IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 11, 2019, accepted October 3, 2019, date of publication October 7, 2019, date of current version October 17, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2945911*

# Two-Stage Topic Extraction Model for Bibliometric Data Analysis Based on Word Embeddings and Clustering

**AYTUĞ ONAN**
Computer Engineering Department, İzmir Katip Çelebi University, 35620 İzmir, Turkey
e-mail: aytug.onan@ikc.edu.tr

**ABSTRACT** Topic extraction is an essential task in bibliometric data analysis, data mining and knowledge discovery, which seeks to identify significant topics from text collections. The conventional topic extraction schemes require human intervention and involve also comprehensive pre-processing tasks to represent text collections in an appropriate way. In this paper, we present a two-stage framework for topic extraction from scientific literature. The presented scheme employs a two-staged procedure, where word embedding schemes have been utilized in conjunction with cluster analysis. To extract significant topics from text collections, we propose an improved word embedding scheme, which incorporates word vectors obtained by word2vec, POS2vec, word-position2vec and LDA2vec schemes. In the clustering phase, an improved clustering ensemble framework, which incorporates conventional clustering methods (i.e., k-means, k-modes, k-means++, self-organizing maps and DIANA algorithm) by means of the iterative voting consensus, has been presented. In the empirical analysis, we analyze a corpus containing 160,424 abstracts of articles from various disciplines, including agricultural engineering, economics, engineering and computer science. In the experimental analysis, performance of the proposed scheme has been compared to conventional baseline clustering methods (such as, k-means, k-modes, and k-means++), LDA-based topic modelling and conventional word embedding schemes. The empirical analysis reveals that ensemble word embedding scheme yields better predictive performance compared to the baseline word vectors for topic extraction. Ensemble clustering framework outperforms the baseline clustering methods. The results obtained by the proposed framework show an improvement in Jaccard coefficient, Folkes & Mallows measure and F1 score.

**INDEX TERMS** Topic extraction, machine learning, cluster analysis, text mining.

## I. INTRODUCTION

Topic extraction is an essential task in bibliometric data analysis, data mining and information retrieval, which aims to identify significant topics from text collections. Topic extraction from scientific literature can be especially essential for exploratory data analysis to get a quick overview of the contents of a collection and to find information objects [1].

Text mining is the computational field of study to explore and analyze the immense quantity of unstructured text documents with the use of tools and techniques from machine learning, data mining and statistics. The application of text mining on bibliometric data analysis is a promising research direction. Topic modelling and graph-based schemes have been successfully employed for topic extraction [2], [3]. The conventional topic extraction schemes require comprehensive pre-processing tasks to represent text collections in an appropriate way. To employ text mining methods on text collections, conventional natural language processing based preprocessing tasks, such as, the identification of synonymous terms, the identification of compound terms, term transformation based on stemming and lemmatization, stop-words and common terms removal must be employed [4]. In addition, conventional topic extraction schemes involve human intervention. The tasks to be needed for topic extraction include the retrieval of links among citations and co-citations and synthesizing technical synonyms [5].

Deep learning is a recent research direction in machine learning in which classification models with high predictive performance have been obtained by multiple layers of

The associate editor coordinating the review of this manuscript and approving it for publication was Michael Lyu.

nonlinear information processing and supervised or unsupervised learning of feature representations in a hierarchical way [6]. Compared to the lower levels of the architecture, higher levels of the hierarchy have more distributed and compact representations toward the data. Deep learning has been successfully employed in a wide range of application fields, such as, computer vision, speech recognition and natural language processing.

Word embedding schemes are applications of deep learning in natural language processing, in which words or phrases are represented in low-dimensional vectors of a continuous space [7]. Word embedding schemes, such as, word2vec and global vectors (GloVe) have been successfully employed for natural language processing tasks, such as, text classification, sentiment analysis and sarcasm identification [8]–[10]. With the use of word embedding schemes, syntactic and semantic relations among the words/phrases can be extracted and represented. Word embedding schemes enable to capture latent features from text collections by layers of the deep neural network architecture. Hence, the stages involved for feature extraction and data preprocessing in the conventional text mining schemes have been eliminated [11].

As stated, word embedding based schemes have been successfully employed for tasks of natural language processing. However, word embedding schemes, such as word2vec and GloVe require very large corpus for training and constructing the vector representation [7]. For this reason, pre-trained word embedding models may be employed on deep learning-based natural language processing tasks. The predictive performance of such schemes may not be promising, since the pre-trained model is not specific to the corresponding task/data [12]. The calculations for vector may not take the context of documents into account and the relationships between words that are not literally co-occurring have not been modelled [7]. To eliminate problems of word embedding schemes, several research contributions incorporated different word embedding schemes into an ensemble vector representation [7], [8].

Clustering (also known as, cluster analysis) is the unsupervised machine learning technique which aims to assign instances into groups based on their similarities. In clustering, instances are deployed into clusters such that the similar instances are grouped into the same cluster and dissimilar instances are deployed out of the cluster [13]. The conventional clustering algorithms may be greatly affected by the characteristics of the dataset and the parameters. The clustering algorithms may suffer from instability. Cluster ensembles (also known as, consensus clustering) is the process of combining results obtained by several base clustering algorithms into a single consolidated clustering to obtain a more robust and stable grouping of instances [14]. Cluster ensembles can improve the quality of final consolidated clustering. In addition, the problems associated by base clustering algorithms can be eliminated and the model's tolerance to noise and outliers can be enhanced [15].

In this research, we present a two-stage framework for topic extraction from scientific literature. The presented scheme employs a two-staged procedure, where word embedding schemes have been utilized in conjunction with cluster analysis. To extract significant topics from text collections, we propose an improved word embedding scheme, which incorporates word vectors obtained by word2vec, POS2vec, word-position2vec and LDA2vec schemes. In the clustering phase, an improved clustering ensemble framework, which incorporates conventional clustering methods (i.e., k-means, k-modes, k-means++, self-organizing maps and DIANA algorithm) by means of the iterative voting consensus, has been presented. In the empirical analysis, we analyze a corpus containing 160,424 abstracts of articles from various disciplines, including economics, engineering and computer science. In the experimental analysis, performance of the proposed scheme has been compared to conventional baseline clustering methods (such as, k-means, k-modes, k-means), LDA-based topic modelling and conventional word embedding schemes. The empirical analysis reveals that ensemble word embedding scheme yields better predictive performance compared to the baseline word vectors for topic extraction. Ensemble clustering framework outperforms the baseline clustering methods. The results obtained by the proposed framework show an improvement in Jaccard coefficient, Folkes & Mallows measure and F1 score.

The main contributions of our two-stage topic extraction framework can be summarized as follows: A novel hybrid scheme based on improved word embeddings and cluster ensemble has been presented. To the best of our knowledge, this is the first study in topic extraction, which extensively analyze the performance of word embedding schemes and present an improved word embedding scheme which incorporates several word vectors. In addition, this is the first study to use cluster ensembles on topic extraction to enhance the performance of base clustering algorithms.

The rest of this paper is structured as follows. We review the related work in Section 2. Section 3 presents the methodology and the proposed framework. We present the empirical procedure and experimental results in Section 4. Finally, we present the concluding remarks of the research in Section 5.

## II. RELATED WORKS

This section briefly presents the earlier research contributions on topic extraction and word embedding based text representation.

### A. TOPIC EXTRACTION SCHEMES

Boyack *et al.* [16] examined the predictive performance of five topic extraction schemes for biomedical text documents from MEDLINE. In the empirical analysis, cosine similarity using term frequency-inverse document frequency vectors, latent semantic analysis (LSA), topic modelling, and two Poisson-based language models have been evaluated.

