph

Several recent studies have explored multi-layered graph-based models for semantic textual similarity (STS). One of the most notable approaches is MNet-Sim [10], which proposes a multi-layered semantic similarity network integrating various similarity measures such as Cosine Similarity, Phrasal Overlap, Euclidean Distance, Jaccard Similarity, and Word Mover's Distance. MNet-Sim applies network science principles and an extended node similarity computation formula to assess sentence similarity.

While both GBNet-STS and MNet-Sim leverage multi-layered graph structures, the fundamental difference lies in how inter-layer relationships are modeled and weighted. MNet-Sim does not prioritize a single dominant layer; instead, it treats all similarity layers equally and aggregates them using a specialized similarity computation method. In contrast, GBNet-STS designates a primary semantic similarity layer (Cosine Similarity) and uses Jensen-Shannon Divergence (JSD) to adjust the contribution of other layers dynamically. This weighted aggregation strategy enables GBNet-STS to achieve more accurate similarity scores, particularly for Vietnamese texts.

Furthermore, the graph construction process differs between the two models. MNet-Sim explores multiple edge selection strategies, including Preceding Adjacent Vertex (PAV), Single Similarity Vertex (SSV), and Multiple Similarity Vertex (MSV), to establish connections between sentence pairs. However, GBNet-STS employs a threshold-based edge formation approach, ensuring that only semantically relevant sentence pairs are connected within the graph. These distinctions highlight the advantages of GBNet-STS in adapting to Vietnamese-specific text characteristics and effectively leveraging multi-layered similarity computations.

3 METHODOLOGY

Traditional semantic similarity graph-based methods evaluate sentence and text similarity by employing centering theory, which posits an inverse relationship between the number of attention shifts within a text and its coherence or semantic

similarity. However, these methods are inherently limited to local contexts, focusing primarily on neighboring sentences (preceding and following), thereby failing to capture semantic similarity between distant or unrelated sentences effectively. These approaches often restrict the application of multiple similarity metrics simultaneously, limiting their ability to evaluate semantically similar sentences within a network comprehensively.

To address these limitations, this research introduces a model designed to compute semantic similarity efficiently for both neighboring and detached sentence pairs while integrating multiple similarity measures. Instead of treating the problem in a flat structure, this study proposes a multi-layered semantic similarity network, which allows for a more comprehensive evaluation of sentence similarity. The paper also presents a novel node similarity computation formula to assess overall similarity within the constructed multi-layer network. We have also endeavored to develop a synonym dataset to directly map the semantics of different words. This effort has led to remarkable improvements in semantic measurement compared to lexical-based methods such as Jaccard or LCS. Further details on the proposed model are elaborated in the subsequent sections.

3.1 Layers Definition

In our multi-semantic layered graph framework, each layer is constructed upon a distinct method for quantifying semantic similarity between textual elements. This approach enables us to capture a multifaceted understanding of semantic relations, leveraging the strengths of individual similarity measures. Each layer reflects unique characteristics in terms of how it perceives semantic relatedness. The specific layers used in this work are detailed in the following subsections:

Cosine Similarity Layer The Cosine Similarity layer quantifies the semantic similarity of textual elements by calculating the cosine of the angle between their vector representations. This approach assumes that the semantic proximity of two text elements is directly proportional to the direction of their vectors in the vector space. The cosine similarity focuses on the orientation of the vectors rather than their magnitudes, making it suitable for comparing documents of varying lengths. The calculation is performed by taking the dot product of the two vectors and normalizing by the product of their Euclidean norms. Formally, given two vectors v_i and v_j representing text elements i and j respectively, the cosine similarity is computed as:

$$\text{Cosine Similarity}(v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \tag{1}$$

where $\cdot$ is the dot product and $\|v\|$ is the Euclidean norm of vector v. A higher cosine value indicates a stronger similarity between the text elements. To create this layer, a weighted edge is established between each pair of text elements, where the weight is the value of the computed cosine similarity, which will be used later to build the graph.

Longest Common Subsequence (LCS) Layer The LCS layer assesses textual similarity by identifying the longest sequence of words that appear in the same order within the compared text elements, irrespective of their contiguity. This measure captures the degree to which word sequences are shared between the documents, revealing common structural patterns. Unlike methods that consider only the presence of words, LCS considers the sequential matching of the text sequences, therefore emphasizing common ordering of words and highlighting shared sequential structures. To calculate the LCS, we employ dynamic programming to build a table indicating the length of the longest common subsequence at each position of the two text sequences. The length of the LCS is calculated via the recurrence:

$$\mathrm{LCS}(i,j) = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 \\ \mathrm{LCS}(i-1,j-1)+1 & \text{if } x_i = y_j \text{ or } x_i \approx y_j \\ \max(\mathrm{LCS}(i-1,j), \mathrm{LCS}(i,j-1)) & \text{if } x_i \neq y_j \end{cases}$$

where $x = x_1, x_2, \ldots, x_m$ and $y = y_1, y_2, \ldots, y_n$ are the two word sequences of the compared text elements. The math symbol "$\approx$" indicates that x_i and y_j are synonymous words, corresponding to mappings in the synonym dictionary we have developed. The introduction of this semantic mapping allows the LCS layer to capture deeper linguistic relationships between sentences, reducing the risk of underestimating similarity due to lexical variations. The length of the LCS between x and y can be calculated as $\mathrm{LCS}(m,n)$. This length is normalized by the length of the longer sequence, resulting in a similarity score between 0 and 1, that is then used as the weight of an edge between text elements.

