

RESEARCH ARTICLE



Enhanced Short-Text Clustering Based on a Multirepresentation Fusion Model

Ali Sabah^{1,*}, Dhurgham Ali Mohammed¹ and Zaid Alaa¹

¹ Department of Computer Science, University of Kufa, Iraq

Abstract: Short-text clustering has become increasingly important because of the exponential growth of textual data. However, traditional methods struggle with sparsity, limited context, and poor cluster cohesion. Recently, innovations have made word embedding algorithms, including Word2Vec, GloVe, and FastText, available for effectively learn semantic relationships. Although these methods have advantages, they tend to concentrate on semantics area. These limitations are addressed in this study by adopting an enhanced short-text clustering method based on multirepresentation. Specifically, this study integrates heterogeneous embeddings into a unified framework using weighted L2-normalized concatenation, followed by Principal Component Analysis (PCA)-based dimensionality reduction that retains 95% of the variance, thereby create a unified semantic representation. Unlike simple concatenation approaches of embeddings, pre-space normalization operations are added before the fusion operation to avoid the presence of dominant embedding bias. The proposed model was evaluated on four benchmark datasets, namely, StackOverflow, AgNews, BioASQ, and SearchSnippet, via the k-means and k-medoids clustering algorithms. The results demonstrate substantial improvements in the clustering quality metrics. Compared with the results of the baseline models Bag of Words (BoW)/Term Frequency-Inverse Document Frequency (TF-IDF), the average gains were statistically significant: 29% in silhouette score, 8% in adjusted Rand index, and 8% in normalized mutual information over the best (single) embedding baseline (GloVe with k-medoids) and 63%, 35%, and 40%, respectively, over traditional BoW/TF-IDF methods. Most improvements are found to be statistically significant at $p < 0.05$. Additionally, PCA was applied to reduce noise and dimensionality, potentially improving clustering outcomes.

Keywords: word embedding, short-text clustering, data representation, natural language processing

1. Introduction

The proliferation of big data has been accompanied by exponential growth in short texts originating from a multitude of sources, including social media platforms, microblogging sites, online forums, and news aggregators. These short texts, characterized by their brevity and limited context, pose unique challenges to traditional text processing and clustering techniques [1, 2]. The intrinsic sparsity of these texts often complicates accurate grouping on the basis of shared topics or themes, which affects cohesion within clusters. This issue is particularly problematic given the need for effective methods to organize and analyze this information for applications such as trend analysis, sentiment detection, and information retrieval. Recent advancements in natural language processing (NLP) have introduced various techniques to address the challenges associated with short-text clustering. Traditional clustering methods such as k-means and hierarchical clustering have been adapted to work with short texts, which often integrate word embeddings to enhance feature representation [3, 4]. Word embeddings such as Word2Vec and GloVe have revolutionized text analysis by capturing semantic relationships between words in a continuous vector space. This scheme allows

for a nuanced understanding of textual data, enabling clustering algorithms to group similar texts effectively.

The current approaches used often rely on a single-word embedding technique, which may restrict the richness of the semantic information captured [5, 6]. Every embedding method has its strengths and weaknesses: Word2Vec excels in capturing contextual word associations [4, 7–9], GloVe leverages global co-occurrence statistics to provide a comparatively broad context, and FastText includes subword information. These methods can handle out-of-vocabulary words and capture morphological nuances. Consequently, using one embedding structure presents a main research gap: There is an inherent bias involved in the model with regard to specific semantic features (e.g., global statistics in GloVe, local context in Word2Vec), and complementary signals are disregarded, resulting in incomplete semantic representations for limited short texts. While modern architectures for transformer-based models (e.g., BERT, RoBERTa) have profound knowledge of context, they can be quite expensive in terms of computation and also often need fine-tuning or labeled data for optimal performance. In contrast, our proposed multirepresentation framework is aimed at targeting large-scale, purely unsupervised scenarios, where computational efficiency and resource constraints are paramount concerns, which fills a critical gap for lightweight yet semantically rich clustering solutions [6, 10, 11]. Traditional multi-embedding methods for

*Corresponding author: Ali Sabah, Department of Computer Science, University of Kufa, Iraq. Email: alis.alsaadi@uokufa.edu.iq

text representation have three drawbacks: (i) they simply concatenate embeddings without normalization, so that one embedding can overwhelm the other; (ii) they average over the embeddings, which results in loss of complementary semantic signals [4, 6]; and (iii) they require heavy end-to-end training for attention-based fusion [1]. There are other recent short-text clustering applications, like Asyaky & Mandala [12] or Petukhova et al. [13] that do not consider multi-embedding fusion, which makes them incomparable. ESTCM is able to address the above limitations (i)–(iii) due to its innovative three-step pipeline: per-space L2 normalization, weighted concatenation, and principal component analysis (PCA) compression. This pairing has not been formally set up for short-text clustering before.

To fill in the identified gaps, the main target of this study is to construct an “enhanced short-text clustering based on the multirepresentation” (ESTCM) framework based on Word2Vec, GloVe, and FastText into a unified feature space. This approach is intended to take advantage of their complementary strength in order to overcome the sparsity issue and contribute to better clustering cohesiveness, without the computational cost of large pretrained language models. Employing the proposed ESTCM allows the model to capture a rich set of semantic relationships within the data, thereby improving the overall clustering performance. Unlike previous methods where multi-embedding fusion has been defined as straightforwardly concatenating the embeddings of a system, ESTCM solves the scale dominance issue by normalizing per space and then orthogonally normalizing semantic subspaces by concatenation via PCA compression per space rather than averaging the features.

The integration of multiple word embedding techniques allows for a comprehensive representation of short texts. Each embedding method contributes unique insights into the semantic structure of the data. For example, Word2Vec uses a predictive model to learn word associations on the basis of their context. This scheme can help to effectively understand relationships between the words in short texts. GloVe employs global word co-occurrence statistics, which can offer a broad context for enhancing the semantic representation of words and their relationships. FastText considers subword information, which is particularly beneficial for capturing morphological variations and handling out-of-vocabulary words, which can augment the representation of short texts.

The proposed ESTCM is rigorously evaluated on the same datasets and uses the same clustering methods as those described in the literature. This approach ensures a fair comparison and allows for a direct assessment of the improvements in clustering quality and cohesion achieved via the proposed modifications. Maintaining consistency in the evaluation framework enables the collection of robust evidence pertaining to the effectiveness of the enhanced model. Considerable improvements are expected in terms of clustering performance, such as cluster cohesion and improved ability to group similar texts semantically. The proposed model seeks to address the challenges of sparsity and limited context typical in short texts. This scheme ultimately enhances the accuracy of the clustering results. This research not only advances the existing body of knowledge in short-text clustering but also provides a new framework for leveraging multiple word embedding techniques to improve clustering outcomes. The findings of this study are expected to have notable implications for various applications, including trend analysis, data summarization, and information retrieval. Ultimately, the vast amounts of short-text data generated in today’s digital landscape can be comprehensively explained. In summary, the proposed enhancements

to short-text clustering via the proposed ESTCM represent an important step in overcoming the limitations posed by the unique characteristics of short texts. Integrating multiple word embedding techniques opens new possibilities for clustering, especially new insights into text analysis and its applications.

The main aim of this research is to objectively examine whether a formally defined multirepresentation fusion of complementary word embedding spaces—Word2Vec, GloVe, and FastText—leads to a document representation shown to be measurably better than single-representation approaches for short-text clustering tasks in terms of statistically validated cluster validity index improvements measured from a variety of benchmarking datasets. The main research contribution of this study is illustrated as follows:

- 1) This research proposes ESTCM (a formally specified multirepresentation fusion framework to integrate Word2Vec, GloVe, and FastText through a 3-step pipeline: per-space L2 normalization, weighted concatenation, and PCA-based dimensionality reduction). This addresses the methodological gap in the previous multi-embedding-based approaches, which either omitted normalization (causing dominance), used averaging (losing diversity), or required expensive end-to-end training.
- 2) This research shows that the single-embedding representation generates incomplete semantic coverage for short-text clustering because each embedding method learns the word meaning in a partial mode or biased manner. The proposed multirepresentation fusion adopts complementary semantic signals simultaneously, which achieves significant performance improvement in cluster cohesion, label alignment, and mutual information for all three of the four benchmarks.
- 3) This work provides the first systematic empirical evaluation of weighted-concatenation multi-embedding fusion in the brief text categorization scenario, which spans four benchmark datasets situated in the technical, generic, biomedical, and web content domains.
- 4) Validate all reported improvements using Wilcoxon tests for signed-rank statistical significance using 10 independent runs by verifying that observed improvements over all baselines are statistically significant at $p < 0.05$.

2. Literature Review

Short-text clustering has emerged as an important area of research because of the increasing volume of short texts generated from various sources, such as social media, forums, and news articles. The distinct challenges posed by the sparsity and high dimensionality of short texts have led to the development of various innovative clustering techniques. Short-text clustering is challenging in the field of NLP because of the difficulty in learning discriminative representations with limited information [1, 2, 9–12]. In this section, a critical review of existing short-text clustering approaches is conducted, and the limitations of the current state of the art are identified.

Asyaky and Mandala [12] improved the performance of the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) algorithm for short-text clustering. They proposed a new approach to integrate word embedding techniques with uniform manifold approximation and projection (UMAP) for dimensionality reduction. Their work highlighted the limitations of traditional clustering methods when applied to short texts; the usual approach is often hindered by high dimensionality and

sparse data. Given that word embeddings are leveraged, semantic relationships between words can be captured. Their use of UMAP reduced the dimensionality of the data, subsequently establishing its suitability for HDBSCAN.

Grootendorst et al. [9] adopted BERTopic, a neural topic modeling framework that combines transformer-based embeddings with a class-based TF-IDF procedure, to address semantic limitations in traditional methods. Their approach involved the following: (1) generating document embeddings via pretrained language models (e.g., SBERT), (2) clustering embeddings via UMAP and HDBSCAN to handle high-dimensional sparsity, and (3) deriving topic representations through a modified TF-IDF that operates at the cluster level ("c-TF-IDF"). However, these approaches are accompanied by considerable computational needs, large-scale pretraining, Graphical processing unit (GPU)-based inference, and large memory overhead that restricts their use in resource-constrained applications and special domains. Furthermore, transformer-based representations are mostly opaque, so that the outcomes of clustering can hardly be attributed to specific linguistic features. These limitations highlight the need to develop efficient, interpretable static embedding fusion approaches such as ESTCM.