In another study, Lu *et al.* [17] presented a classification based recursive soft clustering algorithm for topic extraction. In the presented scheme, text documents have been assigned into topic and topics have been assigned to the clusters. Text documents were allocated to the highest relative clusters and linked with the second most closely related cluster. Ding and Chen [18] examined the predictive performance of latent Dirichlet allocation, hierarchical Dirichlet process, co-word analysis and co-citation analysis for topic extraction and tracking. In another study, Zhang *et al.* [5] presented a term clumping framework to effectively represent text collections by reducing the scale of term-based feature space. Suominen and Toivanen [19] employed the latent Dirichlet allocation model to map scientific text documents to topics. In another study, Zhang *et al.* [20] introduced a hybrid framework for topic analysis and forecasting science, technology and innovation topics. In the presented scheme, an improved k-means clustering based approach has been employed, in which the appropriate value for $k$ parameter has been automatically adjusted. A similarity measure-based function has been employed for topic relationship identification.

The conventional text mining applications generally involve the representation of text documents in terms of bag-of-words representation scheme. However, this scheme suffers from high dimensional feature space and sparsity. In response, Yang *et al.* [21] utilized a text representation scheme based on key terms for text classification. In the presented scheme, six feature selection methods (namely, improved Gini index, information gain, mutual information, odds ratio, ambiguity measure and association factor) have been utilized to extract key terms from the text collections. The empirical analysis with support vector machines and k-nearest neighbor indicated that key terms-based representation enhances the predictive performance for text classification benchmarks. Similarly, Onan *et al.* [22] evaluated the predictive performance of five statistical keyword extraction approaches (most frequent measure-based keyword extraction, term frequency-inverse sentence frequency-based keyword extraction, eccentricity-based keyword extraction and TextRank algorithm) for text classification tasks. The empirical analysis with conventional supervised learning and ensemble learning based classification schemes indicated that keyword-based representation approaches can yield promising results on text classification tasks.

In another study, Yang *et al.* [23] introduced a hierarchical attention network-based framework for text classification. The presented framework consists of two-levels, namely, word-level and sentence-level. The word-level is responsible for extracting significant words and sentence-level is responsible for identifying and relating significant sentences. In another study, Velden *et al.* [24] compared topic extraction from bibliographic data analysis of scientific publications. Recently, Zhang *et al.* [11] presented a deep learning-based framework for topic extraction, which integrates word2vec word embedding scheme and polynomial function-based kernel k-means clustering.

### B. WORD EMBEDDING BASED REPRESENTATION

Maas *et al.* [25] employed word embedding schemes to extract semantic and sentiment meanings of words. In this scheme, an unsupervised probabilistic approach has been utilized to construct vectors. Words cooccurring in the text collections have similar representations and a supervised learning method has been employed for sentiment analysis. In another study, Socher *et al.* [26] introduced an expressive neural tensor network to capture the relationship between entities in a knowledge base. In this scheme, the vector representations of words have been utilized to compute the average word vectors. The presented scheme has been employed for the knowledge base completion task in query expansion, question answering and information retrieval tasks.

Le and Mikolov [27] presented a new representation scheme for paragraphs, sentences and text documents. In this scheme, two models have been introduced based on the vector representations of words. The first model utilizes word vectors to contribute prediction task about the next word in the sentence. In the second model, words sampled from the output has been predicted. In this way, the vector representations for paragraphs, sentences, and documents and the semantic vectors for words have been captured. The empirical analysis indicated that the concatenation of the vectors outperforms for text classification and sentiment analysis tasks. In another study, Tang *et al.* [28] introduced a sentiment specific word embedding framework to capture sentiment information in the continuous representation of words. In this scheme, sentiment information has been encoded into continuous representation of words by enhancing the conventional word embedding scheme by three neural network architectures.

Hong and Zhao [29] utilized word2vec word embedding scheme and the latent Dirichlet allocation to identify anomaly sentences in legal text documents. In another study, Zhang and Wallace [30] employed two pretrained word embedding schemes (namely, GloVE and word2vec vectors) for sentiment analysis. In another study, Wei *et al.* [31] introduced a semi-supervised autoencoder to capture meaningful latent representation of documents. Kamkarhaghighi and Makrehchi [7] presented an enhanced content tree-based word embedding scheme to eliminate word ambiguity and to inject a local context into pretrained word embeddings. Similarly, Laueren *et al.* [32] employed skip-gram and paragraph vectors distributed bag of words schemes to obtain discriminant document embeddings. Recently, Rezaeinia *et al.* [8] presented an ensemble word embedding scheme for sentiment analysis, which incorporates word2vec, pos2vec, lexicon2vec and word-position2vec word embedding schemes.

In another study, Butnaru and Ionescu [33] introduced a clustering based representation scheme for text classification tasks. In this scheme, word embedding schemes have been utilized to embed words into vector space. Then, k-means clustering algorithm has been employed on the vectors to obtain a set of clusters. The empirical results on sentiment analysis and text categorization tasks yield promising results. Similarly, Butnaru and Ionescu [34] presented another representation approach for text classification based on word embeddings and clustering. In this scheme, text collections have been represented by aggregating word embeddings using k-means algorithm.

### C. MOTIVATION AND CONTRIBUTIONS

As mentioned in advance, topic extraction for bibliometric data analysis is a promising research direction. To effectively extract topics from bibliometric text collections, several methods and algorithms have been proposed based on topic modelling [16], [18], [19], cluster analysis [17], [20] and graph models [3]. These models, however, involve human intervention and data preprocessing (such as, the retrieval of links among citations and co-citations, synthesizing technical synonyms, and cleaning words and terms). Co-word based topic extraction schemes, particularly in emerging sectors, cannot deliver promising outcomes in synthesizing technical synonyms [35]. In addition, researchers citing references can be biased towards the sources [36]. These issues degrade the reliability of topic extraction frameworks based on human intervention [11].

Word embedding schemes can be employed to extract syntactic and semantic relations among the words and phrases. With the use of word embedding-based representation, latent features can be captured from text collections and feature extraction and data preprocessing tasks of conventional topic extraction schemes can be eliminated. Recent studies on the use of conventional word embedding schemes yield promising results for topic extraction [25]–[28]. In addition, the ensemble word embedding schemes, which incorporate several word vectors obtained by word2vec, POS2vec, word-position2vec, yield higher predictive performance for several natural language processing tasks, such as sentiment analysis [7], [8], [30].

Though the use of topic modelling schemes, clustering methods and word embedding schemes take great research attention in the literature, the number of works that utilizes clustering methods in conjunction with word embedding schemes is very limited. To fill this gap, this paper presents an empirical analysis and benchmark results for conventional clustering algorithms (i.e., k-means, k-modes, k-means++, self-organizing maps), LDA-based topic modelling and conventional word embedding schemes (i.e., word2vec, POS2vec, word-position2vec, LDA2vec). Motivated by the predictive performance of cluster ensembles for cluster analysis and the predictive performance enhancement obtained by improved word embedding schemes for several tasks of natural language processing, this paper presents a hybrid topic extraction framework based on improved word embeddings and cluster ensemble.

There are several recent contributions dedicated to the use of word embedding schemes in conjunction with clustering to obtain an efficient representation scheme for text classification tasks [33], [34]. Our proposal differs from the earlier schemes in that we utilize an ensemble word embedding scheme obtained by word2vec, POS2vec, word-position2vec and LDA2vec. The clustering algorithms may suffer from instability. In this regard, ensemble word embedding based representation has been integrated to cluster ensemble which contains k-means, k-modes, k-means++, self-organizing maps and DIANA algorithm by iterative voting consensus function.

## III. METHODOLOGY

This section presents the corpus utilized in the empirical analysis, word embedding schemes and the presented improved word embedding scheme, clustering algorithms and the presented clustering ensemble framework and the topic extraction framework.

