IAES International Journal of Robotics and Automation (IJRA)
Vol. 15, No. 4, December 2026, pp. 1051~1061
ISSN: 2722-2586, DOI: 10.11591/ijra.v15i4.pp1051-1061

A hybrid approach for text similarity combining association rules and deep representation learning

Meriem Amina Zingla[1,2], Chérifa Boudia[2,3]
[1]Laboratory of Informatics and Intelligent Systems (LISYS), Mustapha Stambouli University of Mascara, Mascara, Algeria
[2]Department Computer Science, Faculty of Exact sciences, Mustapha Stambouli University of Mascara, Mascara, Algeria
[3]Technological Laboratory in Artificial Intelligence and Food Safety (LABTEC-IA), Mustapha Stambouli University of Mascara, Mascara, Algeria

Article Info

Article history:

Received Aug 14, 2026
Revised Oct 14, 2026
Accepted Nov 10, 2026

Keywords:

Deep learning
Hybrid similarity measure
INEX
Text mining
Text similarity

ABSTRACT

Text similarity plays an important role in several natural language processing applications, such as information retrieval, document clustering, and text summarization. However, traditional approaches often fail to effectively capture both syntactic patterns and semantic relationships between texts. In this paper, we propose a hybrid approach that combines text mining and deep learning techniques to improve text similarity measurement. The text mining component, referred to as the statistical text measure(STR), is based on association rule mining to extract relevant statistical patterns and relationships between terms, while the deep learning component, referred to as the semantic similarity measure (SSM), is used to capture semantic relationships between texts. The proposed approach integrates these two complementary components to compute a more accurate similarity measure. To evaluate its effectiveness, experiments are conducted on the tweet contextualization task, which aims to automatically generate informative summaries from a set of related documents. The evaluation is performed using the standard INEX 2013 and INEX 2014 datasets. Experimental results show that our proposed approach improves similarity performance compared with its individual statistical (STR) and semantic (SSM) components. While the proposed approach does not surpass the best official INEX results, it provides a competitive hybrid alternative that combines statistical evidence with semantic information in a lightweight and interpretable framework.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Meriem Amina Zingla
Department Computer Science, Faculty of Exact sciences, Mustapha Stambouli University of Mascara
Mascara, Algeria
Email: meriem.zingla@univ-mascara.dz

1. INTRODUCTION

Text similarity is a fundamental task in natural language processing (NLP) that aims to estimate the degree of similarity between texts using lexical, syntactic, and semantic information. Despite significant progress, accurate similarity measurement remains challenging, particularly for short texts with limited contextual information [1], [2]. This challenge is also relevant to language-driven robotics and automation, where text similarity can support instruction retrieval, human–robot dialogue, and access to contextual knowledge.

Traditional approaches mainly rely on statistical techniques such as term frequency, n-grams, and vector space models such as term frequency–inverse document frequency (TF-IDF) [3]. Although computationally efficient and relatively interpretable, these methods have limited ability to capture semantic

relationships. Semantic approaches based on distributed representations and deep learning can address this limitation by learning representations from large corpora [4]. However, they may require greater computational resources and generally provide less explicit evidence than statistical approaches. Thus, statistical methods may miss semantic relationships, whereas semantic methods may overlook explicit lexical or corpus-level associations. These limitations are particularly relevant to short-text contextualization, where limited context and differences in wording can make it difficult to identify relevant information. This has motivated hybrid methods that combine statistical and semantic information to benefit from their complementary characteristics [5].

Despite these developments, existing hybrid approaches often rely on complex neural architectures or task-specific training, while transformer-based models provide powerful contextual representations at higher computational cost. This highlights the need for lightweight and interpretable approaches that combine the complementary strengths of statistical and semantic methods.

In this work, we propose SHY, a hybrid similarity measure that combines an association-rule-based statistical component (STR) [6] with a semantic component based on Wikipedia2Vec [4]. STR captures explicit relationships between terms derived from corpus-level co-occurrence patterns, while Wikipedia2Vec captures their semantic relatedness. SHY combines these complementary sources through a simple weighted formulation, preserving the interpretability of statistical associations while incorporating semantic information beyond lexical overlap, without requiring task-specific training or fine-tuning of a large neural model. The main research question is whether this combination can improve text similarity for short-text contextualization.

The proposed approach is evaluated on the tweet contextualization task [7] using the INEX 2013 and INEX 2014 collections. Its performance is compared with the individual STR and SSM components and with the official INEX results using standard evaluation measures. The experimental results show that SHY achieves competitive performance and improves upon its individual components, providing evidence that statistical and semantic information can be effectively combined for short-text similarity.

Although the evaluation focuses on an NLP task rather than a physical robotic platform, the proposed similarity mechanism can support language-based robotics and automation applications, such as human–robot interaction, retrieval of task-related instructions, and access to textual knowledge by autonomous agents. Its lightweight design and independence from task-specific fine-tuning may also be beneficial for resource-constrained or embedded systems. However, further evaluation in robotics-oriented and embedded scenarios is required to assess its practical applicability and generalizability.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed approach. Section 4 describes the experimental settings and discusses the results. Finally, section 5 concludes the paper and presents future research directions.