Given two sentences, S_1 has m words and S_2 has n words, the Longest Common Subsequence (LCS) algorithm determines the longest common string of two sentences S_1 and S_2.

Jaccard Similarity Layer The Jaccard Similarity layer, also known as the Jaccard index, measures the similarity between two text elements based on the ratio of the size of the intersection of their word sets to the size of their union. The Jaccard index, therefore, indicates the level of word overlap, offering a straightforward measure of lexical similarity, without considering the semantics or context of words. Given two text elements with word sets A and B, their Jaccard Similarity is calculated as:

$$\text{Jaccard Similarity}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where $|A \cap B|$ denotes the number of words in common between A and B. $|A \cup B|$ denotes the total number of unique words in sets A and B. In this layer, an edge is created between two text elements and is weighted by the Jaccard similarity between them, expressing the amount of lexical overlap of the texts. However, traditional Jaccard Similarity does not account for semantic equivalence, meaning that two words with similar meanings but different lexical forms would be

Algorithm 1: Longest Common Subsequence Algorithm.

```
def LCS(S1, S2);
  Input: Two sentences S1, S2
  Output : Length of longest common subsequence of S1 and S2
n = len(S1); // the number of words in sentence S1
m = len(S1); // the number of words in sentence S2
int L[n + 1][m + 1] ; // weight matrix
for (int i=0; i≤n; i++) do
    for (int j=0; j≤m; j++) do
        if (i == 0)(j == 0) then
            L[i][j] = 0;
        else
            if S1[i − 1] == S2[j − 1] or S1[i − 1] ≈ S2[j − 1] then
                L[i][j] = L[i − 1][j − 1] + 1;
            else
                L[i][j] = Max(L[i − 1][j], L[i][j − 1]);
            end
        end
    end
end
return L[n][m];
```

treated as distinct, leading to lower similarity scores. To overcome this limitation, GBNet-STS also integrates the synonym dictionary for Jaccard Similarity Layer, redefining the intersection operation $|A \cap B|$ by including the synonymous words.

By improving the intersection function with the synonym dictionary, the Jaccard Similarity layer achieves higher reliability, particularly in cases where sentences express similar meanings using different lexical choices. This enhancement significantly improves the robustness of semantic similarity measurements, especially for Vietnamese, where synonym-rich expressions are common.

3.2 Semantic Similarity Multi-layers Graph

A graph is an intuitive and effective structure for representing relationships between entities (nodes). It enables visualization of the influence of neighboring nodes on the semantic relationship between the two target nodes being measured. In our approach, we employ a total of four semantic graph layers, corresponding to four methods of semantic similarity measurement: Cosine Similarity, Jaccard Similarity, LCS Similarity, and Word Mover's Distance (WMD). The general structure of the graph is illustrated in Figure 1. All semantic layers share the same number of nodes and identical indexing, while the edges differ based on the specific similarity measurement algorithm implemented in each layer.

Notably, we designate one layer as the primary semantic network layer—in this case, the Cosine Similarity layer. This choice assigns greater significance

to the similarity scores calculated within this layer. The remaining layers are considered auxiliary and are utilized to capture additional semantic nuances, complementing the final semantic similarity score.

A. Layers Generating Having established the diverse semantic similarity measures that will form the foundation of our graph layers (as detailed in Section 3.1, 3.1, **??** and 18), this subsection outlines the procedure for constructing each layer and defining the inter-layer relationships. Each layer, representing a different semantic perspective, has a unique topology dictated by the distribution of its similarity scores. The process involves two key steps: edge generation within each layer and assessing the weight of the relationship between the layers.

a) Edge Generation within a Layer To construct the graph structure within each layer, we introduce a thresholding approach based on the distribution of similarity scores. For each layer, we begin by computing the distribution of similarity values between all pairs of nodes. We then determine the most representative value of this distribution. In our implementation, we utilize the mode of the empirical distribution, which is calculated as the value that appears most frequently in the sample. This most representative value serves as a threshold θ. An edge is established between two nodes if their calculated similarity value exceeds this threshold. This method ensures that only strongly semantically related pairs of nodes are directly connected within the graph, reducing the noise and increasing the discriminative power of each layer.

The rationale behind using a threshold derived from the mode is that the mode captures a typical value in the observed similarity distribution, representing a common level of similarity within the layer. By focusing on pairs with higher than this typical similarity, we are only creating edges between the text elements with the stronger relationship on this layer, while removing other edges that will act as noise for our network. This selection process leads to a more focused representation of semantic relationships based on each specific measure.

b) Assessing Inter-Layer Relationship The construction of a multi-layered graph necessitates not only understanding the relationships within each layer, but also the relationships between the layers. To quantify the relationships between layers, we employ the Jensen-Shannon Divergence (JSD), a measure of the similarity between two probability distributions. Each layer's edge weights (i.e. the similarities computed by each method) when properly normalized, may be interpreted as a probability mass function over the edges in this layer. Therefore, the JSD provides a means to assess the degree to which the semantic perspectives captured by the layers are aligned.