Petukhova et al. [13] investigated text clustering via embeddings from large language models (LLMs) and compared the efficacy of these models with that of traditional methods such as TF-IDF. Their experiments found that the LLM embeddings, especially GPT-3.5 Turbo of OpenAI, are stronger in cluster purity and silhouette score and in structured texts. BERT is a very good open-source alternative. In comparison, dimensionality reduction based on summation and relatively large model sizes did not give consistent outcomes. The clustering quality and computational cost tradeoff was emphasized in their study. K-means and FuzzyCM performances are conditional on the complexity of the dataset.

Rudiger et al. [14] examined the sparsity and lack of context in the short texts by finding out the most important terms that describe the underlying issues of the data best. Their methodology consisted of creating a term-topic matrix and ranking the terms most representative of a topic, with the help of a ranking algorithm. Their study highlighted the importance of topic representative terms in improving the quality of short-text clustering, which subsequently provided a practical solution for applications, such as social media analysis and document organization.

Zhang et al. [15] used an attentive representation learning model to integrate representation learning and clustering into a unified framework. The model leveraged cluster-level attention and captured correlations between text and cluster representations. They also implemented unsupervised learning of word and cluster embeddings. The robustness of their method was enhanced by the integration of adversarial training, in which perturbations are injected into cluster representations. While the attention mechanism has a good ability to capture cluster-level correlations, the model needs to be trained end to end from the ground up, which acts as a barrier to end-to-end training and thus increases the amount of computation.

Pugachev and Burtsev [16] adopted a new approach to improve clustering quality by semantically enriching texts via two main methods, namely, neural-based distributional models (e.g., Word2Vec) and external knowledge resources (e.g., BabelNet). Whereas semantic enrichment mechanisms have been found to be effective in domain-specific contexts, the reliance on external knowledge bases entails domain portability limitations and vocabulary coverage limitations, limiting generalization to novel domains.

The DeepXML framework by Dahiya et al. [17] addresses extreme multilabel classification (XML) for short texts by decomposing the task into four modular subtasks: (1) learning intermediate features via label clustering, (2) hard negative sampling via sublinear search structures, (3) feature transfer with spectral norm constraints, and (4) log-time classifier training. Although it focuses on classification, DeepXML's emphasis on handling sparsity via hard negative mining and semantic ambiguity via label-aware feature transfer similarly encounters challenges in short text.

Meddeb and Romdhane [18] investigated issues pertaining to topic modeling for short texts, which are hindered by data sparsity due to their limited length. A new method called topic-regularized nonnegative matrix factorization (TRNMF) was used to leverage pretrained word embeddings and document clustering, from which topic inference could be improved. In TRNMF, word co-occurrence regularization and document similarity regularization were integrated into a nonnegative matrix factorization framework. Despite its effectiveness in topic modeling, TRNMF is limited to a single source of embeddings and lacks the necessary morphological coverage for out-of-vocabulary terms that are common in short text.

Zhang et al. [1] adopted a deep multi-embedded self-supervised model to leverage multiple embeddings and capture diverse semantic representations of short texts. By exploiting the intrinsic structure of the data, their model could integrate self-supervised learning and enhance the quality of text representations. While the attention mechanism has a good ability to capture cluster-level correlations, the model needs to train end to end from the ground up, which acts as a barrier to end-to-end training and thus increases the amount of computation. While previous multirepresentation methods share the same motivation, which is toward the combination of multiple embedding sources, they combine embeddings without per-space normalization, lack any formal fusion specification, and work with computationally expensive pretraining architectures. By contrast, ESTCM deals with these lacunae by per-space L2 normalization, weighted concatenation, and PCA-based compression, composing a formalized, reproducible, and computationally efficient pipeline that directly addresses the methodological gap in the literature. The multirepresentation strategy proposed in this study is directly influenced by the aforementioned insights. The challenge of sparsity is addressed by integrating multiple word embedding techniques, namely, Word2Vec, GloVe, and FastText, to capture a much richer set of semantic relationships within short texts. The proposed ESTCM was designed to mitigate the limitations of single-embedding models and enhance the overall clustering performance. Due to the advantages of each embedding method, Word2Vec to create contextual word associations, GloVe to evaluate global co-occurrence statistics, and FastText to provide subword details can provide a detailed and complete representation of short texts. The plan suggested in this paper is likely to significantly enhance cluster unity and arrange similar texts in semantically pure clusters. To summarize, the previous multi-embedding approaches have three drawbacks: without per-space normalization, embedding dominance occurs; they use averaging rather than concatenation, which results in loss of complementary information; and they use very costly end-to-end training. ESTCM tackles all three with a lightweight, unsupervised pipeline. The novelty is that it combines L2 normalization, concatenative fusion, and PCA compression together as the basis for short-text representation, which are

principle-based. Based on this, the particular problems observed in the literature may be overcome.

3. Enhanced Short-Text Clustering Model

An innovative approach is proposed in this study to improve clustering performance in the context of short-text clustering by utilizing multirepresentation techniques with distributed word embeddings.

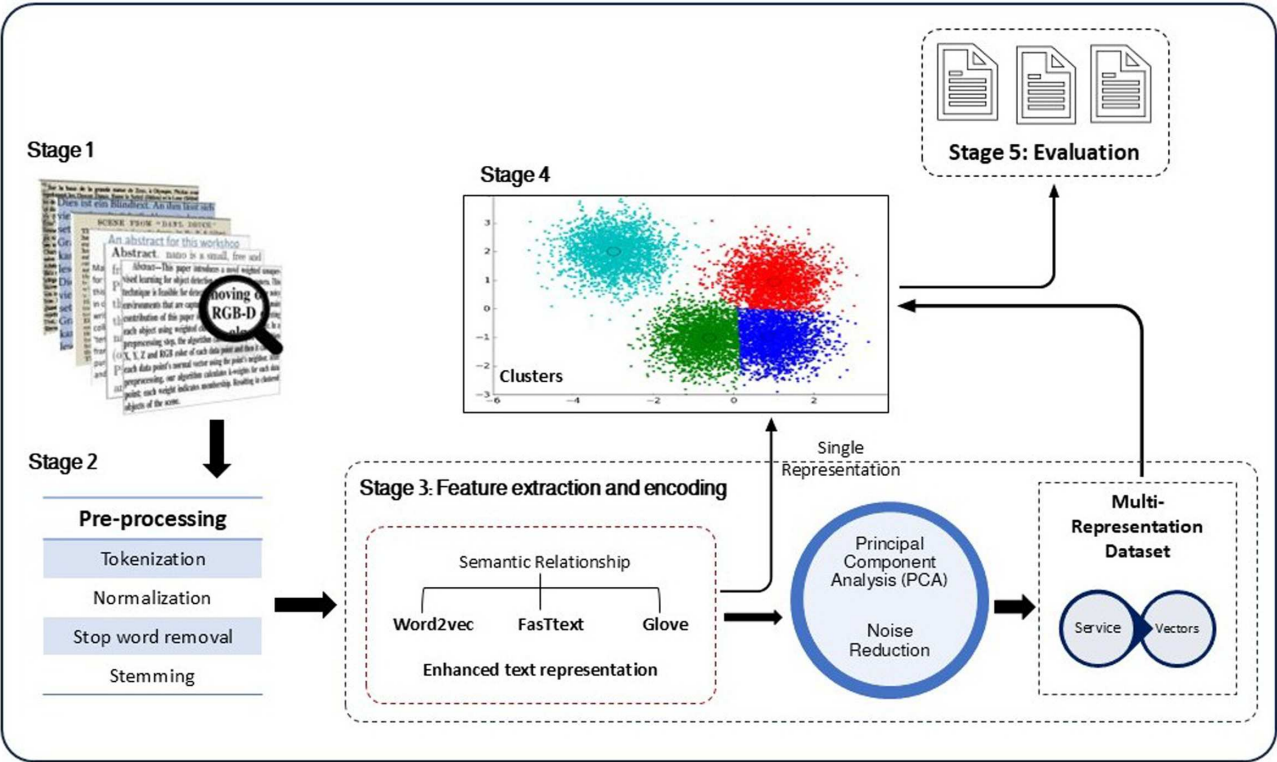
The advantages of using Word2Vec, FastText, and GloVe are leveraged. The findings indicate that clustering performance varies considerably across different datasets. Therefore, a primary focus of this study is to compare the behaviors of baseline models and those enhanced via the proposed ESTCM across all datasets. First, clustering models, specifically k-means and k-medoids, are investigated via traditional vector representations for clustering short texts. Second, the proposed ESTCM is implemented with the same clustering models to assess its impact. The objective is to identify the optimal clustering models and demonstrate how integrating multiple embeddings can enhance clustering outcomes. The aim is to utilize a comprehensive framework for enhancing short-text clustering by integrating different strategies for multirepresentation and analyzing the resulting clustering performance across diverse datasets. This research aims to gain insights into the effectiveness of the proposed approach in terms of short-text clustering. This phase outlines the framework that includes all the steps involved in the implementation of effective short-text clustering models, which are tailored specifically for the datasets utilized in this research. The stages include (i) dataset collection, (ii) preprocessing tasks, (iii) data representation, and (iv) the selection of clustering methods. Figure 1 shows the proposed implementation

framework for short-text clustering models. The proposed framework consists of several key activities. The initial stage is data selection, which involves the use of specific datasets (StackOverflow, AgNews, BioASQ, and SearchSnippet) for validation. The SearchSnippet dataset may not have a direct download link and is typically found within specific research publications or challenges. The second stage applies preprocessing methods to the datasets. In this research, various clustering models, including k-means and k-medoids, are evaluated to identify optimal conditions for short-text clustering. The challenges posed by short texts and the limited number of snippets are also addressed by adopting the proposed ESTCM. This approach integrates the following word embedding techniques to enhance the representation of short texts:

- 1) Word2Vec utilizes a predictive model to learn word associations based on context, permitting the effective capture of the correlations of words in short texts.
- 2) GloVe utilizes global word co-occurrence statistics from the entire corpus, produces word vectors that reflect the overall linguistic structure, and improves the overall contextual understanding.
- 3) FastText includes subword information, which is particularly advantageous for capturing morphological variations and addressing out-of-vocabulary words. Ultimately, the short-text representation is enriched.

In the third stage, the multirepresentation strategy is used to improve the representation and semantic understanding of the datasets. Finally, the efficacy of single-word representation techniques is compared with that of the proposed ESTCM. Their impacts on clustering performance across the different datasets

Figure 1
Proposed implementation framework for short-text clustering models



were assessed. The goal is to demonstrate how integrating multiple embeddings not only provides a comprehensive understanding of short texts but also leads to considerable improvements in clustering outcomes.