2. RELATED WORK

The problem of measuring similarity between texts has attracted significant attention in natural language processing (NLP) due to its importance in information retrieval, text mining, and document understanding. Early approaches mainly relied on statistical and lexical representations, including vector space models, term frequency–inverse document frequency (TF-IDF), and n-gram matching. Although these methods are effective for measuring surface-level similarity and are generally computationally efficient, they are limited in their ability to capture deeper semantic relationships between texts, particularly when semantically related texts have limited lexical overlap [8].

Numerous studies have focused on statistical and extractive techniques for detecting and organizing relevant information from textual data. In [9], the authors addressed the problem of identifying informative and semantically relevant sentences within a document collection by proposing a graph-based clustering approach. Their method groups related phrases into a unified structure and incorporates a reordering mechanism to enhance the coherence of the resulting information. Similarly, authors in [10] investigated the retrieval and organization of relevant textual content for content enrichment. In this approach, a retrieval system is first employed to identify candidate documents, followed by a summarization stage that generates structured outputs using semantic representations. These approaches demonstrate the importance of combining information retrieval and content selection mechanisms for enriching textual information. In [11], the authors addressed the problem of extractive summarization, in which sentence importance is estimated using a greedy optimization strategy that assigns weights based on relevance scores. The proposed method generates summaries by selecting the most informative sentences from the source text.

In [12], automatic text enrichment using external knowledge sources is formulated as a summarization task, where Wikipedia articles are used to provide contextual information. However, the approach relies on the manual selection of relevant articles, which limits scalability and the degree of

automation. In [13], several statistical summarization models are proposed, including graph-based and vector-based methods, for ranking and selecting relevant sentences from text corpora. These approaches demonstrate that statistical representations remain useful for identifying relevant textual information, particularly in retrieval and extractive tasks. More recently, research has focused on combining retrieval and summarization strategies to improve text understanding and content enrichment. Hosking *et al.* [14] proposed a hierarchical retrieval approach for opinion summarization, where semantically relevant sentences are retrieved and used by a pre-trained language model to generate grounded summaries. Although this approach improves the informativeness and coherence of generated text, it focuses on retrieval and summarization rather than directly measuring similarity between short texts and candidate contextual information, highlighting the need for dedicated similarity mechanisms.

In parallel, deep learning approaches have significantly advanced semantic text representation. Aspect-aware models such as AspectCSE [15] learn context-sensitive sentence representations by incorporating structured semantic information, while hybrid models combine conceptual knowledge with distributed embeddings to improve semantic similarity estimation [16]. Deep metric learning architectures, such as 3D Siamese networks [17], further enable fine-grained pairwise similarity modeling. Pre-trained models such as TwHIN-BERT [18] incorporate social and contextual information to enhance multilingual representations. More recently, transformer-based models, including RoBERTa-based approaches [19], have demonstrated strong performance in semantic similarity tasks. Explainable transformers [20] and hybrid deep learning–optimization approaches [21] have also been proposed to improve interpretability, robustness, and accuracy. Recent surveys highlight contrastive learning and hybrid embedding strategies as promising directions for semantic textual similarity research [22].

Despite these advances, there remains a need for lightweight and interpretable hybrid approaches that balance the transparency of statistical methods with the semantic richness of embedding-based methods. Statistical approaches may miss semantic relationships, whereas neural and Transformer-based methods generally require greater computational resources and provide less explicit evidence. This gap is particularly relevant to short-text contextualization, where limited context and lexical variation make it difficult to identify relevant information.

To address this gap, we propose SHY, a hybrid similarity measure that combines an association-rule-based statistical component (STR) with a Wikipedia2Vec-based semantic component (SSM). STR captures explicit term dependencies from corpus-level co-occurrence patterns, whereas SSM captures semantic relatedness beyond lexical overlap. Their weighted combination provides complementary statistical and semantic evidence without requiring task-specific training or fine-tuning of a large neural model, making SHY a lightweight and interpretable approach for short-text similarity.

Table 1 summarizes the main characteristics, advantages, and limitations of the major approaches to text similarity, and positions the proposed SHY framework within this broader context.

Table 1. General comparison of text similarity approaches

Approach	Advantages	Limitations
Statistical/lexical	Simple, computationally efficient, and relatively interpretable	Limited ability to capture deeper semantic relationships; dependent on lexical overlap
Semantic/embedding-based	Captures richer semantic relationships beyond surface-level similarity	May require substantial computational resources and provides less explicit evidence
Transformer-based	Provides rich contextual representations and captures complex semantic relationships	Higher computational and memory requirements; generally less transparent
Hybrid approaches	Combine complementary lexical, statistical, and semantic information	Often involve complex architectures, multiple representations, or task-specific training
SHY (proposed)	Combines explicit statistical relationships with semantic representations in a simple and interpretable framework	Semantic representation is based on pre-trained Wikipedia2Vec and does not directly exploit contextual Transformer representations

3. THE PROPOSED HYBRID TEXT SIMILARITY APPROACH

In this section, we present the proposed approach for text similarity, which combines two complementary measures: a statistical measure and a semantic measure.

3.1. Statistical text similarity measure

This measure (denoted STR) is based on text mining techniques and relies on association rules [6] to estimate similarity between textual units from a statistical perspective. Unlike traditional similarity measures based solely on lexical overlap, STR exploits statistical dependency patterns between terms extracted from a document collection. These dependencies are modeled through association rules, which represent co-occurrence relationships discovered through global corpus analysis.

An association rule is generally expressed in the form:X→Y where X (the antecedent) and Y (the consequent) are disjoint sets of terms (X∩Y=∅). Such a rule indicates that when the set of terms X appears in a document, the terms in Y tend to appear as well. These rules therefore capture associations between terms that may not be evident through direct term matching.