### A. CORPUS

To collect a text corpus containing abstracts of scientific journal papers, we have crawled the Web of Science (WoS) database. In the WoS core collection, research areas are classified into five main categories, as arts and humanities, life sciences and biomedicine, physical sciences, social sciences and technology. In addition, each journal covered by WoS core collection is assigned to at least one WoS subject category. Any scientific journal papers published in a journal has been assigned to the corresponding subject category [37]. We have selected 10 WoS subject categories and collected articles published in journals on 17 April 2019. In this way, abstracts for 160,424 articles from 10 different subject categories have been obtained. The basic descriptive information regarding the corpus is presented in Table 1.

In the construction of corpus, we have intentionally selected diverse subject categories which are not substantially overlap with each other. In the corpus, some subject categories (such as, materials science multidisciplinary, psychology multidisciplinary and biochemistry & molecular biology) have higher number of articles compared to the subject categories (such as, computer science theory & methods, endocrinology & metabolism, and mathematics applied). The imbalanced distribution of articles to subject categories on the corpus can be regarded as an acceptable scenario for real-life case. To make empirical analysis, we have adopted the preprocessing scheme utilized in [11]. Hence, we have constructed row-list containing titles and abstracts of articles without applying any further preprocessing tasks.

### B. IMPROVED WORD EMBEDDING SCHEME

In the proposed improved word embedding scheme, we have incorporated word vectors obtained by word2vec, POS2vec, word-position2vec and LDA2vec schemes to improve the predictive performance for topic extraction.

**TABLE 1.** Descriptive information for the corpus.

| Web of Science Subject Category | Number of Articles | Letters per word | Words per sentence | Max sentence length (words) | Min sentence length (words) |
|---|---|---|---|---|---|
| Computer Science, Theory & Methods | 3,228 | 5,76 | 20.36 | 46 | 1 |
| Radiology, Nuclear Medicine & Medical Imaging | 4,297 | 5,74 | 20.17 | 38 | 1 |
| Agricultural Engineering | 3,842 | 5,05 | 19.80 | 58 | 1 |
| Endocrinology & Metabolism | 1,349 | 5,86 | 19.40 | 43 | 12 |
| Materials Science, Multidisciplinary | 30,536 | 6,19 | 25.40 | 46 | 17 |
| Biochemistry & Molecular Biology | 26,041 | 6,03 | 20.67 | 28 | 16 |
| Psychology Multidisciplinary | 35,088 | 6,8 | 24.63 | 51 | 7 |
| Mathematics Applied | 3,92 | 5,43 | 29.00 | 62 | 12 |
| Mechanics | 23,785 | 5,46 | 16.29 | 27 | 1 |
| Economics | 28,338 | 5,82 | 22.27 | 35 | 14 |

In the improved word embedding based scheme, a sentence of scientific abstracts has been taken as an input to the model. Then, each part-of-speech tag has been assigned to a constant vector. After that, word2vec word embedding scheme has been employed on the input sentence to extract word vector of each input sequence. In case of non-existing words, the vector of the word has been randomly initialized. To obtain the sentence vectors, we have employed average of word2vec vectors scheme. In this scheme, the average of the word vectors for all the words in a sentence has been computed. In this way, the average vector represents the sentence vector. To get sentence vectors, we have evaluated two approaches: removing stop-words and keeping stop-words. Since removing stop-words yields better results, this scheme has been adopted in the experimental analysis. After generating word2vec vectors, part-of-speech tag of each word in the sentence has been identified. Based on part-of-speech tags, the corresponding vectors has been assigned to each part-of-speech tag. Then, word-position2vec vectors have been generated for each word in the sentence and word-position2vec vector has been concatenated to vectors obtained in the earlier stages. Finally, LDA2vec word embedding scheme has been employed on the sentence and the vector obtained has been concatenated to obtain final ensemble word embedding scheme. The general structure for the improved word embedding scheme has been outlined in Figure 1.

The baseline word embedding schemes have been briefly discussed:

### 1) WORD2VEC

The word2vec is a commonly utilized neural network-based approach to learn word embeddings from text documents. It is an unsupervised and efficient method to extract semantic relationships among the words based on their cooccurrence in text documents in a corpus [38].

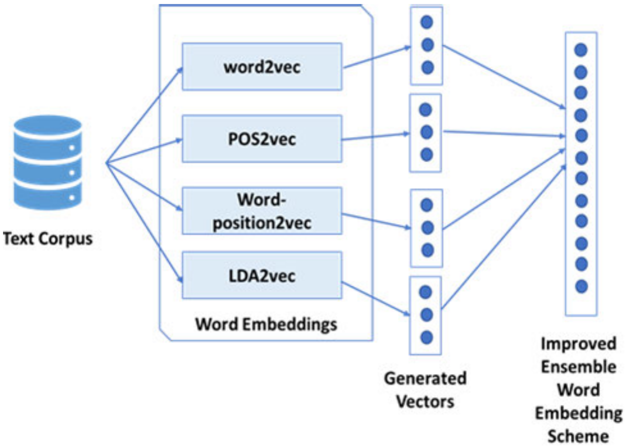

**FIGURE 1.** The general architecture for improved word embedding (IWE) scheme.

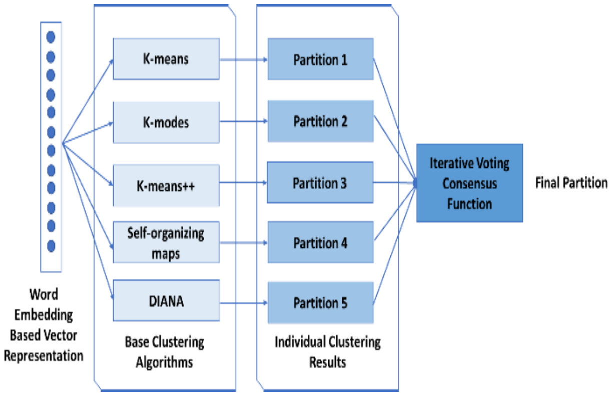
K-means
K-modes
K-means++
Self-organizing maps
DIANA
Partition 1
Partition 2
Partition 3
Partition 4
Partition 5
Iterative Voting Consensus Function
Final Partition
Word Embedding Based Vector Representation
Base Clustering Algorithms
Individual Clustering Results

**FIGURE 2.** The general structure for clustering ensemble framework.

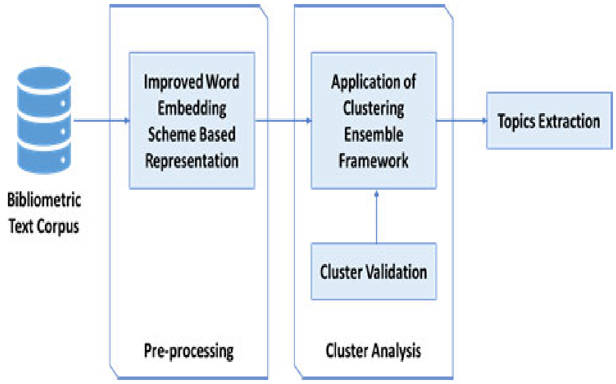
Bibliometric Text Corpus
Improved Word Embedding Scheme Based Representation
Application of Clustering Ensemble Framework
Topics Extraction
Cluster Validation
Pre-processing
Cluster Analysis

**FIGURE 3.** The general structure for topic extraction framework.

### 2) POS2VEC

Part-of-speech tagging (POS tagging) is an essential stage in computational linguistics tasks, in which each word of a text document has been assigned to an appropriate part