Specifically, let P_i and P_j represent the distributions of similarity values from layer i and layer j after a normalization step such that they can be interpreted as a probability mass functions. The JSD between two layers i and j is defined as:

$$JSD(P_i, P_j) = \frac{1}{2} D_{KL}(P_i \| M) + \frac{1}{2} D_{KL}(P_j \| M) \tag{3}$$

where $M = \frac{1}{2}(P_i + P_j)$ is the average distribution, and D_{KL} represents the Kullback-Leibler (KL) divergence. A lower JSD indicates a greater similarity between the two layers, as it implies that their respective similarity distributions are more similar, while a higher JSD suggest that these two layers are capturing diverse semantic relationships. The JSD values are used as weights for representing the inter-layer connections. This means that the more similar the two layers are, less weight will be given to the connection between the two layers, and vice versa.

This method allows us to capture not only the individual perspectives of each semantic layer but also how they interact and relate to one another, providing a more holistic understanding of the semantic space in the context of the multi-semantic layered graph.

B. Determining local layer similarity Considering a layer l in the multi-layered network. This layer represents a specific similarity measure (e.g., cosine similarity, LCS, etc.). We want to compute the local similarity between two sentence vectors A and B within this layer. The edges between nodes in this layer are established if their corresponding similarity score exceeds a predefined threshold θ. The weight of an established edge is the similarity score between the corresponding sentences.

The following formula [10] to calculate the local layer similarity, denoted by $\mathrm{Sim}_l(A, B)$:

$$\mathrm{Sim}_l(A, B) = \frac{\mathrm{sim}(A, B) + \sum_{wtd}(A, B)}{1 + \sum w\_adj(A, B)} \tag{4}$$

where:

- $\mathrm{sim}(A, B)$ represents the weight of the edge directly connecting nodes A and B (i.e., the similarity score between A and B according to the specific measure of layer l).
- $\sum_{wtd}(A, B)$ is the weighted sum of the edges connecting the adjacent nodes of A and B.
- $\sum w\_adj(A, B)$ is the sum of the similarity scores used as weights for the edges connecting the adjacent nodes of A and B.

To compute $\sum w\_adj(A, B)$ and $\sum_{wtd}(A, B)$, we first identify the adjacent nodes of A and B. Let $\mathcal{N}(A)$ and $\mathcal{N}(B)$ denote the sets of adjacent nodes for A and B respectively. Then:

$$\sum w\_adj(A, B) = \sum_{x \in \mathcal{N}(A), y \in \mathcal{N}(B)} \mathrm{sim}(A, x) \times \mathrm{sim}(B, y) \tag{5}$$

$$\sum_{wtd}(A, B) = \sum_{x \in \mathcal{N}(A), y \in \mathcal{N}(B)} \mathrm{sim}(x, y) \times \mathrm{sim}(A, x) \times \mathrm{sim}(B, y) \tag{6}$$

where $x \neq y$ while x and y are connected by a calculated similarity edge. This process is repeated for each layer in the multi-layered network, using the corresponding similarity measure for that layer to compute the edge weights and local layer similarities.

C. Overall sentence pair similarity through layers To compute the overall similarity between two sentences, denoted as $\text{OverallSim}(A, B)$, we introduce a novel aggregation method leveraging the multi-layered network structure by employing two distinct weighting schemes.

First, we calculate a main layer weight, denoted ω_m, representing the average Jensen-Shannon divergence between the designated main layer m and all other layers $l \in L$, where L is the set of all layers in the network excluding the main layer. This main layer weight quantifies the average informational divergence between the chosen main layer and the remaining layers:

$$\omega_m = \frac{1}{|L|} \sum_{l \in L} \text{JSD}(P_m, P_l)$$

Second, we compute a sum weight, ω_s, which represents the total informational divergence across all layers in the network. This sum weight captures the overall inter-layer relationships and accounts for the diversity of the similarity measures encoded within each layer:

$$\omega_s = \sum_{l_1, l_2 \in L, l_1 \neq l_2} \text{JSD}(P_{l_1}, P_{l_2})$$

Finally, the overall sentence similarity is calculated using the following formula:

$$\text{OverallSim}(A, B) = (\omega_s - \omega_m)\, \text{Sim}_m(A, B) + \frac{\prod_{l \in L} \text{Sim}_l(A, B)}{\omega_s \sum_{l \in L} \text{Sim}_l(A, B)}$$

where $\text{Sim}_l(A, B)$ represents the calculated local similarity between sentences A and B in layer l following the Equation 4. This formula incorporates both the main layer's similarity and the product of similarities from other layers, normalized by a weighted sum of these similarities. The weighting factors ω_m and ω_s modulate the influence of the main layer and the overall inter-layer divergence, respectively. The choice of the main layer m can be driven by the specific application or dataset. A layer corresponding to a similarity metric deemed highly relevant can be designated as the main layer. This formulation allows for a balanced consideration of individual layer contributions and their overall relationships within the multi-layered framework.