The extraction of association rules is performed in two steps. First, frequent termsets are mined from the document collection, where each termset corresponds to a subset of the vocabulary whose elements frequently occur together. This extraction is guided by a minimum support threshold in order to retain only statistically significant patterns.

The support of a termset (X) measures how frequently the termset occurs in the corpus and is defined as (1) [6].

$$support(X) = \frac{\text{Number of documents containing X}}{\text{Total number of documents}} \quad (1)$$

In the second step, association rules are generated from these frequent termsets. For a rule R: X→Y its confidence is defined as[6]:

$$conf(\text{X} \rightarrow \text{Y}) = \frac{\text{support(X}\cup\text{Y)}}{\text{support(X)}} \quad (2)$$

Confidence estimates the conditional probability of observing Y given the presence of X, and can therefore be interpreted as a measure of statistical dependency between termsets.

The main advantage of association rules is their ability to capture contextual co-occurrence patterns beyond simple term matching, revealing statistical relationships between terms even when they do not occur within the same textual unit. To compute the statistical similarity between two texts t_1 and t_2, let T_1 and T_2 denote the sets of terms occurring in t_1 and t_2, respectively. Association-rule-based STR exploits the confidence of association rules to quantify the statistical dependency between terms occurring across the two texts. For each pair of terms $X \in T_1$ and $Y \in T_2$, the maximum confidence of an association rule connecting X to Y is defined as (3),

$$C(X,Y) = max\{conf(R) \mid R \in \mathcal{R},\ X \in Antecedent(R),\ Y \in Consequent(R)\} \quad (3)$$

where $\mathcal{R}$ denotes the set of extracted association rules.

Thus, all rules whose antecedent contains X and whose consequent contains Y are considered. For example, rules of the form X → AY and X → BY both contribute to (C(X,Y)), and the maximum confidence among them is selected. This maximum confidence provides the strongest directional evidence of the relationship between X and Y, rather than combining weaker associations. This is consistent with the objective of STR, which is to capture the strongest learned term relationship from the corpus. If no rule satisfying these conditions exists in $\mathcal{R}$, (C(X,Y)) is set to zero.

The STR similarity between t_1 and t_2 is then computed by averaging the maximum confidence values over all possible term pairs between the two texts:

$$\text{STR}(t_1,t_2) = 1/(|T_1||T_2|) \times \sum_{X\in T_1, Y\in T_2} C(X,Y) \quad (4)$$

The normalization factor $|T_1||T_2|$ represents the total number of possible term pairs and makes the similarity score comparable across text pairs of different sizes.

In the tweet contextualization task, t_1 denotes the tweet and t_2 denotes a candidate contextual text. The statistical similarity $\text{STR}(t_1,t_2)$ is therefore computed directionally from the terms of the tweet toward the terms of the candidate text. More specifically, for each term pair (X,Y) where $X \in t_1$ and $Y \in t_2$, the association rule X → Y is considered and its confidence is used to quantify the strength of the relationship. Thus, $\text{STR}(t_1,t_2)$ measures how strongly the terms in the tweet are associated with the terms in the candidate text. Since association-rule confidence is generally asymmetric, $\text{STR}(t_1,t_2)$ is not necessarily equal to $\text{STR}(t_2,t_1)$.

3.2. Semantic text similarity measure

The semantic similarity measure (denoted SSM) is based on Wikipedia2Vec, a neural network-based embedding model that learns dense vector representations of words and entities from both textual content and hyperlink structures in Wikipedia. This model is capable of capturing rich semantic relationships, including contextual similarity, synonymy, and conceptual relatedness between terms.

Unlike traditional surface-based similarity measures such as Jaccard similarity or TF-IDF, which rely on lexical overlap and term frequency, Wikipedia2Vec enables semantic comparison even when texts

share no common words. This makes it particularly suitable for short and noisy texts where lexical matching is insufficient.

To compute similarity, each text is represented as the average of its word embeddings. Although simple, this representation has been widely used in embedding-based similarity tasks as it provides a stable global semantic representation of a text while reducing noise from individual words. The similarity between two texts is then computed using cosine similarity between their vector representations, allowing the system to measure semantic closeness in a continuous vector space.

The cosine similarity is computed as (5),

$$SSM(t_1, t_2) = \frac{\vec{V}t_1 \,.\, \vec{V}t_2}{\|\vec{V}t_1\| . \|\vec{V}t_2\|} \quad (5)$$

where $\vec{V}t_1$ and $\vec{V}t_2$ are the average vector representations of text 1 and text 2, respectively, computed as (6).

$$\vec{V}t_j = \frac{1}{N}\sum_{i=1}^{N} \vec{W}_{tj} \quad (6)$$

N: The total number of words in a text j.$\vec{W}_{tj}$: The vector representation of the word W of text t_j.

3.3. Hybrid text similarity measure

As previously described, the proposed hybrid similarity measure integrates the statistical component (STR) and the semantic component (SSM) using a linear weighting scheme. This combination allows the model to benefit from both statistical co-occurrence information and deep semantic representations.

The hybrid similarity score, denoted *SHY*, is defined as (7),

$$SHY(t_1, t_2) = \alpha . STR(t_1, t_2) + (1 - \alpha) SSM(t_1, t_2) \quad (7)$$

where $\alpha \in [0, 1]$ is a weighting parameter that controls the contribution of each component.