**TABLE 2.** Experimental results for compared models.

| Model | Jaccard Coefficient (JC) | FM Measure | F1-score |
|---|---|---|---|
| KM | 0.2650 | 0.4369 | 0.4231 |
| KMOD | 0.2572 | 0.4241 | 0.3961 |
| KM++ | 0.2678 | 0.4598 | 0.4449 |
| SOM | 0.2712 | 0.4697 | 0.4489 |
| DIANA | 0.2744 | 0.4738 | 0.4561 |
| LDA | 0.2774 | 0.4792 | 0.4646 |
| Clustering Ensemble | 0.3028 | 0.5114 | 0.5169 |
| KM + word2vec (SG) | 0.2814 | 0.4855 | 0.4711 |
| KMOD + word2vec (SG) | 0.2794 | 0.4828 | 0.4653 |
| KM++ + word2vec (SG) | 0.2819 | 0.4899 | 0.4731 |
| SOM + word2vec (SG) | 0.2837 | 0.4952 | 0.4780 |
| DIANA + word2vec (SG) | 0.2859 | 0.4964 | 0.4827 |
| Clustering Ensemble + word2vec (SG) | 0.3163 | 0.5283 | 0.5349 |
| KM + word2vec (CBOW) | 0.3026 | 0.5103 | 0.5123 |
| KMOD + word2vec (CBOW) | 0.3007 | 0.5085 | 0.5115 |
| KM++ + word2vec (CBOW) | 0.3054 | 0.5136 | 0.5225 |
| SOM + word2vec (CBOW) | 0.3069 | 0.5174 | 0.5286 |
| DIANA + word2vec (CBOW) | 0.3090 | 0.5187 | 0.5302 |
| Clustering Ensemble + word2vec (CBOW) | 0.3204 | 0.5318 | 0.5381 |
| KM + POS2vec | 0.2909 | 0.5037 | 0.4919 |
| KMOD + POS2vec | 0.2885 | 0.5030 | 0.4897 |
| KM++ + POS2vec | 0.2944 | 0.5052 | 0.4965 |
| SOM + POS2vec | 0.2959 | 0.5063 | 0.4992 |
| DIANA + POS2vec | 0.2986 | 0.5074 | 0.5067 |
| Clustering Ensemble + POS2vec | 0.3184 | 0.5296 | 0.5369 |
| KM + word-position2vec | 0.3129 | 0.5234 | 0.5338 |
| KMOD + word-position2vec | 0.3109 | 0.5221 | 0.5331 |
| KM++ + word-position2vec | 0.3222 | 0.5324 | 0.5391 |
| SOM + word-position2vec | 0.3236 | 0.5362 | 0.5433 |
| DIANA + word-position2vec | 0.3255 | 0.5383 | 0.5454 |
| Clustering Ensemble + word-position2vec | 0.3307 | 0.5464 | 0.5519 |
| KM + LDA2vec | 0.3438 | 0.5560 | 0.5749 |
| KMOD + LDA2vec | 0.3471 | 0.5580 | 0.5815 |
| KM++ + LDA2vec | 0.3546 | 0.5662 | 0.5883 |
| SOM + LDA2vec | 0.3605 | 0.5765 | 0.5958 |
| DIANA + LDA2vec | 0.3695 | 0.5835 | 0.6040 |
| Clustering Ensemble + LDA2vec | 0.3788 | 0.5955 | 0.6095 |
| KM + IWE | 0.5289 | 0.6398 | 0.6499 |
| KMOD + IWE | 0.5291 | 0.6411 | 0.6601 |
| KM++ + IWE | 0.5551 | 0.6490 | 0.6670 |
| SOM + IWE | 0.5572 | 0.6606 | 0.6609 |
| DIANA + IWE | 0.5506 | 0.6630 | 0.6661 |
| Proposed Topic Extraction Framework | **0.5951** | **0.6849** | **0.6930** |

KM: K-means, KMOD: K-modes, KM++: K-means++, SOM: Self-organizing maps, word2vec (SG): word2vec skip-gram model, word2vec (CBOW): word2vec continuous bag-of-word model, IWE: improved word embedding scheme

of speech, such as noun, verb, adjective, adverb, pronoun, preposition, conjunction, etc. To generate vectors based on POS tagging, we have first utilized NLTK Pos-tagging [39] to identify corresponding part-of-speech tags. Then, each POS tag has been converted to a constant vector and concatenated with word2vec word embedding scheme. In this

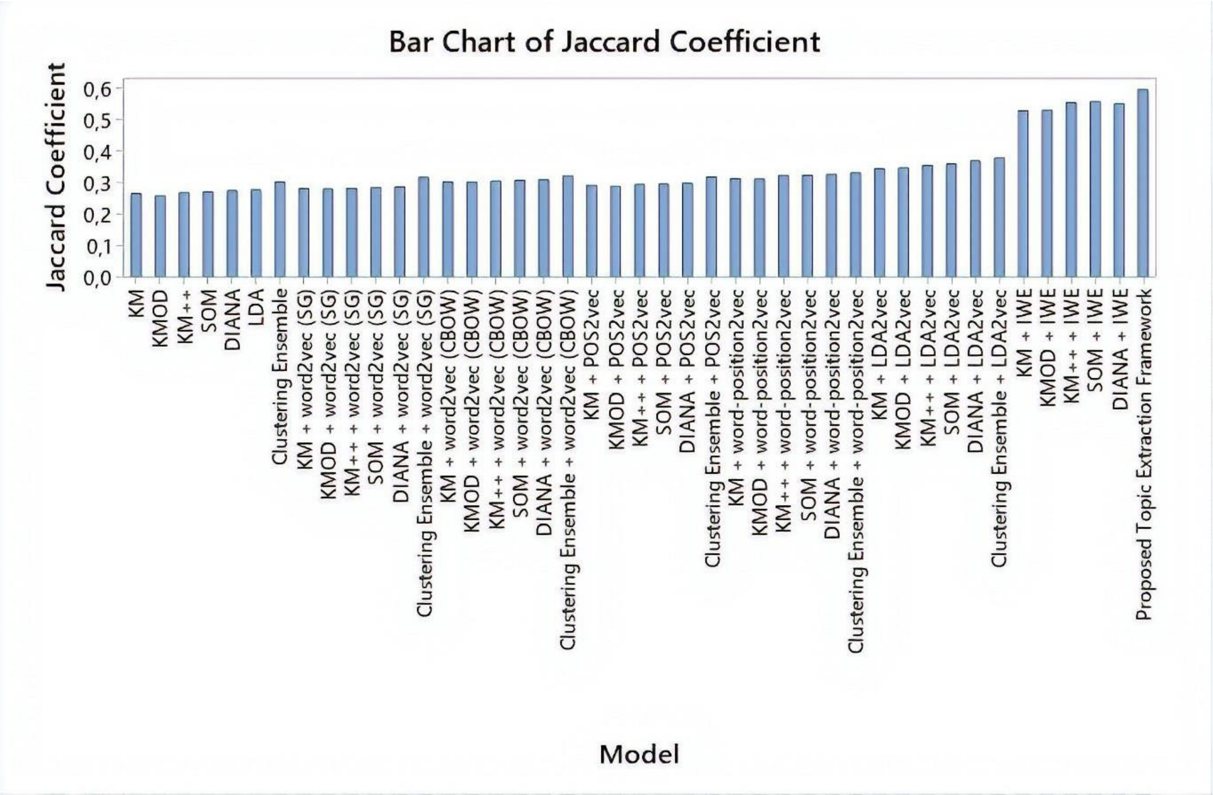

**FIGURE 4.** Bar chart of Jaccard coefficient values for compared models.

way, word2vec based word embedding scheme has been enriched by incorporating syntactic category information into the word representations. In this scheme, two words with the same part-of-speech tag have the same POS2vec vector.

### 3) WORD-POSITION2VEC

The representation of position information for words can be essential stage for computational linguistic tasks [40]. Word-position2vec word embedding scheme aims to represent relative distances of a word to the two ends of one text document. For word-position2vec implementation, we have adopted the framework outlined in [8].

### 4) LDA2VEC

The LDA2vec is another word-embedding scheme based on word2vec word embedding scheme and the latent Dirichlet allocation method [41]. In this scheme, dense word vectors from the latent document-level mixture vectors have been jointly extracted based on the Dirichlet-distribution. In the model, the skip-gram negative sampling has been utilized as the objective function [41].