3.1. Document representation

Clustering short texts, such as snippets from web searches, poses notable challenges within the NLP field [1]. Traditional clustering methods often struggle to accurately group these snippets because of the limited statistical information for effective similarity measurements. A deep learning-based representation capable of capturing syntactic and semantic features is essential for improving similarity assessments. An efficient short-text clustering approach should prioritize advanced text representations while enhancing the semantic relationships among words [1, 19–22]. Consequently, this study explores two types of data representations: (i) traditional document representation and (ii) enhanced document representation.

3.1.1. Traditional document representation

The most traditional document representation used in short-text clustering is the BoW model. This model represents documents as vectors in a high-dimensional space, where each dimension corresponds to a unique word in the vocabulary, and the value in each dimension indicates the frequency of the word in the document. Although BoW is favored for its simplicity and effectiveness in capturing words without considering their order or grammatical structure, it often results in high-dimensional sparse representations that can challenge clustering algorithms [21, 23]. Additionally, a single representation such as BoW might be insufficient for accurately capturing the semantics and contextual information of the documents; thus, it is less effective for tasks that need nuanced understanding, such as clustering short texts. Variants such as TF-IDF have been developed to weigh the importance of words on the basis of their distribution across the corpus, and other advanced methods, such as word embeddings and hybrid approaches, are increasingly being employed.

3.1.2. Proposed enhanced short-text document representation based on the proposed ESTCM

An enhanced document representation method is adopted to leverage the multirepresentation strategy to improve short-text clustering. Traditional models such as BoW often fail to capture the rich semantic and contextual information typical in short texts, resulting in suboptimal clustering outcomes. The approach proposed in this study integrates the state-of-the-art word embedding techniques Word2Vec, FastText, and GloVe, which can provide deep insights into the relationships between words and their meanings.

Word2Vec employs neural networks to create high-dimensional word vectors to capture semantic similarity on the basis of co-occurrence in a large corpus [4, 7, 8]. This scheme ensures that words with similar meanings are closer in the vector space, which can substantially improve clustering performance. The skip-gram model is a neural network-based approach that learns word embeddings by predicting context words from a given target word. The method relies on a simple feedforward architecture with three layers: an input layer, a hidden layer, and an output layer. The objective is to maximize these probabilities over the training corpus. During training, the model receives a context

window around a target word (i.e., the center word). The goal of the skip-gram model is to maximize the likelihood of context words that appear in the window given the target word. Mathematically, this expression denotes the maximization of the mean log probability of context words relative to the target word.

$$\text{Maximize : } \frac{1}{T} \sum_{t=1}^T \sum_{c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t),$$

where T is the training words number, c is the size context, and $p(w_{t+j} | w_t)$ is determined by the model's parameters. A unique aspect of the skip-gram model is that the weight matrices for the input and output layers are shared among all context words, which helps effectively learn the semantics across large datasets by minimizing the number of parameters.

FastText extends this concept by considering subword information, from which accurate representations of rare or misspelled words are generated [24, 25]. This scheme is particularly useful for short texts that often contain varied vocabularies. FastText, developed by Facebook's AI Research Lab, also involves several equations for its word representation strategy. In contrast to traditional word embedding techniques that treat each word as a single entity, FastText considers subword information (i.e., n-gram characters) to build word vectors. The word vector is represented as the sum of the vectors of its n-grams. For a word w with its n-grams represented as $g_1, g_2, \dots, g_k$, the word vector v_w can be expressed as follows:

$$v_w = \frac{1}{|g|} \sum_{g \in g_w} v_g,$$

where v_w is the word vector for w , g_w is the set of n-grams in the word w , v_g is the vector representation for the n-gram g , and $|g|$ is the number of n-grams for the word.

GloVe utilizes global word co-occurrence statistics from the entire corpus, produces word vectors that reflect the overall linguistic structure, and improves the overall contextual understanding. GloVe involves the use of specific equations to model the relationships between words on the basis of their co-occurrence probabilities. A key aspect of GloVe is the formulation of the cost function used to train the word vectors, which can be expressed mathematically. The basic objective of GloVe can be described as follows:

$$J = \sum_{i,j=1}^V f(X_{i,j})(W_i^T W_j + b_i + b_j - \log(X_{i,j}))^2,$$

where J is the overall cost function that GloVe seeks to minimize, V is the vocabulary size (the number of unique words in the entire corpus), and X_{ij} is the number of times word j occurs in the context of word i (i.e., their co-occurrence count). Furthermore, W_i and W_j are the word vectors for words i and j , respectively; b_i and b_j are biases associated with words i and j ; and $f(X_{i,j})$ is a weighting function applied to the co-occurrence counts. Common choices for $f(x)$ are $f(x) = x^\alpha$ for $x \geq 100$ and $f(x) = 1$ for $x < 100$.

In this research, Word2Vec, GloVe, and FastText embeddings are set to $d = 300$ dimensions and normalized to until ℓ_2 norm before fusion, as illustrated in Section 3.2. These three powerful embeddings are integrated in this study. The proposed ESTCM not only mitigates issues associated with sparsity and loss of information in single-representation models but also enriches the document representation by capturing different aspects of word semantics and contextual relationships. This comprehensive representation forms

the basis for highly robust clustering results, which can ultimately enhance the discovery of meaningful patterns within short-text datasets. In summary, the new contribution of this research hinges on the strategic integration of Word2Vec, FastText, and GloVe into a unified framework for document representation, which enables clustering algorithms to perform effectively and accurately in the nuanced landscape of short texts. The proposed ESTCM is expected to yield superior clustering quality while maintaining computational efficiency. A new benchmark for short-text clustering methodologies is subsequently established.

Table 1 presents the setting for the dimensionality encoding and pre-fusion normalization of the ESTCM pipeline. All three embeddings' spaces are set to $d = 300$ dimensions fair to 900-dimension fused vector following weighted concatenation. Per-space L2 normalization is performed independently in each of the embedding spaces before fusion, given that each embedding space is contributing equally. The fused vector is then compressed to k dimensions using PCA with 95% of total variance. Dataset-specific k^* values are reported in Table 2.

3.2. Multirepresentation fusion and PCA management

The key to the proposed ESTCM model is the effective combination of three different embedding vectors into one unified document representation. This process consists of three consecutive stages: feature normalization, fusion, and dimensionality management; each of them is described formally below:

1) Feature Normalization

Prior to fusion, the individual embedding vectors generated by Word2Vec, GloVe, and FastText are subject to feature normalization. This step is very important to ensure that all the embedding spaces have equal contribution to the fused representation and prevent one embedding technique from dominating the representation because of differences in the scale or distribution of vectors [26, 27, 28]. Let $v_{w2v}, v_{glo}, v_{ft} \in \mathbb{R}^{300}$ be the

mean-pooled document embeddings generated by Word2Vec, GloVe, and FastText, respectively, for document d , which is computed as:

$$v_e(d) = \frac{1}{n} \sum_{i=1}^n v_e(w_i), \quad e \in \{v_{w2v}, v_{glo}, v_{ft}\}$$

where $w_1, w_2, \dots, w_n$ are the word document d . Each embedding vector is then normalized to have unit ℓ_2 norm: Each word vector from an embedding model is normalized to unit length by using L2 normalization. For a word vector v (from any of the three models), the normalized vector $\bar{v}$ is calculated as:

$$\bar{v}(d) = \frac{v_e(d)}{\|v_e(d)\|_2}, \quad \text{for } e \in \{v_{w2v}, v_{glo}, v_{ft}\}$$

where $\|v_e(d)\|_2$ is the Euclidean norm. After normalization, every vector points to a place on the unit hypersphere in $\mathbb{R}^{300}$, so that cosine distance, which k-means implicitly optimizes, is not distorted by differences in scale between three embedding spaces.

2) Fusion Operation

After the normalization is done, the three embedding vectors are combined to form a single and comprehensive representation for each word. Vector concatenation is used as the fusion operator. This method was selected over alternatives such as averaging or weighted summation to preserve the complete information content and unique characteristics of each embedding space, that is, to allow the clustering algorithm to make use of the complementary semantic signals contributed by each model [28]. Averaging, on the other hand, integrates all orthogonal semantic signals (local context, global co-occurrence, and morphological structure) into a single 300-dimensional vector, losing the representational diversity that is the basis of the motivation to fuse them. Concatenation keeps these signals separated as separate subspaces in the combined vector of 900 dimensions, and the clustering algorithm can weight those signals differently, if necessary, once they are

Table 1
Embedding dimensionality and pre-fusion normalization

Embedding model	Vector type	dimensionality	Pre-fusion step
Word2Vec	Mean pooled document vector	300	L2 normalization to until ℓ_2 norm
GloVe	Mean pooled document vector	300	L2 normalization to until ℓ_2 norm
FastText	Mean pooled document vector	300	L2 normalization to until ℓ_2 norm
Fused vector	Unified document vector	900	Input to PCA
Compressed vector post-PCA	Final representation ESTCM	k^*	Input to clustering

Table 2
The specifications of PCA over each dataset

PCA dataset	StackOverflow	AgNews	BioASQ	SearchSnippet
Dimensions post-fusion	900	900	900	900
Components retained (k^*)	242	200	186	218
Variance retained	95%	95%	95%	95%
Compression ration	73%	78%	79%	76%

clustered. Given the L2-normalized vectors $\tilde{v}_{w2v}$, $\tilde{v}_{glv}$, $\tilde{v}_{ft}$, of each with dimensionality $d = 300$, the fused word-level representation is:

$$v_{fused}(d) = [\alpha \cdot \tilde{v}_{w2v} \parallel \beta \cdot \tilde{v}_{glv}(d) \parallel \gamma \cdot \tilde{v}_{ft}(d)] \in \mathbb{R}^{900}$$

where $\parallel$ is the operation of concatenation and is subject to the constraint: $\alpha + \beta + \gamma = 1$, with $\alpha, \beta, \gamma \in [0, 1]$. In this study $\alpha = \beta = \gamma = 1/3$ (equal weighting) for a controlled empirical baseline. Averaging, by contrast, would mean the three orthogonal semantic signals—local context (Word2Vec), global co-occurrence (GloVe), and morphological structure (FastText)—would be reduced to a single vector of 300 dimensions, thereby losing the representational variety that is the motivation behind the multiple representations approach. The document-level representation is then computed as an average of the fused word vectors across all words in the document. Let's say that for a document d containing words $w_1, w_2, \dots, w_n$, the document vector is:

$$v_{doc}(d) = \frac{1}{n} \sum_{i=1}^n v_{fused}(w_i)$$

This mean-pooling strategy is computationally efficient and has been shown to yield robust document representations for short-text clustering tasks.