4 Construction of a Vietnamese Semantic Textual Similarity Dataset

4.1 Overview of the ViSTS Dataset

To date, research on the construction of monolingual Semantic Textual Similarity (STS) datasets has been primarily conducted for the English language, with only a limited number of studies focusing on other languages. Since 2005, several STS datasets have been developed to support various natural language processing tasks. Lee et al. created a dataset consisting of 65 sentence pairs, with similarity scores ranging from 0 to 4, though its scale remained limited [11]. Li et al. expanded this work by introducing a dataset of 50 news documents, evaluating similarity on a scale of 1 to 5; however, their focus was more on document-level rather than sentence-level similarity [13]. The SemEval STS dataset was developed using various sources such as news articles, video descriptions, and forum discussions [1]. Subsequent versions from 2013 to 2016 further increased the dataset size and diversity of sentence genres, while also incorporating Spanish-language data. Marelli et al. introduced the SICK dataset, which comprises 10,000 sentence pairs annotated with similarity scores from 1 to 5, along with entailment relations [18]. Additionally, Khukrit Osathanunkul proposed an STS dataset for the Thai language [22], and Jakub Sido et al. developed a dataset for the Czech language [24].

Existing monolingual STS datasets still exhibit several limitations, which negatively impact the effectiveness of NLP model training. First, small scale and lack of diversity are major issues, as many datasets focus primarily on English and specific text genres, such as news articles or image captions, without encompassing conversational or domain-specific language. Additionally, the distribution of similarity scores is highly imbalanced, with the majority of sentence pairs receiving low scores (0 or 1), making it difficult for models to learn fine-grained distinctions between highly similar sentences.

Moreover, current datasets lack interpretability, as they provide only similarity scores without explanations. Furthermore, inter-annotator agreement remains low, with correlation rates ranging from 62% to 87% [2], leading to inconsistencies in human-labeled data. Most notably, Vietnamese still lacks a high-quality STS dataset to support the development of NLP systems, highlighting the need for a well-constructed semantic similarity corpus for this language.

In this study, we construct a Vietnamese Semantic Textual Similarity (ViSTS) dataset. The ViSTS dataset is developed through a rigorous and reliable process. First, we select a high-quality and trustworthy monolingual Vietnamese corpus. Next, we carefully choose sentence pairs based on similarity distribution, domain diversity, and sentence length variations to ensure dataset balance and representativeness. The similarity assessment of sentence pairs is conducted meticulously by responsible annotators. The similarity scores assigned by the annotators are validated using statistical measures such as correlation and inter-annotator agreement. The detailed steps of the ViSTS dataset construction are presented below.

4.2 Data Collection Process

The construction of the ViSTS dataset was carried out through a structured and systematic process. The dataset was derived from the Vietnamese Corpus (Vcorpus), a large-scale corpus comprising 28 million sentences and approximately 560 million words, sourced from various online newspapers. The following steps outline the data collection process:

- Initial Sentence Extraction - Sentences were extracted from the Vcorpus, ensuring linguistic diversity and contextual richness. - The selection focused on domains such as news, sports, economy, education, science, and tourism to provide a comprehensive representation of real-world text.
- Sentence Pairing - To ensure even coverage across different levels of semantic similarity, pairs of sentences were calculated using the Jaccard similarity coefficient ($Sim_{Jaccard}$) based on word overlap. - Sentence pairs were categorized into ten intervals within the range of [0, 1], with 100 pairs selected from each interval.
- Length Filtering - Sentence length constraints were applied to exclude excessively short or long sentences, which might lack meaningful context or pose challenges in annotation. - Only sentences with a word count between 10 and 50 were retained for pairing.
- Domain Balancing - The selected sentence pairs were curated to ensure a balanced distribution of domains, capturing a wide variety of language usage patterns. - This step ensures the dataset reflects diverse contexts, enhancing its applicability to various NLP tasks.

The resulting dataset consists of 600 sentence pairs annotated for semantic similarity, offering an evenly distributed and domain-diverse resource. This process ensures the ViSTS dataset is robust and reliable for training and evaluating Vietnamese NLP models.

4.3 Annotation Process

The annotation process for the ViSTS dataset was meticulously designed to ensure consistency and reliability across all annotated sentence pairs. This section details the steps and guidelines provided to annotators, who were tasked with evaluating semantic similarity on a predefined scale.

- Recruitment of Annotators - The annotation team consisted of linguistic experts and domain specialists proficient in Vietnamese. Their expertise ensured accurate and consistent evaluations.
- Annotation Scale - Annotators were instructed to score the semantic similarity of sentence pairs on a five-point scale ranging from 0 to 4. This scale captures varying degrees of similarity, from completely unrelated to very similar.
- Guidelines for Annotation - Detailed guidelines were provided to annotators, outlining the criteria for each similarity score. These guidelines aimed to standardize the evaluation process and minimize subjective interpretation. The guidelines are summarized in Table 1.

Table 1. Annotation guidelines provided to annotators.

Title	Scale	Description
Very Similar	4	Two sentences are completely similar in meaning. They refer to the same object or concept, using words with semantic similarity or synonyms to describe them. The lengths of the two sentences are equivalent.
Somewhat Similar	3	Two sentences with slight similarities in meaning, referring to the same object or concept. The lengths of the two sentences may vary slightly.
Somewhat Related but Not Similar	2	Two sentences that are related in meaning, each referring to objects or concepts that are related but not identical. The lengths of the two sentences may vary slightly.
Slightly Related	1	Two sentences that differ in meaning but have slight semantic relatedness, possibly sharing the same topic. The lengths of the two sentences can vary greatly.
Unrelated	0	The two sentences are completely different in meaning, and their content is not related. The lengths of the two sentences can vary greatly.