## *C. CLUSTERING ENSEMBLE FRAMEWORK*

Cluster ensembles (also known as, consensus clustering) is the process of combining the different decisions obtained by several clustering algorithms to overcome the constraints of individual clustering algorithms, to enhance the robustness and the quality of clustering results [42]. Cluster ensemble requires a two-phased procedure, i.e., cluster ensemble generation and consensus function. In the cluster ensemble generation phase, the members to be included in the ensemble have been determined. One critical issue to achieve improvement in clustering quality is providing diversity among the members of ensemble [43]. To obtain different clustering results from the same data, there are several ways, such as, homogenous ensemble, random subsampling and heterogeneous ensemble. In the consensus function, the final partition of ensemble has been obtained by combing the individual clustering algorithms. Consensus functions can be classified into four groups, as direct methods, feature-based methods, pairwise similarity-based methods and graph-based methods [42].

Motivated by the enhanced clustering quality obtained on several data mining and information retrieval tasks, we present a heterogeneous clustering ensemble framework for topic extraction. The presented clustering ensemble framework utilizes k-means, k-modes, k-means++, self-organizing maps and DIANA algorithm. In the framework, iterative voting consensus function has been employed to combine the clustering results of individual clustering algorithms. The general structure for the clustering ensemble framework has been outlined in Figure 2.

The rest of this section presents the main components (base clustering algorithms and the consensus function) of the clustering ensemble framework:

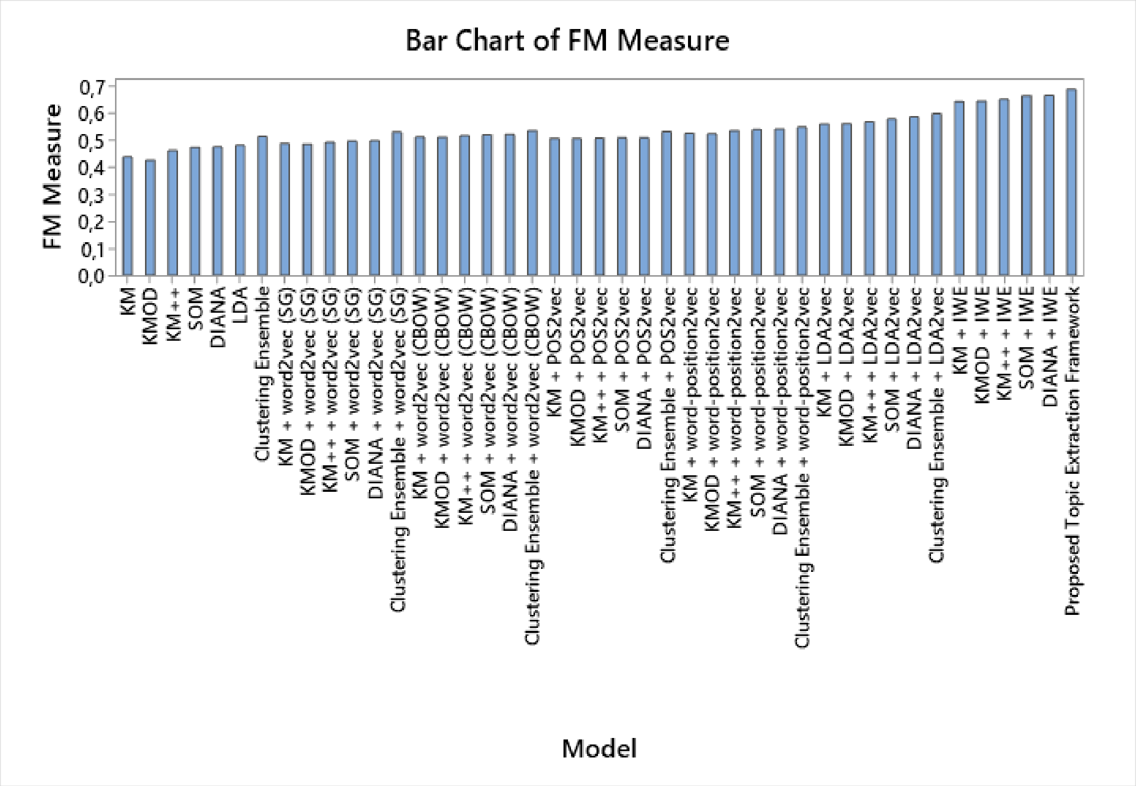

**FIGURE 5.** Bar chart of FM values for compared models.

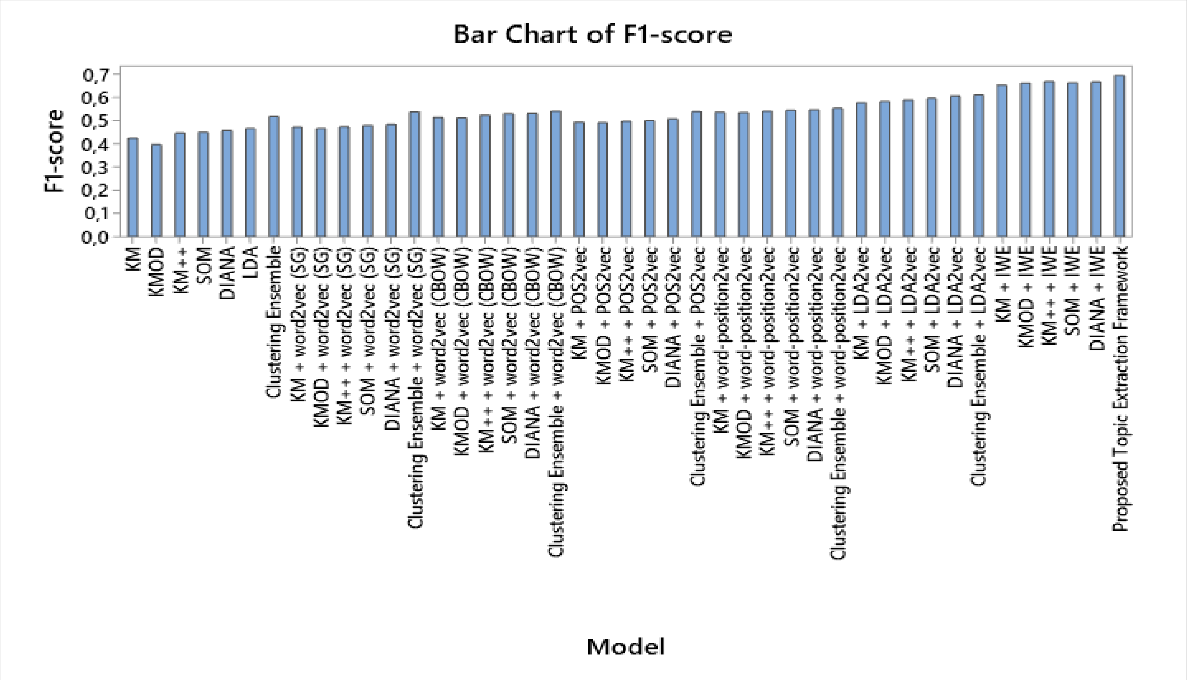

**FIGURE 6.** Bar chart of F1-score values for compared models.

### 1) K-MEANS ALGORITHM

K-means algorithm (KM) is a partition-based clustering algorithm, which is frequently employed in the cluster analysis, due to its easy implementation, simplicity and efficiency [44].

### 2) K-MODES ALGORITHM

K-modes algorithm (KMOD) is an extension of conventional k-means algorithm for cluster analysis on categorical data [45].