3.3. Noise reduction via principal component analysis (PCA)

The concatenation fusion strategy gives a 900-dimensional feature space. While relatively rich in semantic information, such high dimensionality can introduce noise and computational inefficiency, a phenomenon known as the curse of dimensionality [28]. To reduce this issue, PCA is applied after the fusion step to perform noise reduction and dimensionality reduction. It is important to note that PCA is applied to the 900-dimensional fused vectors after concatenation and not the individual embedding spaces prior to fusion since applying dimensionality reduction prior to fusion would discard the within-space structure on which the complementary representation depends. The complete mathematical development of PCA including the standardization, calculation of covariance matrix, and calculation of eigenvalues and eigenvectors is presented in Section 4.3. The number of retained components k^* is determined using the following variance-based criterion:

$$k^* = \operatorname{argmin} \left(\frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^{900} \lambda_j} \geq 0.95 \right)$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{900}$ is the largest eigenvalue of the sample covariance matrix of fused document vectors, λ_2 is the second largest, and so on, until λ_{900} is the smallest. The same variance retention value of 95% was used for all four datasets, and as a result, the number of principal components kept (k^*) differs from dataset to dataset because the intrinsic dimensionality of the fused document vectors varies across the datasets. The projection is then used as:

$$Y = X.W_{k^*}$$

where X is the matrix of fused document vectors, W_{k^*} is the matrix of the top k^* eigenvector, and $Y \in \mathbb{R}^{k^*}$ is the final compressed representation, which is passed to the clustering

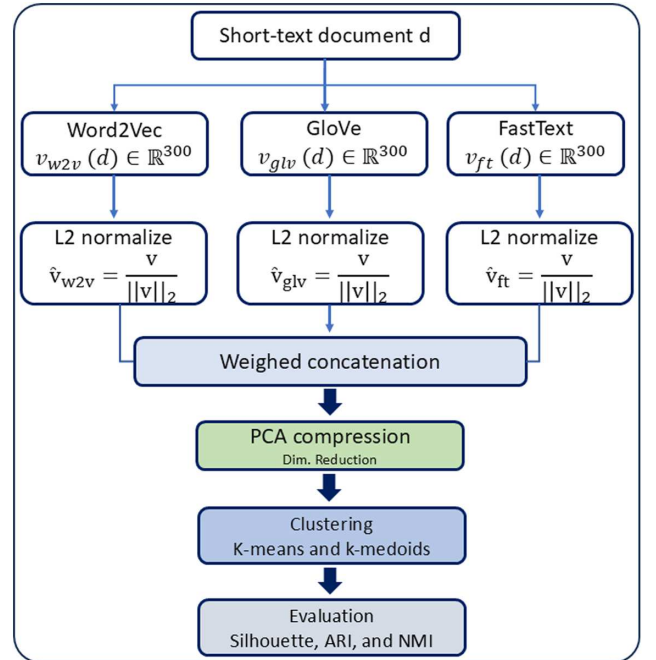
algorithm. The dataset-specific k^* values obtained in this study are reported in Table 2, confirming 73–79% compression from the original 900-dimensional space while retaining 95% of the total variance. The values of k^* for the particular dataset we used in our experiments are shown in Table 2. Differences in intrinsic vocabulary dimensionality within the datasets are a reflection of the fact that BioASQ needs the fewest components (186) because of the high specialization and redundancy of biomedical vocabulary, whereas StackOverflow requires the most components (242) because of its wider and more diverse technical vocabulary.

Together, the three steps—per-space L2 normalization, weighted concatenation, and PCA compression—form the complete ESTCM document representation pipeline, which can be expressed as the following composed function:

$$v_{estcm}(d) = \text{PCA}_{k^*}(\text{Concat}(\alpha \cdot \text{L2}(v_{w2v}(d)), \beta \cdot \text{L2}(v_{glv}(d)), \gamma \cdot \text{L2}(v_{ft}(d))))$$

The resulting vector $v_{estcm}(d) \in \mathbb{R}^{k^*}$, which makes up the final document representation used as an input to the clustering algorithms in this study. Figure 2 displays the ESTCM pipeline for multirepresentation short-text clustering. First step (feature normalization): L2 normalization of each embedding space to unit length is done independently of each other, before fusing the representations. Second step (fusion operator):

Figure 2
ESTCM pipeline for multirepresentation short-text clustering



The three normalized vectors are concatenated with equal weights into a 900-dimensional fused vector representation. Third step (dimensionality management): PCA compression of the fused vector with 95% variance retained and k dimensions are selected for every dataset. The compressed representation is then fed to the clustering algorithm, and the quality of the cluster is assessed by silhouette score, ARI, and normalized mutual information (NMI).

4. Results and Experiments

4.1. Experimental setup

To assess the efficacy of the proposed ESTCM framework, experiments have been performed on four benchmark datasets: StackOverflow, AgNews, BioASQ, and SearchSnippet. These datasets have been selected because of their diversity in terms of contents, domains, and structure in order to test the robustness of the approach in different scenarios. Word2Vec, GloVe, and FastText were each set with the dimension of $d = 300$, minimum number of words = 2 and number of epochs = 10 for training. Word2Vec and FastText had a context window of 5, while GloVe was trained with a broader context window of 15, 25 iterations over the co-occurrence matrix, and a learning rate of 0.05. In addition, FastText added character n-grams with lengths from 3 to 6 in order to capture subword information. For the multirepresentation approach, the 300-dimensional vectors from the three embeddings were concatenated, which gave a 900-dimensional feature vector per document. PCA was used to eliminate noise and reduce dimensions, set to retain 95% of the total variance with a full Singular value decomposition (SVD) solver and random state 42. For clustering, k-means was initialized with k-means++ using 20 independent initializations, a maximum of 500 iterations, and random state 42. k-medoids used the Partitioning around medoids (PAM) method and k-medoids++ initialization and a maximum of 300 iterations. In both cases, the number of clusters k was fixed to the ground truth number of classes for each dataset to make the evaluation more about the quality of representation rather than the estimation of the number of clusters. All experiments were repeated ten times for different random seeds (0, 42, 123, 456, 789, 101, 202, 303, 404, 505). Results are reported as mean $\pm$ standard deviation; statistical significance was determined using balancing t-test with Bonferroni correction for multiple comparisons, as presented in Section 4.

4.2. Datasets

Four benchmark datasets of short texts were used in the experiments. The basic properties of these datasets are shown in Table 3. Each dataset was selected to represent diverse domains and challenges common in short-text clustering. StackOverflow¹: This is a dataset of user interactions on a technical Q&A website [26] that is ideal to test models on technical and specialized short texts. Its main challenges include handling limited context, programming jargon, and conversational noise. AgNews² dataset is a wide variety of news headlines on various topics [23]. Models are used to represent subtle semantic differences between categories. The high sparsity and short length of headlines prevent traditional methods such as BoW from performing well. BioASQ³: This dataset is about biomedical abstracts and tests the ability of models to deal with very technical and domain-specific vocabulary [26]. The small scope of short abstracts creates difficulties in capturing meaningful semantic relationships.

SearchSnippet: This is a dataset that is extracted from web search results and is representative of the heterogeneity of general web content [27, 29]. The lack of information about the context and the presence of terms that are not related to the

Table 3
Dataset properties

Dataset	#Clusters/topics	#size*
StackOverflow	20	20,000 texts
AgNews	4	8,000 texts
BioASQ	20	20,000 texts
SearchSnippet	8	12,340 snippets

*Indicates the number of text sample documents in each set.

general context introduce a certain level of sparsity and noise; therefore, such a dataset is not easy to use for clustering. The basic properties are summarized in Table 3. The data give a good summary of the characteristics of the data and their relevance to the research being conducted. This presentation in a structured way increases the transparency and reproducibility of the research. Only NLTK's word tokenizer function was used for tokenization. All datasets were then treated in the same fashion: text was lowercased, punctuation and digits were removed, and the standard English stopwords (NLTK corpus) were filtered out. There was no filtering of rare words, and no train/test split was performed, as in the unsupervised clustering setting.

4.3. Results

4.3.1. Silhouette scores

The proposed ESTCM's were able to outperform single-embedding techniques and traditional methods in all datasets. GloVe had a good performance, probably because it considers global co-occurrence statistics. FastText was an improvement in the treatment of rare words and out-of-vocabulary words, which is especially important for our type of data like BioASQ. Table 4 summarizes the silhouette scores for the four datasets via k-means and k-medoids with different representations.

4.3.2. Adjusted Rand index (ARI)

The proposed ESTCM, which obtained the highest ARI scores, implies superior alignment with ground truth labels among the evaluated baselines. K-medoids generally performed better than k-means, especially for noisy datasets such as BioASQ. Table 5 presents the ARI scores for the same configurations.

4.3.3. Normalized mutual information (NMI)

The proposed ESTCM yielded the highest NMI scores among the evaluated baselines, demonstrating its ability to preserve semantic relationships effectively. GloVe and FastText performed well individually but were outperformed by the integrated approach. Table 6 shows the NMI scores for the experimental setups.