- Quality Control - To ensure annotation quality, a subset of sentence pairs was annotated by multiple annotators. The inter-annotator agreement was measured to assess the consistency of the annotations.
- Final Annotation - After completing the annotation process, the scores for each sentence pair were finalized based on the consensus among annotators. Discrepancies were resolved through discussion or by involving additional annotators.

This structured annotation process ensures the dataset's reliability and utility for evaluating and training NLP models on Vietnamese semantic textual similarity tasks.

The selected sentence pairs are randomly divided into ten sets, each containing 100 sentence pairs. Each set is evaluated by 12 information technology students (Annotators) for similarity according to five levels as shown in Table 1. Before annotators assessing the similarity of sentence pairs, they were instructed on similarity levels including totally different, different, slightly similar, similar, very similar, as well as how to estimate the similarity of sentence pairs. Annotators conduct the evaluation of the similarity of sentence pairs independently.

4.4 Statistical Analysis

To evaluate the quality and characteristics of the ViSTS dataset, we performed a detailed statistical analysis on the annotated sentence pairs. The dataset consists of 600 sentence pairs, each assigned an average similarity score based on assessments by multiple annotators.

The similarity scores range from 0 (completely unrelated) to 4 (very similar). The distribution of scores across this range is well-balanced: 15% of pairs scored

0, 18.3% scored 1, 19.3% scored 2, 27.2% scored 3, and 20.2% scored 4. This even distribution ensures comprehensive coverage of different similarity levels, making the dataset versatile for evaluating semantic similarity models.

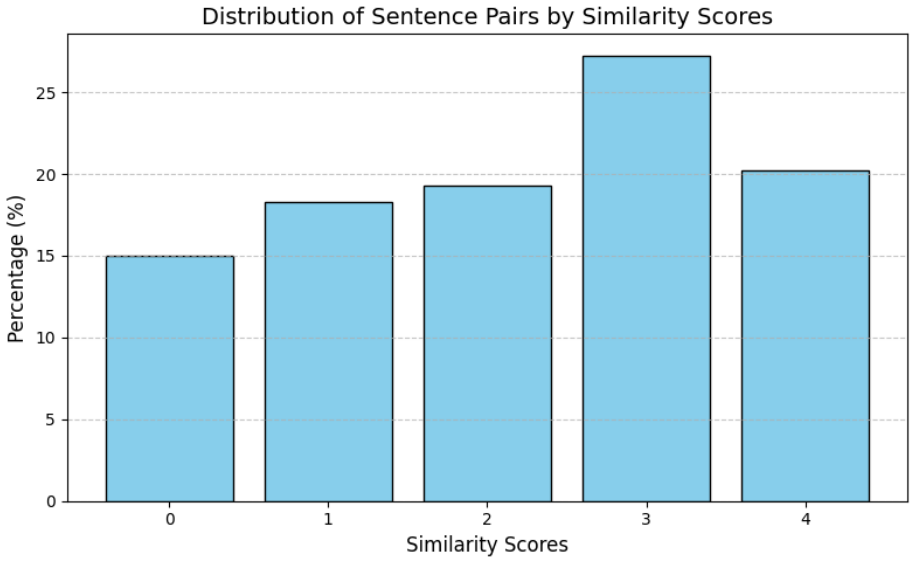

Fig. 1. Distribution of Sentence Pair Similarity Scores in the ViSTS Dataset.

The length of the sentences in the dataset varies from 6 to 94 words. This broad range ensures that sentences provide sufficient context for meaningful similarity evaluation. The inclusion of moderately long sentences contributes to reducing bias while maintaining diversity in sentence structure and complexity.

The dataset encompasses sentence pairs from a wide range of domains, including news, sports, literature, science, technology, tourism, and daily life. Specifically, domains like "Tin t´c" (News) account for 197 pairs, while "Th Thao" (Sports) and "Đ`i S´ng" (Lifestyle) represent 75 and 60 pairs, respectively. This domain diversity enhances the dataset's applicability for evaluating semantic textual similarity across various real-world contexts.

The similarity scores for each pair were computed as the average of evaluations from multiple annotators. The consistency of these annotations is reflected in a standard deviation of 1.35 across all scores. This indicates a high degree of agreement among annotators, ensuring reliable ground-truth labels.

Higher similarity scores (3–4) are often associated with pairs describing the same topic using semantically similar or synonymous expressions. For example, a pair of sentences comparing geometric objects or restating factual information tends to score highly. Conversely, lower scores (0–1) are assigned to pairs with

distinct topics or negligible semantic overlap, such as unrelated descriptions of physical processes or speculative scenarios.

These statistics demonstrate that the ViSTS dataset is well-constructed, with balanced coverage across similarity levels, diverse sentence lengths, and robust annotations. This makes it a reliable resource for training and evaluating semantic similarity models tailored to Vietnamese.

The ViSTS dataset represents an essential step towards enabling Vietnamese NLP models to perform more effectively in tasks involving semantic textual similarity.