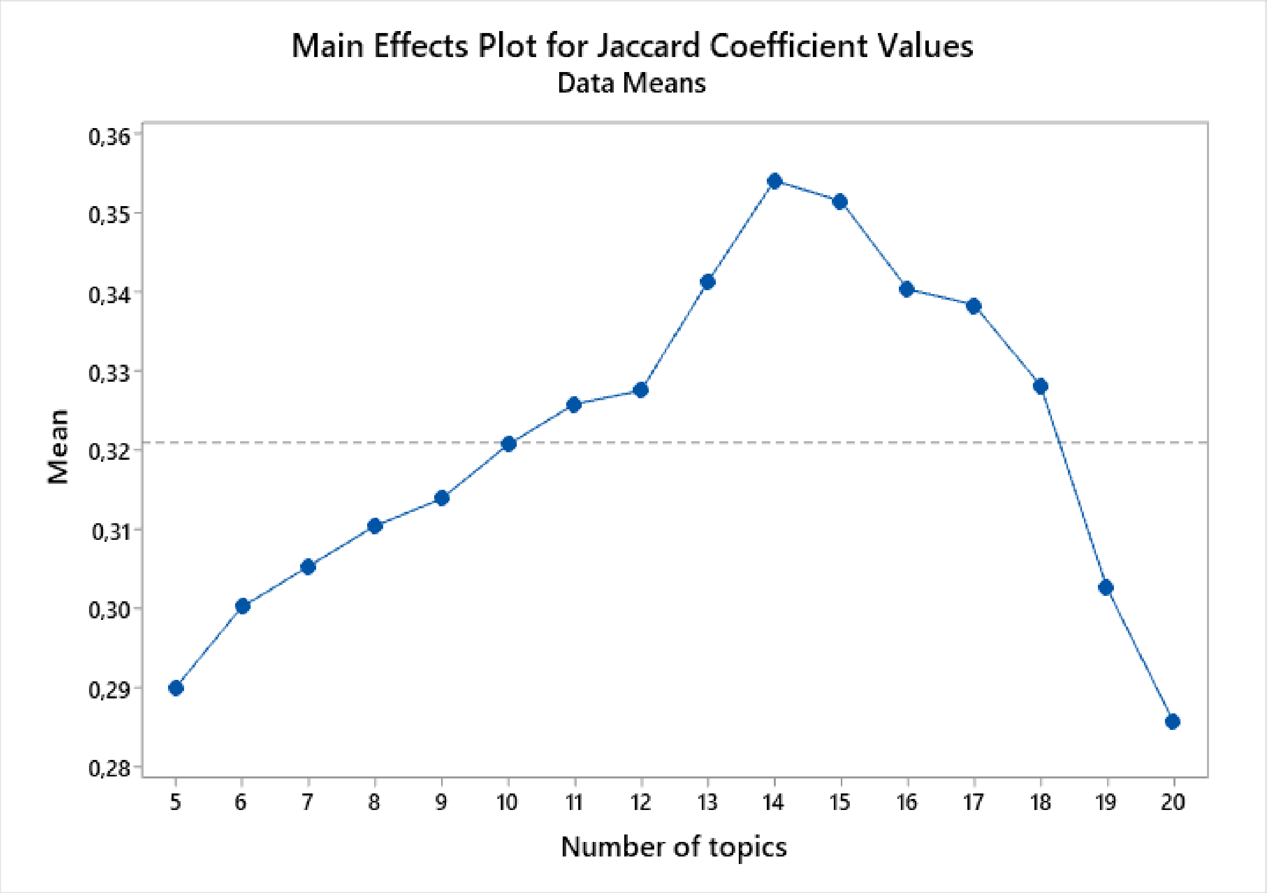

**FIGURE 7.** Main effects plot for Jaccard coefficient values based on different number of topics.

### 3) K-MEANS++ ALGORITHM

The performance of K-means algorithm is greatly influenced by the initialization of random seeds that are selected as the cluster centers. K-means++ algorithm (KM++) utilizes a heuristic function to find the initial cluster centers of K-means algorithm [46].

### 4) SELF-ORGANIZING MAPS ALGORITHM

Self-organizing map (SOM) is a partition-based clustering algorithm, which is based on clustering and dimension reduction [45]. SOM algorithm can be utilized in a wide range of applications, including sampling, supervised learning and cluster analysis [48], [49].

### 5) DIANA ALGORITHM

Divisive analysis clustering (DIANA) is a divisive hierarchical clustering algorithm [50].

### 6) ITERATIVE VOTING CONSENSUS FUNCTION

Iterative voting consensus (IVC) is a feature-based consensus function, which seeks to identify the consensus partition of data instances from the label assignment or categorical data matrix which is induced by a cluster ensemble [51]. Let we assume that we are given a set of $N$ data instances $X = \{x_1, x_2, \ldots, x_n\}$ and a set of $C$ clustering partitions $\prod = \{\pi_1, \pi_2, \ldots, \pi_C\}$ of data instances in $X$. Each clustering $\pi_i$ seeks to map from $X$ to $\{1, \ldots, n_{\pi I}\}$, where $n_{\pi I}$ denotes the number of clusters in $\pi_i$. The consensus function seeks to find a final consolidated partitioning $\pi*$ with the higher quality of clustering. IVC algorithm employs an iterative process, in which generated feature-vectors $Y = \{y_1, y_2, \ldots, y_n\}$ where $y_i$ is being specified as given by Equation 1:

$$y_i = \langle {}_1(\ x_i), {}_2\,(x_i)\,, \ldots, {}_C\,(x_i)\rangle \tag{1}$$

Each iteration of the algorithm involves a two-staged procedure. First, the cluster center of each cluster in the target consensus clustering has been identified. Then, each data instance has been assigned to the closest cluster center. Initially, the $i^{th}$ feature of the cluster center vector has been taken as the majority of all $i^{th}$ features of data instances of the cluster. Then, each data instance has been reassigned to its closest center by computing the center of the minimum distance to the examined instances. To compute the distance, the Hamming distance between two vectors has been employed. The general framework of iterative voting consensus (IVC) has been adopted from [51].