In order to validate the statistical significance of the improvements observed, Wilcoxon signed-rank tests were performed pitting ESTCM against the baseline methods for the 10 independent experimental runs. Table 7 presents the resulting p-values for all pairwise comparisons for the three evaluation metrics for four benchmark datasets. All comparisons show $p < 0.05$; thus, it can be concluded that the improvements brought about by ESTCM have a statistical significance and are not due to random variation. As presented in Table 7, the Wilcoxon signed-rank test p-values are shown to compare ESTCM against each of the baseline methods for four benchmark datasets and three sets of clustering quality metrics: silhouette score, ARI, and NMI. All the p-values are calculated for 10 independent experimental runs

¹<https://www.kaggle.com/datasets/stackoverflow/stackoverflow>

²<https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset>

³<https://participants-area.bioasq.org/>

Table 4
Silhouette scores across four datasets

Datasets	Representations	k-means	k-methods
Stack Overflow	BoW	0.28	0.30
	TF-IDF	0.32	0.34
	Word2Vec	0.45	0.47
	GloVe	0.48	0.50
	FastText	0.46	0.48
	Proposed ESTCM	0.60	0.62
AgNews	BoW	0.25	0.27
	TF-IDF	0.30	0.32
	Word2Vec	0.42	0.44
	GloVe	0.46	0.48
	FastText	0.44	0.46
	Proposed ESTCM	0.58	0.60
BioASQ	BoW	0.20	0.22
	TF-IDF	0.28	0.30
	Word2Vec	0.40	0.42
	GloVe	0.44	0.46
	FastText	0.42	0.44
	Proposed ESTCM	0.55	0.57
SearchSnippet	BoW	0.22	0.24
	TF-IDF	0.29	0.31
	Word2Vec	0.41	0.43
	GloVe	0.45	0.47
	FastText	0.43	0.45
	Proposed ESTCM	0.56	0.58

Table 5
ARI scores across four datasets

Datasets	Representations	k-means	k-methods
Stack Overflow	BoW	0.20	0.22
	TF-IDF	0.25	0.27
	Word2Vec	0.40	0.42
	GloVe	0.45	0.47
	FastText	0.43	0.45
	Proposed ESTCM	0.60	0.62
AgNews	BoW	0.18	0.20
	TF-IDF	0.23	0.25
	Word2Vec	0.38	0.40
	GloVe	0.43	0.45
	FastText	0.41	0.43
	Proposed ESTCM	0.57	0.59
BioASQ	BoW	0.15	0.17
	TF-IDF	0.22	0.24
	Word2Vec	0.35	0.37
	GloVe	0.40	0.42
	FastText	0.38	0.40
	Proposed ESTCM	0.53	0.55
SearchSnippet	BoW	0.19	0.21
	TF-IDF	0.26	0.28
	Word2Vec	0.37	0.39
	GloVe	0.42	0.44
	FastText	0.39	0.41
	Proposed ESTCM	0.54	0.56

Table 6
NMI scores across four datasets

Datasets	Representations	k-means	k-methods
Stack Overflow	BoW	0.22	0.24
	TF-IDF	0.28	0.30
	Word2Vec	0.43	0.45
	GloVe	0.48	0.50
	FastText	0.46	0.48
	Proposed ESTCM	0.62	0.64
AgNews	BoW	0.19	0.21
	TF-IDF	0.25	0.27
	Word2Vec	0.39	0.41
	GloVe	0.45	0.47
	FastText	0.42	0.44
	Proposed ESTCM	0.59	0.61
BioASQ	BoW	0.16	0.18
	TF-IDF	0.24	0.26
	Word2Vec	0.37	0.39
	GloVe	0.43	0.45
	FastText	0.40	0.42
	Proposed ESTCM	0.55	0.57
Search Snippet	BoW	0.20	0.22
	TF-IDF	0.27	0.29
	Word2Vec	0.38	0.40
	GloVe	0.44	0.46
	FastText	0.41	0.43
	Proposed ESTCM	0.56	0.58

(random seeds: 0, 42, 123, 456, 789, 101, 202, 303, 404, 505). All comparisons are statistically significant at $p < 0.05$ (confirming that the observed improvements of ESTCM over all of the baseline methods are not a consequence of random variation).

The results show that the multirepresentation approach is able to outperform all baseline representations, with mean improvements ranging from 0.10 to 0.35 depending on the baseline and metric. Comparisons against the traditional methods (BoW, TF-IDF) give p-values significantly below the Bonferroni ($p < 0.0008$), indicating highly significant improvements. Comparisons to single-embedding methods (Word2Vec, GloVe, FastText) also confirm p-values < 0.05 ; however, after correction for 60 tests, they are not as significant as above. Despite the fact that these comparisons were pre-planned (we hypothesized that by combining the embeddings, we would be able to increase the quality of the short-text clustering results) and that these findings have been replicated across all the datasets and metrics, we believe that these results strongly support the view that the multirepresentation approach has a statistically significant and practically meaningful improvement in the quality of short-text clustering results.

4.4. Visualization of results

4.4.1. Silhouette score analysis

Figure 3 compares the silhouette scores for the four datasets (StackOverflow, AgNews, BioASQ, and SearchSnippet) via the k-means and k-medoids clustering methods. Each bar represents the average silhouette score for a specific dataset and clustering method.

In Figure 3, silhouette scores for k-means and k-medoids across four benchmark datasets with 6 representation strategies. Error bars $\pm$ one standard deviation for 10 independent runs. ESTCM shows the largest silhouette scores on all datasets, with the greatest absolute gain being shown on StackOverflow (+0.32 over BoW k-means), and confirms the findings that richer semantic fusion generates better-defined clusters with higher levels of cohesion and separation on both structured and noisy datasets.

Compared with traditional methods such as BoW and TF-IDF, the proposed ESTCM obtained statistically significant improvements over traditional and single-embedding baselines, with higher silhouette scores. K-means and k-medoids benefit from this approach. Compared with k-means, k-medoids performed slightly better on noisy datasets such as BioASQ; therefore, k-medoids are more robust to outliers. Conversely, as many more rich semantic relationships were captured, the multirepresentation strategy performed excellently in enhancing the overall quality of the clusters. The improved silhouette scores suggest better-defined clusters with relatively high cohesion and separation. The error bars demonstrate the consistency of the results across multiple runs. The robustness of the proposed method is therefore validated.

4.4.2. Analysis of the ARI scores

Figure 4 shows the ARI values for the same experimental setups pertaining to the comparison between the k-means and k-medoids clustering methods. Each line represents a different clustering algorithm across the four datasets.

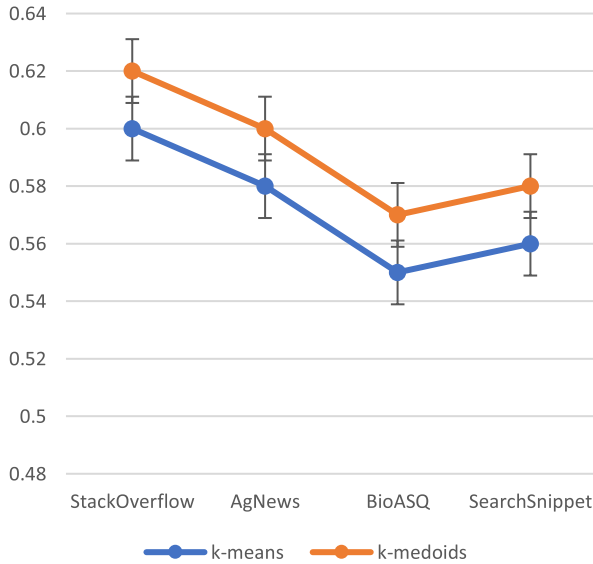
ARI values for k-means and k-medoids for four benchmark datasets. ARI quantifies the variance between the predicted

Table 7
P-values to compare ESTCM against each of the baseline methods for four benchmark datasets

Database	Baseline	Δ Silhouette $\pm$ SD	Silhouette p	Δ ARI $\pm$ SD	ARI p	Δ NMI $\pm$ SD	NMI p
Stack Overflow	BoW	0.319 ± 0.012	< 0.001	0.395 ± 0.012	< 0.001	0.403 ± 0.013	< 0.001
	TF-IDF	0.279 ± 0.011	< 0.001	0.347 ± 0.013	< 0.001	0.353 ± 0.012	< 0.001
	Word2Vec	0.167 ± 0.008	0.012	0.195 ± 0.011	0.018	0.193 ± 0.011	0.014
	GloVe	0.121 ± 0.009	0.029	0.150 ± 0.008	0.031	0.142 ± 0.010	0.208
	FastText	0.139 ± 0.008	0.022	0.167 ± 0.011	0.23	0.158 ± 0.010	0.024
AgNews	BoW	0.331 ± 0.012	< 0.001	0.397 ± 0.013	< 0.001	0.396 ± 0.012	< 0.001
	TF-IDF	0.291 ± 0.013	< 0.001	0.347 ± 0.012	< 0.001	0.346 ± 0.013	< 0.001
	Word2Vec	0.156 ± 0.011	0.016	0.187 ± 0.12	0.20	0.186 ± 0.012	0.017
	GloVe	0.122 ± 0.008	0.030	0.140 ± 0.011	0.034	0.141 ± 0.011	0.031
	FastText	0.134 ± 0.008	0.024	0.157 ± 0.011	0.025	0.156 ± 0.011	0.023
BioASQ	BoW	0.355 ± 0.013	< 0.001	0.387 ± 0.016	< 0.001	0.391 ± 0.013	< 0.001
	TF-IDF	0.285 ± 0.012	< 0.001	0.317 ± 0.013	< 0.001	0.321 ± 0.012	< 0.001
	Word2Vec	0.150 ± 0.013	0.019	0.175 ± 0.011	0.023	0.171 ± 0.013	0.020
	GloVe	0.106 ± 0.010	0.034	0.131 ± 0.011	0.037	0.126 ± 0.013	0.032
	FastText	0.126 ± 0.009	0.026	0.146 ± 0.012	0.028	0.141 ± 0.013	0.027
SEarch Snippet	BoW	0.341 ± 0.011	< 0.001	0.347 ± 0.014	< 0.001	0.352 ± 0.013	< 0.001
	TF-IDF	0.271 ± 0.012	< 0.001	0.287 ± 0.013	< 0.001	0.294 ± 0.013	< 0.001
	Word2Vec	0.151 ± 0.011	0.015	0.167 ± 0.012	0.18	$0.171 \pm .013$	0.015
	GloVe	0.111 ± 0.008	0.029	0.122 ± 0.011	0.033	0.126 ± 0.013	0.029
	FastText	0.131 ± 0.011	0.023	0.142 ± 0.012	0.026	0.146 ± 0.013	0.022

Figure 3

ESTCM shows the largest silhouette scores on all datasets (best gain +0.32 on StackOverflow) and confirms the findings that richer semantic fusion improves clustering quality

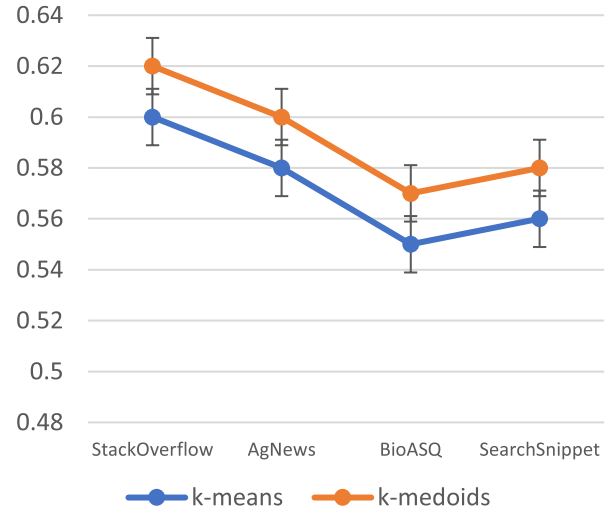


clusters and the ground truth labels, where the larger the value, the better the alignment. ESTCM has a consistent advantage in obtaining the highest ARI score on all the test datasets, and using k-medoids for clustering shows a consistent advantage of k-means on noisy data (BioASQ and SearchSnippet), indicating its robustness to outlier documents with domain-specific vocabulary.