A weighted linear combination is adopted because it provides a simple and interpretable fusion mechanism while preserving the contribution of each component. Although STR and SSM are expressed as bounded similarity scores, their distributions may differ because they are derived from different sources of evidence. Therefore, normalization does not by itself guarantee identical score distributions or balanced contributions. The relative contribution of the two components is controlled by the weighting parameter α in the proposed hybrid formulation.

4. EXPERIMENTATIONS, RESULTS AND DISCUSSION

In this section, we present the experimental setup used to evaluate the proposed hybrid similarity approach. The evaluation is conducted on the task of short text contextualization [7], [23], which can be formulated as a text similarity and ranking problem.

Given a set of tweets $\tau = \{tw_1, \ldots, tw_n\}$ and document collection $\delta = \{d_1, \ldots, d_m\}$ the objective is to generate a context c_i for each tweet $tw_i . \in \tau$. This is achieved by identifying and ranking relevant textual passages from the document collection based on their similarity to the input tweet.

For each tweet tw_i, a subset of relevant documents $\delta_p \subset \delta$ is first retrieved using an information retrieval system. Then, candidate passages extracted from δ_p are ranked according to their similarity scores computed using the proposed SHY measure. The top-ranked passages are selected to construct the context c_i.

This task is particularly suitable for evaluating text similarity methods, as it involves comparing short and often ambiguous texts with longer and information-rich documents, requiring both lexical and semantic matching capabilities. To validate the effectiveness of the proposed approach, experiments are conducted on the INEX 2013 and INEX 2014 tweet contextualization benchmarks [7], [23]. The document collection is based on a Wikipedia dump from November 2012, containing 3,902,346 articles after removing non-informative content such as notes and bibliographic references. Only non-empty articles are retained.

The two datasets differ in terms of tweet collections. The INEX 2013 dataset includes 580 English tweets covering various topics, while the INEX 2014 dataset consists of 240 English tweets selected from the RepLab campaign, focusing on four domains: automotive, banking, music, and university. While the INEX 2013 and 2014 collections are established datasets, they remain the gold-standard benchmarks for tweet contextualization, ensuring fair comparison with existing literature. Their adoption facilitates reproducible experimentation and mitigates dataset-specific bias by leveraging two independent evaluation campaigns. Furthermore, as no official benchmarks have succeeded the 2014 edition, these collections continue to serve

as a reliable and appropriate testbed for assessing the effectiveness and generalizability of the proposed similarity approach.

4.1. Evaluation metric

Contexts are evaluated according to informativeness measure. It aims at measuring how well the summary explains the tweet or how well the summary helps a user to understand the tweet's content. informativeness measure is based on lexical overlap between a pool of relevant passages and participant contexts. It is utilized for comparing the frequency distributions between a set of reference passages from an extensive document collection and the summaries generated by extracting these documents. This method serves as a compromise between Rouge and KL measures and is computed by (8) [7],

$$DIS(L,S) = \sum_{t \in L}(P-1) \times \left(\frac{\min(\log(P),\log(Q))}{\max(\log(P),\log(Q))}\right) \quad (8)$$

where

L: a set of query terms present in reference summary.

S: a set of query terms present in a submitted summary.

$f_L(t)$: the frequency of term t in reference summary.

$f_S(t)$: the frequency of term t in a submitted summary.

$P = \frac{f_L(t)}{f_L} + 1$

$Q = \frac{f_S(t)}{f_S} + 1$

f_L: the total frequency of terms in reference summary.

f_S: the total frequency of terms in a submitted summary. As the Informativeness is a dissimilarity, a low value or informativeness indicates that two summaries are similar.

There are different distributions for the reference summaries, namely: i) Unigrams made of single lemmas (after removing stop-words), ii) Bigrams made of pairs of consecutive lemmas (in the same sentence), and iii) Bigrams with 2-gaps also made of pairs of consecutive lemmas but allowing the insertion between them of a maximum of two lemmas (Also referred to as skip distribution).

These distributions are more or less strict: Unigrams only consider simple overlap, where Bigrams take into account successive words. It is important to remember for the analysis of the results that this evaluation metric is good when it is low. Figure 1 illustrates the overall workflow of the proposed tweet contextualization approach.