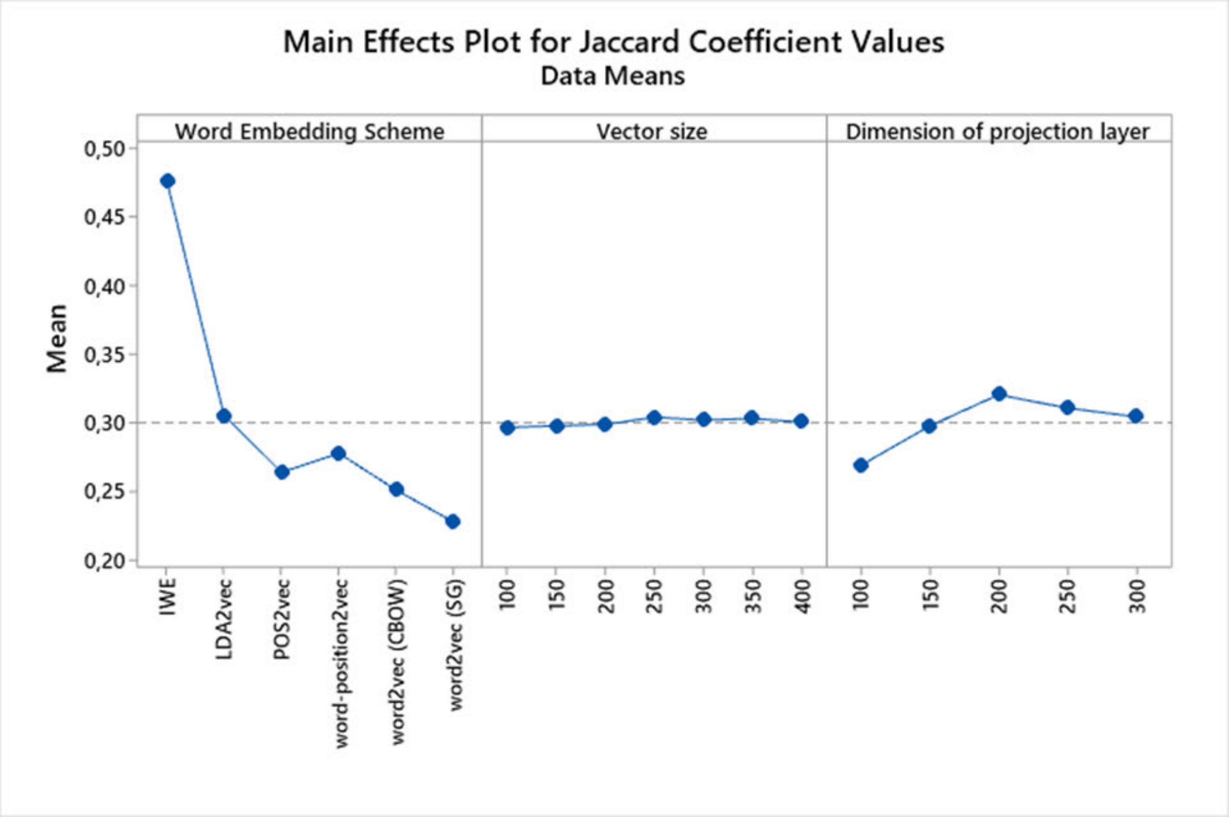

**FIGURE 8.** Main effects plot for Jaccard coefficient values based on different vector sizes and dimensions of projection layer.

### D. TOPIC EXTRACTION FRAMEWORK

The presented topic extraction framework utilizes a two-staged procedure for bibliometric data analysis. In the first stage, text corpus has been represented by a word embeddings-based approach. In this scheme, we have incorporated word vectors obtained by word2vec, POS2vec, word-position2vec and LDA2vec schemes, as outlined in Figure 1. In this way, the preprocessing tasks, feature engineering and human intervention involved in the conventional machine learning-based bibliometric data analysis have been eliminated. After extracting small set of key features based on the improved word embedding scheme, word vectors have been subject to the clustering ensemble framework, as outlined in Figure 2. As a result of employing clustering ensemble framework on improved word embedding scheme based representation, we have obtained a number of clusters. The clusters that have been generated by clustering ensemble framework have been regarded as topics (i.e. topics are the clusters). The general framework of the topic extraction framework has been summarized in Figure 3.

## IV. EXPERIMENTAL PROCEDURE AND RESULTS

In this section, the clustering evaluation metrics utilized in the empirical analysis, experimental procedure, empirical results and discussion have been presented.

### A. CLUSTERING EVALUATION METRICS

To evaluate the clustering quality of the proposed topic extraction scheme, three clustering evaluation (validation) metrics, namely, Jaccard coefficient, Folkes & Mallows (FM) measure and F1 score have been utilized. The evaluation metrics are counting-pair based clustering evaluation metrics. Let $a$ denote the number of pairs of instances that are grouped in the same cluster by the clustering algorithm and fall within the same class in the original dataset. Let $b$ denote the number of pairs of instances that are not grouped in the same cluster but fall within the same class in the original dataset. Let $c$ denote the number of pairs of instances that are grouped in the same cluster but do not fall within the same class in the original dataset. Based on $a$, $b$and $c$,Jaccard coefficient (JC), Folkes & Mallows (FM) measure and F1-score have been computed as given by Equations 2-4, respectively:

$$JC = \frac{a}{a+b+c} \tag{2}$$

$$FM = \left(\frac{a}{a+b}\frac{a}{a+c}\right)^{1/2} \tag{3}$$

$$F1 - score = \frac{2a^2}{2a^2+ac+ab} \tag{4}$$

### B. EXPERIMENTAL PROCEDURE

In the empirical analysis, we have compared the clustering quality obtained by the proposed scheme with seven

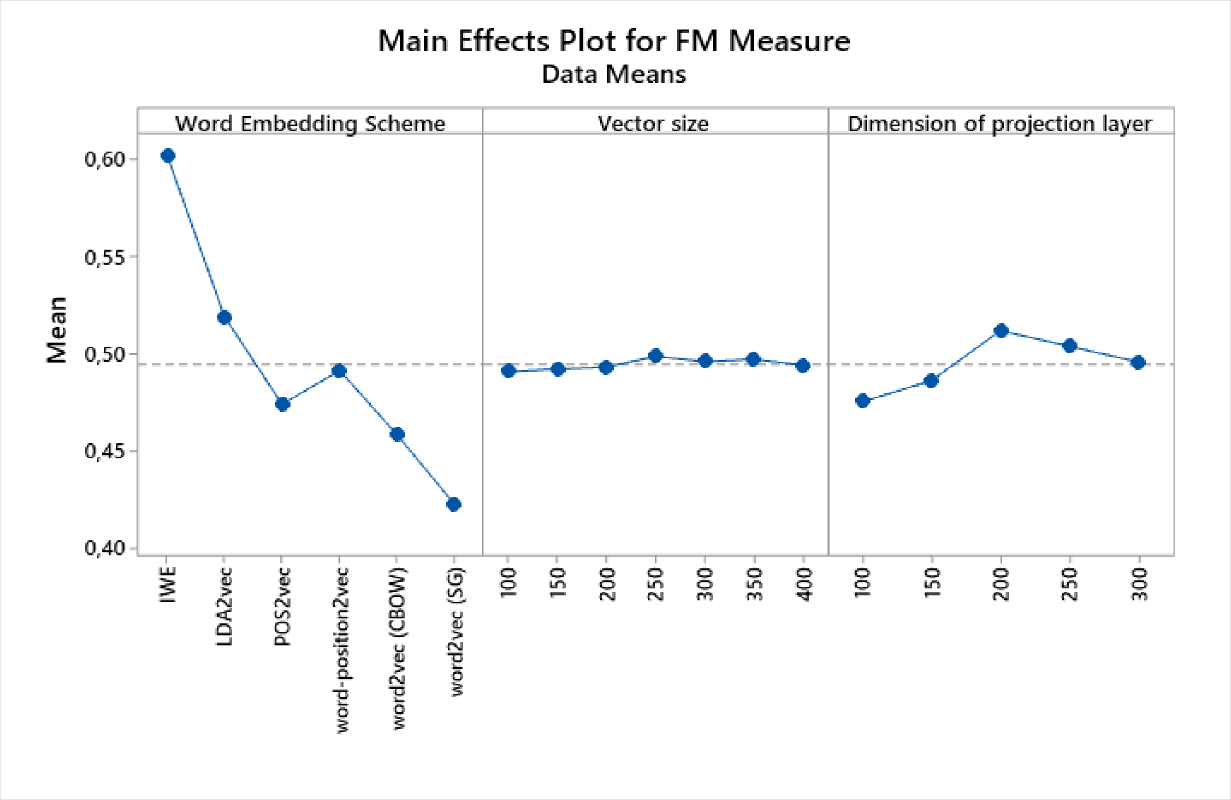

**FIGURE 9.** Main effects plot for FM values based on different vector sizes and dimensions of projection layer.

clustering-based baseline models. The clustering-based baseline models include k-means, k-modes, k-means++, self-organizing maps, DIANA algorithm, latent Dirichlet allocation and the clustering ensemble framework. In the clustering ensemble framework, only ensemble framework outlined in Figure 2 has been employed. All the baseline models have been implemented in Python. For clustering algorithms which involve the number of clusters ($k$) as the input parameter, we have made empirical analysis with different number of clusters, ranging from 5 to 20, and the highest clustering quality obtained by the base clustering algorithm has been reported in this section. For the latent Dirichlet allocation algorithm, we have utilized symmetric Dirichlet priors in the LDA estimation with $\alpha = /50$ K and $\beta = 0.01$, which are frequently utilized parameter values for LDA. In addition to the baseline models, we have considered word2vec (skip-gram model) based models. For word2vec based models, we have integrated baseline clustering methods (namely, k-means, k-modes, k-means++, self-organizing maps and DIANA algorithm and clustering ensemble framework) with word2vec word embedding scheme. In other words, we have employed a two-staged procedure, in which bibliometric text corpus has been initially represented by word2vec (skip-gram model) followed by application of any clustering algorithm. The same procedure has been also repeated for other word-embedding schemes. Namely, we have also integrated word2vec (continuous bag-of-words model), pos2vec, word-position2vec, LDA2vec and the improved word embedding scheme (IWE) word embeddings in conjunction with clustering algorithms. Taking all these configurations into account, we have 43 configurations in total. For LDA2vec, we have made empirical analysis with several parameter sets. Since the best predictive performance has been obtained by the negative sampling exponent ($\beta \in 0.75$), we have reported the results for this parameter value. For the word-embedding schemes, vector size has been set to 250.