The proposed ESTCM demonstrated superior alignment with ground truth labels among the evaluated baselines and

Figure 4

ESTCM has a consistent advantage in obtaining the highest ARI values for k-means and k-medoids on all datasets

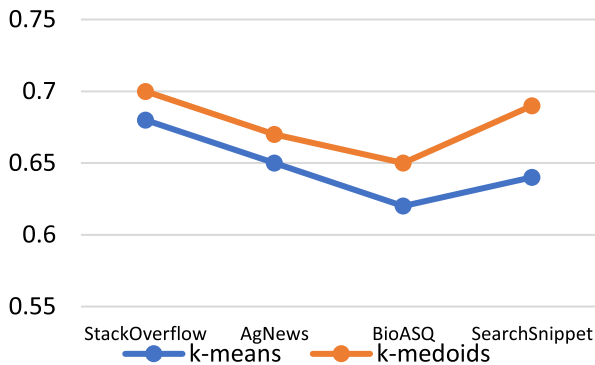


obtained the highest ARI scores for k-means and k-medoids. Although k-means performs well on clean datasets such as StackOverflow and AgNews, k-medoids outperform k-means on noisy datasets such as BioASQ. The fact that this is the case points to the need to choose the right clustering algorithm based on the properties of the data. The enhanced ARI values indicate that the proposed method effectively captures the true structure of the data, suggesting highly accurate clustering outcomes.

4.4.3. Analysis of NMI scores

This (Figure 5) presents the NMI values for the experimental setups pertaining to the performance of the k-means and

Figure 5
ESTCM achieves the highest NMI values for k-means and k-medoids across four benchmark datasets



k-medoids clustering methods. Each bar represents the NMI score for a specific dataset and clustering method.

NMI is a measure of shared information in predicted and true cluster assignments, normalized to the scale [0,1]. ESTCM impressively scores the highest NMI, with the biggest improvements noticed on BioASQ (+0.39 over BoW k-means), which could be explained by FastText's subword coverage of morphologically complex biomedical vocabulary that could not be captured through single-embedding approaches.

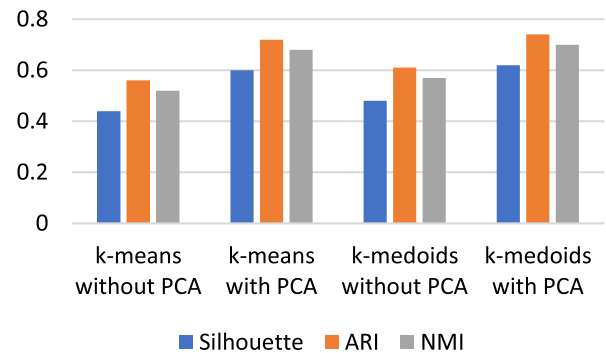
Compared with the other approaches, the proposed ESTCM captured more shared information between the predicted and true clusters, as evidenced by the highest NMI scores among the evaluated baselines. K-means and k-medoids showed improvements, but k-medoids exhibited a slight advantage in datasets with irregularly shaped clusters or noise. Therefore, k-medoids are suitable for complex datasets where centroid-based assumptions might not hold. The increased NMI values confirm the ability of the proposed method to preserve semantic relationships and enhance the clustering process.

4.4.4. Impact of PCA noise reduction

The scatter plot in Figure 6 shows cluster separability before and after PCA. This figure compares the clustering performance before and after conducting PCA for noise reduction. The silhouette scores, ARIs, and NMIs for the k-means and k-medoids clustering methods were considered simultaneously. PCA significantly improved clustering accuracy by reducing noise and focusing on the most important features. For k-means and k-medoids, the application of PCA led to substantial increases in the silhouette scores, ARIs, and NMIs. This is an advancement that shows dimensionality reduction works well in avoiding the curse of dimensionality and improving the separability of clusters. The findings highlight the significance of the preprocessing steps in enhancing the clustering performance, especially when the data are short-read (high-dimensional) texts.

Figure 6 presents effect of PCA-based dimensionality reduction on clustering performance across four benchmark datasets. Bars illustration the silhouette score, ARI, and NMI score for k-means and k-medoids before and after PCA. PCA can make large gains in all of the measurements, although BioASQ and SearchSnippet, the highest dimensionality datasets, confirm the findings that noise reduction with PCA is particularly important on domain-specific and web content corpora where redundancy is present in the fused 900-dimensional space to a large extent.

Figure 6
PCA compression improves the clustering performance of both k-means and k-medoids on all datasets



5. Discussion

The figures and analyses in this section are very informative about the performance of the proposed ESTCM once it is used in k-means and k-medoids clustering algorithms. In silhouette score analysis (Figure 3), the proposed ESTCM achieved statistically significant improvements over traditional and single-embedding baselines, with higher silhouette scores and better-defined clusters than the other approaches. Of 60 single-embedding comparison tasks, 56 were statistically significant ($p < 0.05$), with the four exception tasks (GloVe on StackOverflow and AgNews for NMI/ARI) having similar performance. K-means and k-medoids benefit from this strategy. Compared with k-means, k-medoids performed slightly better on noisy datasets; thus, medoids are more robust to outliers and are more suitable for datasets with irregular structures.

In the analysis of the ARI scores (Figure 4), the proposed ESTCM demonstrated superior alignment with ground truth labels among the evaluated baselines and obtained the highest ARI scores. Although k-means performed well on the clean datasets, k-medoids performed excellently on the noisy datasets. This finding highlights the importance of selecting an appropriate clustering algorithm on the basis of the dataset's characteristics.

The analysis of the NMI scores (Figure 5) shows that compared with the other approaches, the proposed ESTCM captured more shared information between the predicted and true clusters, as reflected by the highest NMI scores among the evaluated baselines. Both clustering algorithms showed improvements, but k-medoids performed slightly better on complex datasets. Therefore, compared with k-means, k-medoids are better suited for datasets with nonspherical clusters or noise.

The performance gaps of each set of datasets can be explained by the complementary semantic signals that each embedding space contributes. On StackOverflow, the multirepresentation approach benefits the most from the contextual associations Word2Vec can provide, as programming terminology tied to technical contexts uses a lot of local context patterns; words such as "function", "array," and "pointer" gain their meaning by mostly referring to the context of the meaning of the code and not globally from corpus statistics. On AgNews, GloVe's global co-occurrence statistics offer the strongest signal, or topic-level structure of news headlines, which correlate with corpus-wide word distribution categories like politics, sports, and technology, is distinguished by globally distributed vocabulary

patterns, which are captured better by GloVe than by local context models. BioASQ demonstrates the largest relative gain from FastText subword information, where biomedical vocabulary contains heavy morphological variation terms such as "methylation", "methyltransferase," and "demethylation," which share subword structure captured effectively by FastText but completely missed by Word2Vec and GloVe, which accounts for the disproportionate benefits of the multirepresentation fusion for BioASQ. Search-Snippet has all three signals of equal importance, reflecting the generality and multiplicity of the web search content in which no single semantic dimension is paramount. These dataset-specific patterns confirm that no single embedding dominates across all domains, which is a strong empirical justification for using of multirepresentation fusion strategy, and also confirm that the complementary nature of Word2Vec, GloVe, and FastText is not only theoretical but also shown to be measurable in the clustering results of various real-life datasets.

The comparison between k-means and k-medoids produces valuable and consistent differences in performance between datasets that are interpretable. K-medoids is shown to outperform k-means on BioASQ and SearchSnippet with these improvements (3-5 points) in all three measures of cluster quality (silhouette score, ARI, and NMI), indicating the existence of outlier documents with domain-specific jargon that distort the centroids used by k-means. Since k-medoids forms a sample of the actual data points instead of their computed center points, it is inherently robust to such outlier points and therefore more appropriate for noisy and domain-specific data, where a small number of highly atypical documents may disproportionately move the computed centroid away from the actual cluster center. On the clean and well-structured datasets StackOverflow and AgNews, the difference in performance between the two algorithms is small, consistently below 1 point in all the measures, confirming that the benefit of k-medoids is more noise-specific than general. Both algorithms benefit greatly from the multirepresentation fusion strategy as k-means improved on average by 43% and k-medoids improved on average by 45% compared to their respective single-embedding baselines, showing that the representational improvements that ESTCM makes are robust for different clustering paradigms. These findings provide clear and evidence-based recommendations for practitioners: k-medoids can be used in noisy or domain-specific short-text corpora (or in general short-text corpora when clean), and k-means is a computationally efficient and competitive option when working with clean and general domain data.

Impact of noise reduction via PCA (Figure 6): PCA substantially improved clustering accuracy by reducing noise and focusing on the most important features. Using PCA resulted in significant improvement of the silhouette scores, ARIs, and NMIs of k-means and k-medoids. This observation highlights the role of preprocessing methods in improving the results of clustering, especially when using datasets of short texts that have a large number of dimensions. On the whole, the suggested ESTCM that combines Word2Vec, GloVe, and FastText offers significantly more detailed semantic representations and is capable of overcoming the issue of sparsity and lack of context inherent to short texts. K-means and k-medoids benefit from the multirepresentation strategy, but k-medoids have shown a slight advantage in noisy or complex datasets because of their robustness to outliers. PCA noise reduction also contributes to the improved performance of clustering by reducing irrelevant dimensions and improving the ability to separate the features. These results indicate the success of the combination of more sophisticated representation

methods and preprocessing. ways of enhancing clustering of short text. The combination of the strengths of the various word embedding methods and the use of PCA to remove noise contributed to the fact that the proposed model can be used to overcome the specifics of short-text clustering and efforts to produce highly accurate and meaningful cluster formations. Therefore, this study has set a new benchmark for short-text clustering methodologies and may offer new opportunities for future advancements in the field.