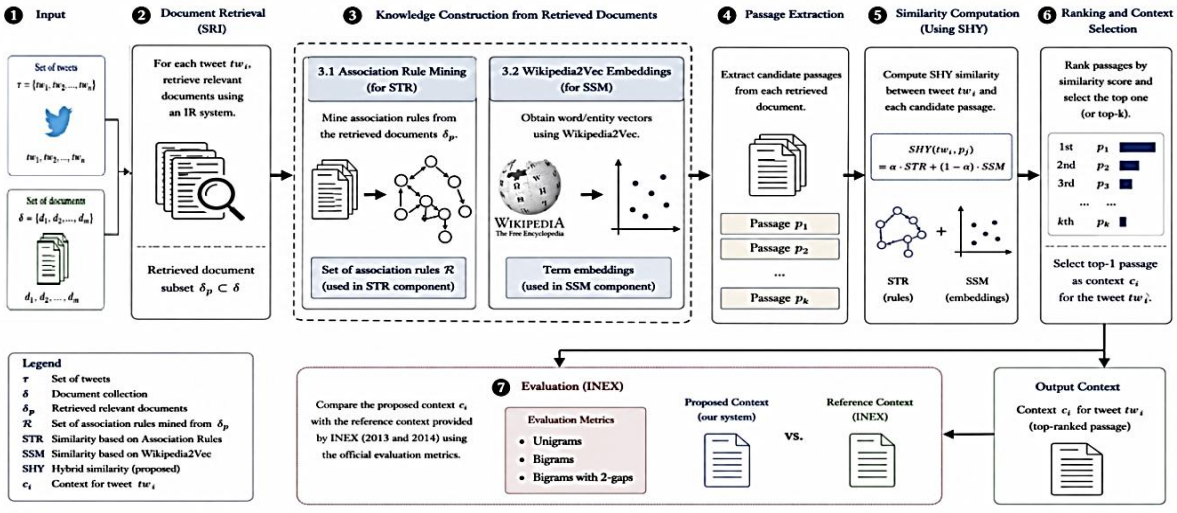


Figure 1. Workflow of the proposed approach showing how the hybrid text similarity measure is exploited for tweet contextualization

4.2. Computational complexity analysis

This section analyzes the computational complexity of the proposed tweet contextualization framework. Let Nd denote the total number of documents in the collection, Ndr the number of documents

retrieved by the information retrieval system, Np the number of candidate passages, At the average number of terms per text unit, M the number of generated association rules, and L the dimensionality of the Wikipedia2Vec embeddings.

The first stage retrieves relevant documents using a standard information retrieval (SRI) system. With indexing structures, the cost for the returned results can be approximated as O(Ndr).

The second stage performs association rule mining on the retrieved document subset. Although frequent itemset mining can have exponential complexity with respect to the number of distinct terms, minimum support and confidence thresholds, together with the reduced search space, limit the number of candidates itemsets and generated rules in practice. The cost of processing the resulting rules is approximately O(M), where M denotes the number of generated association rules. The overall mining cost is therefore denoted by O(TARM), where TARM depends on the characteristics of the transaction database and the mining parameters.

The third stage computes similarity between the tweet and candidate passages. STR requires $O(At^2)$ per passage, resulting in $O(NpAt^2)$ overall, while SSM based on Wikipedia2Vec requires O(L) per passage, resulting in O(NpL). The SHY combination adds O(Np), and ranking the candidate passages requires O(Np log Np).Thus, the overall computational complexity can be expressed as:

$$O(TARM) + O(NpAt^2) + O(NpL) + O(Np \log Np) \quad (9)$$

Since at is relatively small for short texts and passages, the similarity computation remains limited. The computational cost is mainly influenced by the association-rule mining stage and the number of candidate passages Np.

The proposed framework does not require transformer fine-tuning and relies on association-rule mining and pre-trained Wikipedia2Vec embeddings, which provides interpretability and avoids the training cost of large transformer models. However, empirical evaluation of runtime, memory consumption, and scalability is beyond the scope of this study and is left for future work.

4.3. Runs

The evaluation follows a one-group comparative design in which the individual STR and SSM components are first evaluated separately as baseline conditions, followed by the evaluation of their hybrid combination, SHY. This setup allows the contribution of the proposed integration to be assessed by comparing the performance of SHY with that of its individual components under the same experimental conditions and datasets. Accordingly, the comparison focuses on whether combining statistical and semantic information provides a measurable benefit over using either component independently.

We conducted the following runs:

- Run STR ($\alpha=1$): Passage selection is based on the STR measure, without considering the semantic measure.
- Run SSM ($\alpha=0$): Passage selection is based on the SSM measure, without considering the statistical one.
- Run SHY ($0<\alpha<1$): Passage selection is based on both STR and SSM measures.

To construct association rules between terms, a subset of 50,000 Wikipedia articles was retrieved according to the themes of the tweets using the Terrier 4.0 information retrieval platform [24]. The BM25 model was employed for document retrieval. The association rules were generated using the Charm algorithm [25], which is designed for efficient mining of closed frequent itemsets.

The selection of threshold parameters plays a crucial role in the quality of the extracted rules. The minimum support (minsupp) threshold was empirically adjusted according to the document collection size and term distribution. Considering the Zipfian distribution of textual data [26], minsupp was set to filter out infrequent terms and retain statistically meaningful associations. Similarly, the minimum confidence threshold (minconf) was varied in the range [0.4, 1.0] with a step of 0.1. Based on experimental observations, the value of minconf was set to 0.7, ensuring that the generated rules reflect strong correlations between terms. The optimal values of these parameters (minsupp and minconf) for the different datasets are reported in Table 2.

Table 2. Description of the IR test collections considered

#documents	Minsupp	Minconf	#Rules
Wikipedia corpus for INEX 2014 50000	15	0.7	378,212
Wikipedia corpus for INEX 2013 50000	15 (0.03%)	0.7	390,257

Note: The minimum support threshold is specified as an absolute occurrence count in the Charm implementation. Thus, the value 15 means that an itemset must occur in at least 15 transactions/documents. For the 50,000-document collection, this corresponds to 0.03% of the documents.

To investigate the influence of the weighting parameter α, values ranging from 0.1 to 0.9, with a step of 0.1, were evaluated within the INEX experimental setting, using the bigram 2-gaps metric to assess performance where lower scores indicate better informativeness. The corresponding results are reported in Table 3. Based on these results, α = 0.4 was selected for the subsequent analysis. Table 4 presents the obtained results on the INEX 2014 collection. The performance is evaluated using the bigram 2-gaps metric.