5 Experiment

5.1 Experimental Setup

This section details the experimental setup used to evaluate the effectiveness of our proposed multi-layered graph approach for sentence similarity.

A. Dataset: We utilize the ViSTS dataset, as described in Section 4, for our experiments. This dataset provides a diverse range of sentence pairs and corresponding ground truth similarity scores, enabling a robust evaluation of our method.

B. Embedding Extraction: We leverage three distinct types of Large Language Models (LLMs) to extract embeddings for each sentence in the ViSTS dataset:

- **Vietnamese-fine-tuned LLMs:** These model PhoBERT v2 has been specifically fine-tuned for the Vietnamese language and are expected to capture nuances specific to Vietnamese text.
- **Multilingual LLMs (Vietnamese-capable):** This category includes models like Qwen family, which are multilingual and demonstrate proficiency in Vietnamese despite not being explicitly fine-tuned for it.
- **Multilingual LLMs (Not Vietnamese-focused):** We also include models like Llama family, which are multilingual but not specifically designed or trained for Vietnamese. This allows us to investigate the robustness of our approach when applied to LLMs with varying degrees of Vietnamese language proficiency.

These embeddings serve as the initial representation of the sentences and are used as input for both the baseline comparison and our proposed multi-layered graph method.

C. Multi-layered Graph Construction: The extracted sentence embeddings are then used to construct the multi-layered graph. Each layer of the graph represents a different similarity measure, nodes in the graph represent sentences, and edges are weighted based on the corresponding similarity scores between the connected sentences within each layer. The overall similarity score between a sentence pair is then computed using the method described in Section 3.

D. Evaluation Metrics: We compare the performance of our multi-layered

graph approach against a baseline of using the raw LLM embeddings directly. Specifically, we calculate the Pearson correlation, Spearman correlation between:

- The cosine similarity of the original LLM embeddings for each sentence pair and their corresponding ground truth similarity scores.
- The overall similarity scores derived from our multi-layered graph and their corresponding ground truth similarity scores.

This comparison allows us to assess the effectiveness of our approach in capturing sentence similarity relative to the inherent similarity captured by the LLM embeddings. Higher correlation values indicate stronger agreement with the ground truth and, thus, better performance.

5.2 Evaluation

We hypothesize that integrating multiple semantic similarity measures through a layered graph framework will enhance performance compared to using raw LLM embeddings directly. The evaluation is based on Pearson and Spearman correlation coefficients between predicted similarity scores and the ground truth annotations from the ViSTS dataset. We compare our approach against a baseline using cosine similarity of raw LLM embeddings, including Vietnamese fine-tuned model like PhoBERT v2, multilingual Vietnamese-capable (QWEN 2.5), and multilingual not Vietnamese-focused (Llama 3.1).

The experimental results clearly indicate that our GBNet-STS model outperforms existing sentence similarity methods across all tested models. The Pearson and Spearman correlation scores demonstrate a significant improvement over traditional embedding-based approaches. Notably, GBNet-STS shows strong consistency in capturing semantic similarities, even for sentence pairs that exhibit complex linguistic variations.

The Table 2 and 3 show that GBNet-STS provides consistent improvements in semantic similarity measurement across different models, with varying degrees of enhancement depending on the model's Vietnamese language capability. For Vietnamese-trained models like PhoBERT v2, GBNet-STS slightly enhances performance, demonstrating that graph-based similarity can refine even strong Vietnamese embeddings. The Pearson correlation improves from 0.668 to 0.672, and the Spearman correlation increases from 0.676 to 0.680, reinforcing that the multi-layered graph approach refines sentence similarity beyond direct embedding-based methods. For multilingual models with Vietnamese capability, such as Qwen 2.5 and Vistral 7B, improvements are observed but depend on how well the model already represents Vietnamese semantics. Specifically, GBNet-STS boosts Qwen 2.5's Pearson correlation from 0.642 to 0.655 and Spearman from 0.658 to 0.661, while for Vistral 7B, Spearman remains unchanged at 0.619, with Pearson showing a slight drop from 0.619 to 0.602. This suggests that while GBNet-STS is beneficial, its impact varies based on the quality of the model's initial embeddings. For general multilingual models like LLaMA 3.1 8B, the Spearman correlation improves from 0.614 to 0.626, indicating that GBNet-STS

enhances ranking consistency even when absolute similarity scores (Pearson) do not show significant improvements. Overall, these findings confirm that GBNet-STS consistently enhances ranking-based similarity measurements, making it a valuable approach for models that struggle with Vietnamese semantic similarity tasks.

Table 2. The proposed model GBNet-STS is compared with other LLMs in measuring semantic similarity performance through Pearson Correlation.

Model Pair Comparing	**Pearson Correlation**
GBNet-STS / phoBERT v2	**0.672** / 0.668
GBNet-STS / Vistral 7B	0.602 / **0.619**
GBNet-STS / Llama 3.1 8B	0.589 / **0.621**
GBNet-STS / Qwen 2.5 7B	**0.655** / 0.642

Table 3. The proposed model GBNet-STS is compared with other LLMs in measuring semantic similarity performance through Spearman Correlation.