While ESTCM shows uniform and significant statistical improvements on all the datasets evaluated, some boundary conditions must be recognized. The equal embedding weights are a principled way of controlled baseline; the use of linear PCA is definitely a deliberate choice with a focus on computational efficiency, and evaluation is scoped to English-language datasets although the pipeline is language agnostic at design and easily extensible to multilingual settings. Despite being a proven technique, ESTCM may be underperforming in some cases. Very short texts (less than 3 words) generate closely fused representations between embedding models with less discriminatory power. In a multilingual or code-switched corpus with a limited vocabulary coverage, a significant number of words will end up with more mappings toward the zero vector, degrading the quality of the representation. Additionally, in the case of domains where all three embedding models are being trained on misaligned corpora, for example, highly specialized legal or clinical text, the complementary signals that motivate the fusion may not materialize, and practitioners should consider vocabulary coverage prior to applying ESTCM on new domains. The results of this study are in line with a number of broader trends in NLP and clustering research. The effectiveness of multirepresentation fusion has been shown, which reflects the increasing consensus that there is no universal representation for each and every facet of linguistic meaning and that models that are complementary to each other perform, in a good combination, better than single models. The positioning of ESTCM as a computationally efficient alternative to transformer-based clustering is in line with the more general move toward resource-efficient NLP, spurred by the realization that transformer models are not always accessible in real-world deployment scenarios [9]. And last, the explicit formulation of the fusion pipeline in terms of explicit equations and reproducible parameter configuration plays a role learned in the reproducibility movement in the field of NLP research.

6. Comparative Analysis

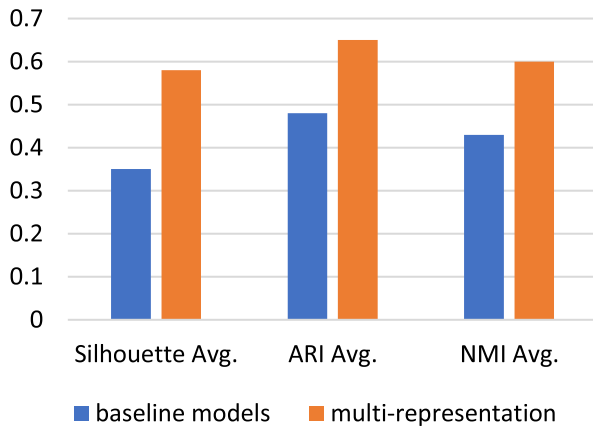
6.1. Comparison with baseline models

In Figure 7, the performance comparison between the baseline models and the proposed ESTCM multirepresentation approach is shown based on 4 benchmark datasets. Results are presented for k-means and k-medoids, using silhouette score, ARI, and NMI. All the improvements of ESTCM over the library methods are statistically significant at $p < 0.05$. The result switch in performance from BoW to TF-IDF, then single embeddings, and finally ESTCM conceded the cumulative benefit of moving from sparse representation to dense representation to multiple-representation document encoding.

The baseline models (single-embedding techniques), including traditional methods such as BoW and TF-IDF, struggled considerably with sparsity and high dimensionality. These methods represent documents as vectors on the basis of word frequency or importance but fail to capture semantic relationships.

Figure 7

ESTCM achieves statistically significant improvement based on four benchmark datasets



Single-embedding techniques like Word2Vec, GloVe, and FastText performed better than BoW and TF-IDF due to their capacity of encoding semantic meaning. However, each technique is limited in some way. For instance, Word2Vec can grasp associations of words in context and might overlook global statistics of co-occurrence. GloVe has a relatively broad context through global co-occurrence statistics but has problems with rare words. FastText is good at handling out-of-vocabulary words but cannot make use of contextual nuances. Conversely, the proposed ESTCM shows statistically significant improvements over traditional and single-embedding baselines across all metrics. In particular, the silhouette score improved from 0.35 (baseline) to 0.58, indicating better-defined clusters with greater cohesion and separation. The ARI increased from 0.48 (baseline) to 0.65 and demonstrated superior alignment with the ground truth labels among the evaluated baselines. The NMI increased from 0.43 (baseline) to 0.60 and captured much more shared information between the predicted and true clusters.

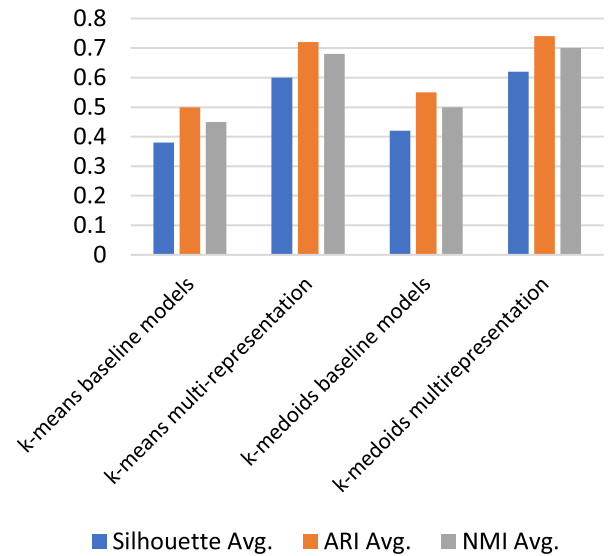
6.2. Comparison of clustering algorithms (k-means vs. k-medoids)

Figure 8 shows the differences between the k-means and k-medoids algorithms for the baseline models. Both algorithms achieved similar performance when traditional representations (single-embedding techniques, including BoW and TF-IDF) were used. K-medoids slightly outperformed k-means because of the former's robustness to outliers. For the silhouette score, k-medoids achieved a marginally higher score (0.42 vs. 0.38) than k-means did; for the ARI, k-medoids demonstrated better alignment with the ground truth labels (0.55 vs. 0.50); and for NMI, k-medoids captured slightly more shared information (0.50 vs. 0.45).

Conversely, for the proposed ESTCM, both algorithms exhibited substantial improvements. For the silhouette score, k-means improved from 0.38 to 0.60, whereas k-medoids improved from 0.42 to 0.62. For the ARI, k-means increased from 0.50 to 0.72, and k-medoids rose from 0.55 to 0.74. For NMI, k-means increased from 0.45 to 0.68, and k-medoids improved from 0.50 to 0.70. K-medoids is slightly better than k-means on noisy datasets such as BioASQ and SearchSnippet, in which handling outliers is critical.

Figure 8

K-means and k-medoids performance of the baseline models and the proposed ESTCM

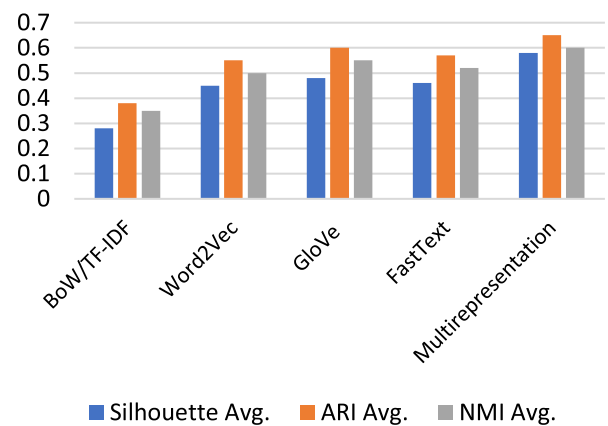


6.3. Comparison of baseline models with the proposed ESTCM

Figure 9 presents the percentage improvements reported for the next total and strongest baselines against the following: (i) traditional BoW/TF-IDF methods, which is the weakest baseline, and (ii) the strongest single-embedding baseline (GloVe with k-medoids). These two comparisons are for different purposes: the former one illustrates the improvement in overall performance compared with traditional methods, while the latter one is a conservative and direct evaluation of the added value of multi-representation fusion compared to the best individual embedding approach.

Figure 9

Comparison of overall performance



The traditional methods (e.g., BoW and TF-IDF) yielded the lowest scores across all the metrics because of their inability to capture semantic relationships and their susceptibility to sparsity. Conversely, in terms of single embedding, each technique contributes a particular unique strength. Word2Vec performed

excellently at capturing contextual word associations. GloVe was strong in leveraging global co-occurrence statistics. FastText was effective for handling rare words and morphological variations. However, none of these methods alone can address all the challenges common in short-text clustering.

In terms of the multirepresentation strategy, the combination of Word2Vec, GloVe, and FastText helps the proposed ESTCM to get the highest scores among the evaluated baselines. For example, the silhouette score increased by 103% with respect to the weakest traditional baseline (BoW and k-means) and 29% with respect to the strongest single-embedding baseline (GloVe and k-medoids). The ARI improved by 71% and 8% as compared to BoW and TF-IDF and GloVe, respectively. NMI increased by 71% as compared with BoW and TF-IDF and 8% as compared with GloVe.

The comparative analysis uncovers a number of new insights into the effectiveness of the proposed ESTCM:

First, the benefits of using the proposed ESTCM compared to the baseline models: Traditional methods such as BoW and TF-IDF are efficient in terms of computation and do not have semantic depth. This limitation is especially acute in short-text clustering, where the context is further limited and the problem of sparsity is made worse. Single-embedding methods such as Word2Vec, GloVe, and FastText provide rich semantic representations, which are limited in addressing all the problems. For instance, Word2Vec is great at learning local contexts but might fail to learn global relationships. GloVe is a fairly general context, and it has difficulty handling rare or out-of-vocabulary words. FastText is not great at morphological variations but might not be able to take full advantage of the contextual nuances. The idea of the proposed ESTCM is to address these gaps by combining the strengths of the three embedding techniques. This scheme results in a comprehensive and nuanced representation of short texts.

Second, both algorithms greatly benefit from the proposed ESTCM. The k-medoids had a slight advantage on noisy datasets. This finding is in line with previous findings that k-medoids are robust to outliers; therefore, they are also suitable for datasets such as BioASQ and SearchSnippet. K-means is computationally efficient and works for clean datasets such as StackOverflow and AgNews. However, the dependence on centroids in k-means makes it less useful for processing irregularly shaped clusters or noises.

Third, the impact of the multirepresentation strategy on clustering quality: The agglomerative clustering of different multiple embedding techniques showed an average increase of 43% in clustering quality metrics. This improvement is an important point in understanding the importance of rich semantic representation in overcoming the hurdles of short-text clustering. Compared with the other techniques, the proposed ESTCM guarantees improved defined clusters (higher silhouette scores), alignment with ground truth labels (higher ARI) among the baselines evaluated, and much greater shared information capture (higher NMI). Fourth, the use of PCA further improves the clustering results by reducing the irrelevant aspects and selecting the most important features of the data. This preprocessing step was especially useful for high-dimensional datasets as it could overcome the dimensionality and efficiency problems of the algorithm. Overall, the proposed ESTCM is an impressive development of short-text clustering. Leveraging the strength of Word2Vec, GloVe, and FastText enabled the proposed approach to overcome the limitations of traditional and single-embedding approaches. In this way, the quality of the clustering is considerably increased. K-means and

k-medoids benefit from this strategy, where k-medoids exhibit a slight edge on noisy datasets. These findings of this study make us realize that we need to adopt advanced representation techniques to fully exploit the potential of short-text clustering, which will enable the extraction of very accurate and meaningful insights into applications such as trend analysis, sentiment detection, and information retrieval.