Table 3. Effect of α on the contributions of STR and SSM

	α = 0.1	α = 0.2	α = 0.3	α = 0.4	α = 0.5	α = 0.6	α = 0.7	α = 0.8	α = 0.9
STR contribution	10%	20%	30%	40%	50%	60%	70%	80%	90%
SSM contribution	90%	80%	70%	60%	50%	40%	30%	20%	10%
bigram 2-gaps	0.9450	0.9430	0.9434	0.9346	0.9373	0.9424	0.9461	0.9365	0.9365

Table 4. Obtained results on the INEX 2014 collection

Similarity Measure	Run	Unigram	Bigram	Bigram 2-gaps
Statistic	Association-rule-based STR	0.7984	0.9400	0.9367
Semantic	Wikipedia2Vec based SSM	0.8041	0.9545	0.9499
Hybrid	SHY hybrid measure	0.7978	0.9397	0.9346
INEX2014 official runs	INEX2014 best run(361)	0.7939	0.8908	0.8943
	INEX2014 median run(356)	0.8415	0.9696	0.9702
	INEX2014 worst run(362)	0.8686	0.9828	0.9840

The results show that the proposed SHY approach achieves competitive performance in terms of informativeness. In particular, for the Bigram 2-gaps metric, SHY obtains a score of 0.9346, which is lower than those obtained by both individual components, STR (0.9367) and SSM (0.9499). This indicates that the hybrid combination of statistical and semantic similarity provides an improvement over using either component independently. Furthermore, the SHY score of 0.9346 is lower than the official INEX2014 median (0.9702) and worst (0.9840) runs, demonstrating that the proposed approach achieves competitive performance compared with the official INEX2014 results. However, it remains higher than the best official run (0.8943), and therefore it should not be considered superior to the best official INEX2014 run for this metric. Figure 2 further illustrates the comparison between SHY and its individual components, namely the statistical method (STR) and the semantic method (SSM), as well as the official INEX 2014 runs [23].

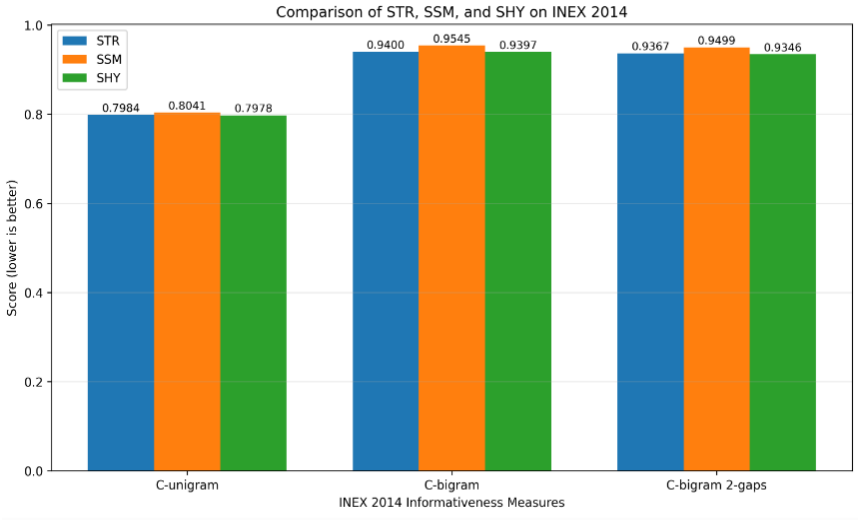


Figure 2. Comparison of our runs with the official INEX 2014 results

The strong performance of SHY can be attributed to its ability to integrate both explicit lexical features and implicit semantic relations. This complementary modeling improves the ranking of relevant passages and enhances the overall quality of the generated context.

In addition, the STR component alone also achieves competitive results. This can be explained by the effectiveness of association rule mining in capturing frequent term co-occurrences, which may reveal latent thematic relationships even in the absence of direct lexical overlap.

In light of the results obtained on the INEX 2014 collection, we further evaluate the robustness of the proposed approach on a different test set, namely INEX 2013. This evaluation aims to verify whether the best-performing runs (SHY and STR) maintain consistent performance across datasets with different characteristics. It is important to note that the tweets in INEX 2013 differ significantly from those in 2014 and are generally more challenging to process.

In particular, many tweets in INEX 2013 are highly time-sensitive, referring to events occurring in early 2013. This temporal dependency makes information retrieval more difficult when using a static Wikipedia snapshot (November 2012). In addition, some tweets are highly abstract or metaphorical, which increases the difficulty of automatic contextualization.

Table 5 presents the results obtained by the proposed SHY approach on the INEX 2013 collection. It reports the performance of the two individual components, STR and SSM, as well as their hybrid combination, SHY, across the considered evaluation metrics. For comparison, the best, median, and worst results reported by the official INEX 2013 runs [7] are also included. These results provide an assessment of the contribution of the proposed hybridization and its performance relative to the official INEX 2013 results. Both SHY and STR achieve performance above the median of the official runs. Moreover, according to the official ranking, SHY and STR rank 8th and 9th, respectively, among the 21 submitted runs [7].