Model Pair Comparing	**Spearman Correlation**
GBNet-STS / phoBERT v2	**0.680** / 0.676
GBNet-STS / Vistral 7B	**0.627** / 0.619
GBNet-STS / Llama 3.1 8B	**0.626** / 0.614
GBNet-STS / Qwen 2.5 7B	**0.661** / 0.658

Furthermore, the integration of multiple similarity layers within our framework ensures that different aspects of sentence similarity are accounted for. While cosine similarity provides a general measure of vector-based closeness, the inclusion of Jaccard Similarity and LCS allows for a more fine-grained analysis of word overlap and syntactic structure. This multi-faceted approach results in a more holistic and accurate similarity assessment.

5.3 Impact of Multi-layered Approach

The influence of different similarity layers is further analyzed in Table 4, which compares the performance of various layer combinations. The best-performing configuration is Cosine Similarity + Jaccard Similarity + LCS, achieving a

Spearman correlation of 0.680 with PhoBERT v2, which is higher than Cosine + LCS (0.671) and Cosine + Jaccard (0.675). Similar trends are observed across other models, where Qwen 2.5 improves from 0.640 to 0.661 when incorporating all three similarity layers.

These results confirm that Cosine Similarity alone provides a solid foundation, but adding Jaccard Similarity and LCS allows for a more nuanced evaluation of sentence similarity. Specifically, Jaccard Similarity captures lexical overlap, while LCS identifies shared syntactic structures, helping the model distinguish semantically close sentence pairs more effectively. This combination is particularly beneficial for handling paraphrased or reordered sentences, which traditional embedding-based methods may struggle with. Additionally, the results suggest that the impact of different similarity layers varies across models, indicating that optimal layer selection should be adapted based on the characteristics of the underlying embeddings. A typical piece of evidence for this is the result when applying only the Cos + Jaccard configuration to the LLaMA 3.1 8B model, which achieves a Spearman correlation of 0.638, even higher than the three-layer configuration, which scores 0.626.

Table 4. The performance of GBNet-STS with different combinations of layers through Spearman Correlation.

Model Pair Comparing	Cos + LCS	Cos + Jaccard	Cos + Jaccard + LCS
GBNet-STS / phoBERT v2	0.671 / **0.676**	0.675 / **0.676**	**0.680** / 0.676
GBNet-STS / Vistral 7B	0.598 / **0.619**	0.618 / **0.619**	**0.627** / 0.619
GBNet-STS / Llama 3.1 8B	**0.625** / 0.614	**0.638** / 0.614	**0.626** / 0.614
GBNet-STS / Qwen 2.5 7B	0.640 / **0.658**	0.647 / **0.658**	**0.661** / 0.658

The findings suggest that Cosine Similarity serves as a strong foundational measure, effectively capturing semantic meaning from embeddings. However, incorporating Jaccard Similarity and LCS allows for a more fine-grained comparison, addressing lexical and syntactic similarities that pure embeddings might overlook. This confirms the hypothesis that a multi-layered approach leads to a more comprehensive semantic similarity measurement.

5.4 Discussion

Our experimental results confirm that GBNet-STS outperforms both embedding-based baselines and existing graph-based approaches such as Mnet-SIM. This improvement is particularly noticeable in cases where sentence pairs exhibit complex paraphrasing, synonym usage, or varying syntactic structures.

One key observation is that our model mitigates the limitations of raw embedding similarity. While traditional cosine similarity performs well for direct

semantic alignment, it struggles with nuanced linguistic variations. By integrating Jaccard Similarity and LCS, GBNet-STS effectively addresses these gaps, achieving higher agreement with human annotations.

Furthermore, the results reinforce the adaptability of our approach. The model enhances performance across different types of embeddings, making it a viable solution for diverse NLP applications. In particular, its ability to improve multilingual models highlights its potential for low-resource languages, where pre-trained embeddings may lack domain-specific optimization.

5.5 Limitations and Future Work

Despite its strong performance, GBNet-STS has certain limitations. First, the computational overhead of constructing and processing the graph structure is higher than that of direct embedding-based approaches. While this complexity is justified by the performance gains, future research could explore optimization techniques to enhance efficiency.

Another limitation is the static nature of the similarity layers. While our results show that combining Cosine Similarity, Jaccard Similarity, and LCS is effective, the optimal layer configuration may vary for different datasets. Future work could investigate dynamic weighting mechanisms to adaptively determine the most relevant similarity metrics for each dataset.

Additionally, our experiments focus primarily on Vietnamese, but the approach could be extended to other low-resource languages. Evaluating GBNet-STS on different linguistic datasets would provide insights into its generalizability and potential for cross-lingual applications.

By addressing these challenges, we aim to further refine GBNet-STS, making it a more efficient and versatile solution for semantic similarity measurement in NLP. The promising results presented in this study pave the way for future innovations in graph-based similarity computation, particularly for languages that lack extensive annotated resources.