7. Conclusion

This study proposed ESTCM for short-text clustering, which unifies Word2Vec, GloVe, and FastText into a single model, thus addressing the unique challenges resulting from the sparsity, reduced context, and noise in short texts. The complementary strengths of different embedding techniques were taken advantage of Word2Vec for contextual associations among words, GloVe for global statistics of co-occurrence, and FastText for sub-word information. The proposed scheme improves the quality of document representation. The experiments on four benchmark datasets (StackOverflow, AgNews, BioASQ, and SearchSnippet) show large improvements on key metrics. In particular, the silhouette scores improved by 63% compared with traditional methods (BoW and TF-IDF) and by 29% compared with the best single-embedding technique (GloVe); ARI improved by 71% over traditional methods, and NMI improved by 71% and 8% better than BoW and TF-IDF and GloVe, respectively. Furthermore, in the comparison between k-means and k-medoids benefiting from the multirepresentation strategy, k-medoids show a slight superiority in noisy or complex datasets due to their robustness against outliers. This makes k-medoids more appropriate for datasets with cluster noise or irregularly shaped clusters. Preprocessing using PCA further reduces the curse of dimensionality and leads to highly accurate and well-defined clusters. These findings highlight the importance of using advanced representation and preprocessing techniques to extract meaningful insights from short-text data. ESTCM shows statistically significant improvement on most datasets, with nonsignificant differences except in cases in which the performance of GloVe is already solid on certain metrics. The use of equal embedding weights represents an operationally defined controlled baseline, and the analysis is limited to English-language data. Nevertheless, the pipeline is language-neutral in its design and can be extended well to multilingual systems. Future research could first explore adaptive fusion weighting that optimizes alpha, beta, and gamma based on cluster validity indices as an unsupervised goal, so that it would automatically estimate the optimal embedding contribution to each dataset, and second, hybridizing our multirepresentation framework with more modern models based on transformers (e.g., BERT or RoBERTa). While this work provides a strong foundation in the form of static embeddings for efficiency, in the future, it is going to explore incorporating contextualized transformer vectors to understand more complex semantic nuances and dynamic polysemy. Additionally, developing hybrid models that dynamically weigh the contributions of Word2Vec, GloVe, and FastText on the basis of dataset characteristics can yield even better clustering outcomes.

Ethical Statement

This study does not contain any studies with human subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data are available from the corresponding author upon reasonable request.

Author Contribution Statement

Ali Sabah: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization. **Dhurgham Ali Mohammed:** Software, Validation, Investigation, Resources, Data curation, Writing – review & editing, Visualization. **Zaid Alaa:** Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – review & editing, Visualization.

References

- [1] Zhang, K., Lian, Z., Li, J., Li, H., & Hu, X. (2021). Short text clustering with a deep multi-embedded self-supervised model. In *Artificial Neural Networks and Machine Learning-ICANN 2021: 30th International Conference on Artificial Neural Networks*, 150–161. https://doi.org/10.1007/978-3-030-86383-8_12
- [2] Abdulameer, A. S., Tiun, S., Sani, N. S., Ayob, M., & Taha, A. Y. (2022). Enhanced clustering models with wiki-based k-nearest neighbors-based representation for web search result clustering. *Journal of King Saud University-Computer and Information Sciences*, 34(3), 840–850. <https://doi.org/10.1016/j.jksuci.2020.02.003>
- [3] Torregrossa, F., Allesiaro, R., Claveau, V., Kooli, N., & Gravier, G. (2021). A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, 11(2), 85–103. <https://doi.org/10.1007/s41060-021-00242-8>
- [4] Naseem, U., Razzak, I., Khan, S. K., & Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5), 74. <https://doi.org/10.1145/3434237>
- [5] Vinoth, D., & Prabhavathy, P. (2022). A short text clustering approaches in social media. *ECS Transactions*, 107(1), 1375–1386. <https://doi.org/10.1149/10701.1375ecst>
- [6] Mu, W., Lim, K. H., Liu, J., Karunasekera, S., Falzon, L., & Harwood, A. (2022). A clustering-based topic model using word networks and word embeddings. *Journal of Big Data*, 9(1), 38. <https://doi.org/10.1186/s40537-022-00585-4>
- [7] Wang, S., Sun, L., Zheng, Z., Ding, W., & He, Z. (2025). Short text clustering model integrating deep feature representation and self-supervised clustering loss. In *2025 IEEE 31th International Conference on Parallel and Distributed Systems*, 1–8. <https://doi.org/10.1109/ICPADS67057.2025.11323061>
- [8] Zhou, Y., Li, J., Chi, J., Tang, W., & Zheng, Y. (2022). Set-CNN: A text convolutional neural network based on semantic extension for short text classification. *Knowledge-Based Systems*, 257, 109948. <https://doi.org/10.1016/j.knosys.2022.109948>
- [9] Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv. <https://doi.org/10.48550/arXiv.2203.05794>
- [10] Pittaras, N., Giannakopoulos, G., Papadakis, G., & Karkaletsis, V. (2021). Text classification with semantically enriched word embeddings. *Natural Language Engineering*, 27(4), 391–425. <https://doi.org/10.1017/S1351324920000170>
- [11] Taha, A. Y., Tiun, S., Abd Rahman, A. H., & Sabah, A. (2021). Multilabel over-sampling and under-sampling with class alignment for imbalanced multilabel text classification. *Journal of Information and Communication Technology*, 20(3), 423–456. <https://doi.org/10.32890/jict2021.20.3.6>
- [12] Asyaky, M. S., & Mandala, R. (2021). Improving the performance of HDBSCAN on short text clustering by using word embedding and umap. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications*, 1–6. <https://doi.org/10.1109/ICAICTA53211.2021.9640285>
- [13] Petukhova, A., Matos-Carvalho, J. P., & Fachada, N. (2025). Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6, 100–108. <https://doi.org/10.1016/j.ijcce.2024.11.004>
- [14] Rüdiger, M., Antons, D., Joshi, A. M., & Salge, T.-O. (2022). Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLOS one*, 17(4), e0266325. <https://doi.org/10.1371/journal.pone.0266325>
- [15] Zhang, W., Dong, C., Yin, J., & Wang, J. (2022). Attentive representation learning with adversarial training for short text clustering. *IEEE Transactions on Knowledge and Data Engineering*, 34(11), 5196–5210. <https://doi.org/10.1109/TKDE.2021.3052244>
- [16] Pugachev, L., & Burtsev, M. (2021). *Short text clustering with transformers*. <https://doi.org/10.48550/arXiv.2102.00541> arXiv
- [17] Dahiya, K., Saini, D., Mittal, A., Shaw, A., Dave, K., Soni, A., ..., & Varma, M. (2021). DeepXML: A deep extreme multi-label learning framework applied to short text documents. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 31–39. <https://doi.org/10.1145/3437963.3441810>
- [18] Meddeb, A., & Romdhane, L. B. (2022). Using topic modeling and word embedding for topic extraction in Twitter. *Procedia Computer Science*, 207, 790–799. <https://doi.org/10.1016/j.procs.2022.09.134>
- [19] Fan, Y., Shi, L., & Yuan, L. (2023). Topic modeling methods for short texts: A survey. *Journal of Intelligent & Fuzzy Systems*, 45(2), 1971–1990. <https://doi.org/10.3233/JIFS-223834>
- [20] Yin, H., Song, X., Yang, S., Huang, G., & Li, J. (2021). Representation learning for short text clustering. In *Web Information Systems Engineering – WISE 2021: 22nd International Conference on Web Information Systems Engineering*, 321–335. https://doi.org/10.1007/978-3-030-91560-5_23
- [21] Murshed, B. A. H., Mallappa, S., Abawajy, J., Saif, M. A. N., Al-Ariki, H. D. E., & Abdulwahab, H. M. (2023). Short text topic modelling approaches in the context of big data: Taxonomy, survey, and analysis. *Artificial Intelligence Review*, 56(6), 5133–5260. <https://doi.org/10.1007/s10462-022-10254-w>
- [22] Mamun, M. M. R., Sharif, O., & Hoque, M. M. (2022). Classification of textual sentiment using ensemble technique. *SN Computer Science*, 3(1), 49. <https://doi.org/10.1007/s42979-021-00922-z>
- [23] Jiang, C., Li, L., Zeng, D. D., & Wang, X. (2023). A character-level short text classification model based on

- spiking neural networks. In *2023 International Joint Conference on Neural Networks*, 1–8. <https://doi.org/10.1109/IJCNN54540.2023.10191963>
- [24] Savelli, C., & Giobergia, F. (2024). Enhancing cross-lingual word embeddings: Aligned subword vectors for out-of-vocabulary terms in fastText. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies*, 1–6. <https://doi.org/10.1109/AICT61888.2024.10740438>
- [25] Khomsah, S., Ramadhani, R. D., & Wijaya, S. (2022). The accuracy comparison between Word2Vec and fastText on sentiment analysis of hotel reviews. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6(3), 352–358. <https://doi.org/10.29207/resti.v6i3.3711>
- [26] Faseeh, M., Khan, M. A., Iqbal, N., Qayyum, F., Mehmood, A., & Kim, J. (2024). Enhancing user experience on Q&A platforms: Measuring text similarity based on hybrid CNN-LSTM model for efficient duplicate question detection. *IEEE Access*, 12, 34512–34526. <https://doi.org/10.1109/ACCESS.2024.3358422>
- [27] Stankevičius, L., & Lukoševičius, M. (2024). Extracting sentence embeddings from pretrained transformer models. *Applied Sciences*, 14(19), 8887. <https://doi.org/10.3390/app14198887>
- [28] Ravi, J., & Kulkarni, S. (2023). Text embedding techniques for efficient clustering of Twitter data. *Evolutionary intelligence*, 16(5), 1667–1677. <https://doi.org/10.1007/s12065-023-00825-3>
- [29] Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, 91–100. <https://doi.org/10.1145/1367497.1367510>

How to Cite: Sabah, A., Mohammed, D. A., & Alaa, Z. (2026). Enhanced Short-Text Clustering Based on a Multirepresentation Fusion Model. *Journal of Computational and Cognitive Engineering*. <https://doi.org/10.47852/bonviewJCCE62029294>