Table 5. Obtained results on the INEX 2013 collection

Similarity Measure	Run	Unigram	Bigram	Bigram 2-gaps
Statistic	Association-rule-based STR	0.8022	0.9310	0.9298
Semantic	Wikipedia2Vec based SSM	0.8049	0.9343	0.9331
Hybrid	SHY hybrid measure	0.7954	0.9298	0.9283
INEX2013 official runs	INEX2013 best run	0.7820	0.8810	0.8861
	INEX 2013 median run	0.8673	0.9540	0.9575
	INEX2013 worst run	0.9981	0.9999	0.9999

The results obtained on the INEX 2013 collection show that the proposed SHY method maintains a similar overall behavior across the two datasets. Although its performance is slightly lower than that obtained on INEX 2014, this difference may be related to the characteristics of the INEX 2013 tweets, which can be more time-sensitive and semantically ambiguous. Despite these challenges, the obtained results are above the median of the official runs, indicating that the proposed approach can handle noisy and complex short texts to a certain extent. However, the current evaluation does not include comparisons with recent transformer-based or sentence-embedding approaches. Therefore, the results should not be interpreted as evidence of superiority over modern semantic similarity methods. Instead, they provide evidence that combining the statistical STR component with the semantic SSM component can improve performance over the individual components.

The results have potential implications for short-text retrieval and information-filtering systems, particularly in language-based robotics. In contextual retrieval of task instructions, SHY could help identify relevant instructions or knowledge entries from textual repositories, even when the query and the retrieved content use different lexical expressions. Similarly, in dialogue-based human–robot interaction, it could support the retrieval of contextually related responses or knowledge from previous dialogue turns. The same characteristics could be useful in other applications, such as social media monitoring and customer-support knowledge retrieval, where short queries may not share the exact wording of relevant content.

SHY can also serve as a reusable teaching example for hybrid NLP methods in robotics and automation courses. Its simple formulation allows students to implement and compare statistical, semantic, and hybrid similarity approaches while examining the trade-offs between interpretability, semantic richness, and computational cost. However, these applications remain prospective, as the present study was evaluated only on the INEX tweet contextualization task. Future work should therefore validate SHY on domain-specific and robotics-oriented datasets, including instruction retrieval and human–robot dialogue, and evaluate its computational requirements on larger or resource-constrained systems.

5. CONCLUSION

In this paper, we proposed SHY, a hybrid text similarity measure that combines a statistical component based on association rule mining with a semantic component based on Wikipedia2Vec embeddings. The main objective of this hybridization is to exploit the complementary strengths of statistical and semantic information for similarity computation, particularly for short and sparse texts.

The statistical component captures explicit term relationships through corpus-level co-occurrence patterns, while the semantic component captures latent semantic relationships using distributed

representations. Their combination provides two complementary sources of evidence for estimating similarity between a short text and candidate contextual information. The proposed framework also provides reusable methodological knowledge for designing lightweight hybrid NLP methods, including the integration of explicit corpus-level associations with pretrained semantic representations without task-specific fine-tuning.

The proposed approach was evaluated on the INEX 2013 and INEX 2014 tweet-contextualization datasets. The results show that SHY consistently improves over its individual STR and SSM components, although the observed improvements are relatively modest and were not statistically significant. Moreover, SHY remains below the best official INEX runs on both datasets. These results therefore provide evidence that combining statistical and semantic information can improve the performance of the individual components, but they do not establish superiority over the best existing INEX results or modern semantic similarity approaches.

Future work will focus on evaluating SHY against recent transformer-based and sentence-embedding models, investigating adaptive weighting strategies, and assessing its generalization across additional datasets and textual domains. In particular, robotics-specific validation will be conducted on instruction-retrieval and human–robot dialogue datasets to determine whether SHY can support language-based robotic applications. Further research will also investigate the integration of textual similarity with multimodal information, such as visual, speech, and sensor-based representations, to support richer contextual understanding in robotics and automation.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Meriem Amina Zingla	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Chérifa Boudia		✓				✓		✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing -**O**riginal Draft
E : Writing - Review &**E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are not publicly available due to the original source website being no longer accessible, but are available from the corresponding author on reasonable request.

REFERENCES

[1] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, "A survey on the techniques, applications, and performance of short text semantic similarity," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 5, Mar. 2021, doi: 10.1002/cpe.5971.

[2] Z. H. Amur, Y. Kwang Hooi, H. Bhanbhro, K. Dahri, and G. M. Soomro, "Short-text semantic similarity (STSS): Techniques, challenges and future perspectives," *Applied Sciences (Switzerland)*, vol. 13, no. 6, p. 3911, 2023, doi: 10.3390/app13063911.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.

[4] I. Yamada *et al.*, "Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia," in *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations*, Oct. 2020, pp. 23–30, doi: 10.18653/v1/2020.emnlp-demos.4.

[5] P. Yang, H. Wang, J. Yang, Z. Qian, Y. Zhang, and X. Lin, "Deep learning approaches for similarity computation: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 7893–7912, 2024, doi: 10.1109/TKDE.2024.3422484.