Acknowledgments

References

1. Agirre, E., Cer, D.M., Diab, M.T., Gonzalez-Agirre, A.: Semeval-2012 task 6: A pilot on semantic textual similarity. In: International Workshop on Semantic Evaluation (2012), https://api.semanticscholar.org/CorpusID:12549805
2. Agirre, E., Cer, D.M., Diab, M.T., Gonzalez-Agirre, A., Guo, W.: *sem 2013 shared task: Semantic textual similarity. In: International Workshop on Semantic Evaluation (2013), https://api.semanticscholar.org/CorpusID:10241043
3. Aliguliyev, R.M.: A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Syst. Appl. **36**(4), 7764–7772 (2009), http://dblp.uni-trier.de/db/journals/eswa/eswa36.htmlAliguliyev09
4. Burke, R., Hammond, K., Kulyukin, V., Tomuro, S.: Question Answering from Frequently Asked Question Files. AI Magazine **18(2)**, 57–66 (1997)

5. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of ACL. pp. 8440–8451 (2020)
6. Deshpande, A., Kamoi, S., Asai, A., Lin, J.: Qumse: Quadruplet-based metric learning for conditional semantic textual similarity. In: Proceedings of the 31st International Conference on Computational Linguistics (COLING) (2023)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL. pp. 4171–4186 (2019)
8. Farouk, M., Ishizuka, M., Bollegala, D.: Graph matching based semantic search engine. In: Garoufallou, E., Sartori, F., Siatri, R., Zervas, M. (eds.) MTSR. Communications in Computer and Information Science, vol. 846, pp. 89–100. Springer (2018), http://dblp.uni-trier.de/db/conf/mtsr/mtsr2018.htmlFaroukIB18
9. Ferreira, R., Lins, R.D., Simske, S.J., Freitas, F., Riss, M.: Assessing sentence similarity through lexical, syntactic and semantic analysis. Comput. Speech Lang. **39**, 1–28 (2016), http://dblp.uni-trier.de/db/journals/csl/csl39.htmlFerreiraLSFR16
10. Jeyaraj, M.N., Kasthurirathna, D.: Mnet-sim: A multi-layered semantic similarity network to evaluate sentence similarity. CoRR **abs/2111.05412** (2021), https://arxiv.org/abs/2111.05412
11. Lee, M.D., Pincombe, B.M., Welsh, M.: An empirical evaluation of models of text document similarity (2005), https://api.semanticscholar.org/CorpusID:645710
12. Lee, M.C., Zhang, J.W., Lee, W.X., Ye, H.Y.: Sentence similarity computation based on pos and semantic nets. In: Kim, J., Delen, D., Park, J., Ko, F., Rui, C., Lee, J.H., Wang, J., Kou, G. (eds.) NCM. pp. 907–912. IEEE Computer Society (2009), http://dblp.uni-trier.de/db/conf/ncm/ncm2009.htmlLeeZLY09
13. Li, Y., Mclean, D., Bandar, Z., O'Shea, J.D., Crockett, K.A.: Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering **18**, 1138–1150 (2006), https://api.semanticscholar.org/CorpusID:12007882
14. Liu, X., Qin, Z., Wang, Z., Liang, W., Zong, L., Xu, B.: Conditional semantic textual similarity via conditional contrastive learning. In: Proceedings of the 31st International Conference on Computational Linguistics (COLING) (2025)
15. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint, arXiv:1907.11692 (2019)
16. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development **1**, 309–317 (1957)
17. Manning, C.D., MacCartney, B.: Natural language inference. In: Natural language inference (2009)
18. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R.: A sick cure for the evaluation of compositional distributional semantic models. In: International Conference on Language Resources and Evaluation (2014), https://api.semanticscholar.org/CorpusID:762228
19. Morris, A.C., Maier, V., Green, P.D.: From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In: INTERSPEECH. ISCA (2004), http://dblp.uni-trier.de/db/conf/interspeech/interspeech2004.htmlMorrisMG04
20. Nguyen, D.Q., Nguyen, A.T.: Phobert: Pre-trained language models for vietnamese. In: Cohn, T., He, Y., Liu, Y. (eds.) EMNLP (Findings). pp. 1037–1042. Association for Computational Linguistics (2020), http://dblp.uni-trier.de/db/conf/emnlp/emnlp2020f.htmlNguyenN20

21. Nguyen, D.Q., Nguyen, A.T.: Phobert: Pre-trained language models for vietnamese. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1037–1042 (2020)
22. Osathanunkul, K.: Semantic similarity framework for thai conversational agents (2014), https://api.semanticscholar.org/CorpusID:5976464
23. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1410, https://aclanthology.org/D19-1410/
24. Sido, J., Seják, M., Praák, O., Konopík, M., Moravec, V.: Czech news dataset for semantic textual similarity. ArXiv **abs/2108.08708** (2021), https://api.semanticscholar.org/CorpusID:237213540
25. Wang, Z., Mi, H., Ittycheriah, A.: Sentence similarity learning by lexical decomposition and composition. In: Calzolari, N., Matsumoto, Y., Prasad, R. (eds.) COLING. pp. 1340–1349. ACL (2016), http://dblp.uni-trier.de/db/conf/coling/coling2016.htmlWangMI16
26. Yang, M., Wang, R., Chen, K., Utiyama, M., Sumita, E., Zhang, M., Zhao, T.: Sentence-level agreement for neural machine translation. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) ACL (1). pp. 3076–3082. Association for Computational Linguistics (2019), http://dblp.uni-trier.de/db/conf/acl/acl2019-1.htmlYangWCUSZZ19