[6] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. of the 20th International*

Conference on Very Large Data Bases (VLDB'94), 1994, vol. 1215, pp. 487–499.
[7] P. Bellot, V. Moriceau, J. Mothe, E. S. Juan, and X. Tannier, "Overview of INEX tweet contextualization 2013 track," in *CEUR Workshop Proceedings*, Sep. 2013, vol. 1179.
[8] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *Journal of Artificial Intelligence Research*, vol. 37, pp. 141–188, 2010, doi: 10.1613/jair.2934.
[9] A. Dhokar, L. Hlaoua, and L. Ben Romdhane, "Tweet contextualization: combining sentence extraction, sentence aggregation and sentence reordering to enhance informativeness and readability," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–16, 2021, doi: 10.1007/s13278-021-00724-4.
[10] E. L. Pontes, J. M. Torres-Moreno, S. Huet, and A. C. Linhares, "Tweet contextualization using continuous space vectors: Automatic summarization of cultural documents," in *CEUR Workshop Proceedings*, 2016, vol. 1609, pp. 1238–1245.
[11] A. C. Linhares, "An automatic greedy summarization system at INEX 2013 tweet contextualization track," in *CEUR Workshop Proceedings*, Sep. 2013, vol. 1179.
[12] R. Deveaud and F. Boudin, "Contextualisation automatique de Tweets à partir de Wikipédia," in *CORIA 2013 - Conference in Research and Applications Informations - 10th French Information Retrieval Conference*, Apr. 2013, pp. 125–140.
[13] J. M. Torres-Moreno, "Three statistical summarizers at CLEF-INEX 2014 Tweet contextualization track," in *CEUR Workshop Proceedings*, Sep. 2014, vol. 1180, pp. 565–573.
[14] T. Hosking, H. Tang, and M. Lapata, "Hierarchical indexing for retrieval-augmented opinion summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1533–1555, 2024, doi: 10.1162/tacl_a_00703.
[15] T. Schopf, E. Gerber, M. Ostendorff, and F. Matthes, "AspectCSE: Sentence embeddings for aspect-based semantic textual similarity using contrastive learning and structured knowledge," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2023, pp. 1054–1065, doi: 10.26615/978-954-452-092-2_113.
[16] Z. Qiu, G. Huang, X. Qin, Y. Wang, J. Wang, and Y. Zhou, "A hybrid semantic representation method based on fusion conceptual knowledge and weighted word embeddings for English texts," *Information (Switzerland)*, vol. 15, no. 11, p. 708, 2024, doi: 10.3390/info15110708.
[17] J. Zang and H. Liu, "Improving text semantic similarity modeling through a 3D Siamese network," in *Frontiers in Artificial Intelligence and Applications*, Sep. 2023, vol. 372, pp. 2970–2977, doi: 10.3233/FAIA230612.
[18] X. Zhang *et al.*, "TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations at Twitter," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023, pp. 5597–5607, doi: 10.1145/3580305.3599921.
[19] E. Hassan, A. S. Talaat, and M. A. Elsabagh, "Intelligent text similarity assessment using Roberta with integrated chaotic perturbation optimization techniques," *Journal of Big Data*, vol. 12, no. 1, 2025, doi: 10.1186/s40537-025-01233-3.
[20] A. Vasileiou and O. Eberle, "Explaining text similarity in transformer models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, Jun. 2024, vol. 1, pp. 7852–7866, doi: 10.18653/v1/2024.naacl-long.435.
[21] S. Saeed, Q. Rajput, and S. Haider, "SUMEX: A hybrid framework for semantic textual similarity and explanation generation," *Information Processing and Management*, vol. 61, no. 5, p. 103771, 2024, doi: 10.1016/j.ipm.2024.103771.
[22] L. Kumar, N. S. Upadhye, and K. Piedy, "Advances and challenges in semantic textual similarity: A comprehensive survey," *arXiv preprint arXiv:2601.03270*, 2025.
[23] P. Bellot, V. Moriceau, J. Mothe, E. S. Juan, and X. Tannier, "Overview of INEX tweet contextualization 2014 track," in *CEUR Workshop Proceedings*, Sep. 2014, vol. 1179, pp. 494–500.
[24] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson, "Terrier information retrieval platform," in *Lecture Notes in Computer Science*, 2005, vol. 3408, pp. 517–519, doi: 10.1007/978-3-540-31865-1_37.
[25] M. J. Zaki and C.-J. Hsiao, "CHARM: An efficient algorithm for closed itemset mining," in *Proceedings of the 2nd SIAM International Conference on Data Mining (SDM)*, 2002, pp. 457–473, doi: 10.1137/1.9781611972726.27.
[26] A. B. Hollingshead and G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*, vol. 14, no. 6. Cambridge, MA, USA: Addison-Wesley Press, 1949.

BIOGRAPHIES OF AUTHORS



Meriem Amina Zingla received the master's degree in computer science from the University of Mascara, Algeria, in 2013. She was awarded a scholarship to pursue her doctoral studies at the University of Tunis El Manar, Tunisia, where she obtained the Ph.D. degree in Computer Science in 2018. She is currently an associate professor with the Department of Computer Science, University of Mascara, Algeria. Her research interests include data mining, information retrieval, text mining, image processing, and healthcare data analytics. She can be contacted at meriem.zingla@univ-mascara.dz.



Chérifa Boudia received her Ph.D. degree in Computer Science from Oran 1 Ahmed Ben Bella University, Algeria, in 2020. She is currently an associate professor with the Department of Computer Science, Mustapha Stambouli University, Mascara, Algeria. She has authored over 25 international journal and conference papers. Her research focuses on artificial intelligence, generative AI in computer science education, collaborative programming, and ICT-enhanced learning. She is interested in the application of AI and Generative AI to improve computer science education and is an active member of the academic community promoting innovative teaching practices and the integration of emerging AI technologies in higher education. She can be contacted at cherifa.boudia@univ-mascara.dz.