### C. EXPERIMENTAL RESULTS

Forty-three different configurations were compared in the empirical analysis, as shown in Table 2, based on three cluster evaluation measures. Based on the observations collected from empirical analysis, there are several managerial insights for topical extraction. Regarding the clustering quality obtained by conventional clustering algorithms, DIANA algorithm outperforms the other clustering baselines. The second highest clustering quality in terms of cluster validation measures has been obtained by self-organizing maps algorithm, which is followed by k-means++ algorithm. The lowest clustering quality has been obtained by k-modes clustering algorithm. As it can be observed from the results presented in Table 2, the latent Dirichlet allocation algorithm outperforms the conventional clustering baselines. In addition, empirical results indicate that combining the

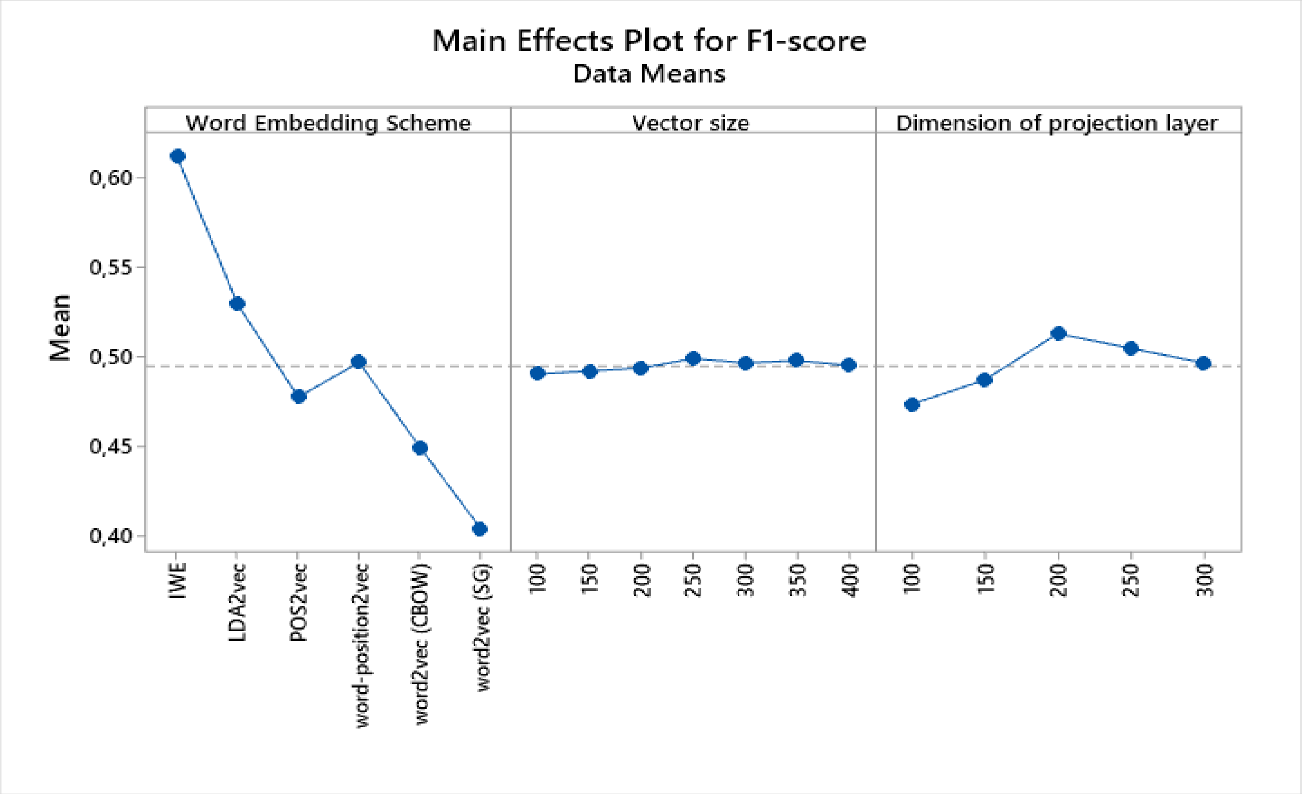

**FIGURE 10.** Main effects plot for F1-score values based on different vector sizes and dimensions of projection layer.

cluster partitions obtained by several clustering algorithms can achieve more stable and robust clustering results.

To summarize the main findings of the experimental analysis, bar charts for Jaccard coefficient, FM measure and F1-score values have been presented in Figures 4-6, respectively.

Regarding the performance of different word embedding schemes taken into consideration, the highest clustering quality has been obtained by the integration of LDA2vec with clustering algorithms. The second highest performance has been achieved by the utilization of word-position2vec word embedding scheme in conjunction with clustering methods.

The clustering ensemble framework which incorporates conventional clustering methods (i.e., k-means, k-modes, k-means++, self-organizing maps and DIANA algorithm) by means of the iterative voting consensus outperforms the baseline clustering methods and the latent Dirichlet allocation model. The second concern of the empirical analysis is to evaluate whether clustering quality can be enhanced with the integration of word embedding schemes to clustering methods. As it can be seen from the empirical results, word embedding schemes significantly enhance the performance of clustering algorithms.

Regarding the clustering quality obtained by different pre-trained word embedding schemes, POS2vec outperforms word2vec based schemes. word2vec continuous bag-of-words model yields higher clustering quality compared to word2vec skip-gram model. The proposed topic extraction framework, which employs improved word embedding scheme-based representation on bibliometric data, followed by the utilization of clustering ensemble framework generated the highest (best) performance metric values of all the compared groups. The highest performance with a Jaccard coefficient value of 0.5951, Folkes & Mallows (FM) value of 0.6849 and F1-score value of 0.6930 has been achieved by the proposed scheme. Hence, the empirical analysis indicates that clustering ensemble improves the clustering quality of conventional clustering algorithms, the utilization of word embedding schemes enhance the clustering quality of clustering algorithms and the integration of several word embedding schemes can yield higher predictive performance. As outlined in Figure 3, the proposed topic extraction framework consists of improved word embedding scheme based representation and clustering ensemble framework. To validate the usefulness of the components of the proposed ensemble, we have employed two ablation tests, one is dedicated to clustering algorithms and the other one is dedicated to word embedding schemes. For ablation test on clustering algorithms, we have considered five baseline clustering algorithms (namely, k-means, k-modes, k-means++, self-organizing maps, and DIANA algorithm) and their ensemble combinations. In this way, we have examined $2^5 - 1$ (31 configurations) in total. For ablation test on word embedding schemes, we have considered five baseline word embedding schemes (namely,

**TABLE 3.** Ablation test for compared clustering algorithms.

| Model | Jaccard Coefficient | FM Measure | F1-score |
|---|---|---|---|
| KM | 0.2650 | 0.4369 | 0.4231 |
| KMOD | 0.2572 | 0.4241 | 0.3961 |
| KM++ | 0.2678 | 0.4598 | 0.4449 |
| SOM | 0.2712 | 0.4697 | 0.4489 |
| DIANA | 0.2744 | 0.4738 | 0.4561 |
| KM + KMOD | 0.2696 | 0.4570 | 0.4353 |
| KM + KM++ | 0.2722 | 0.4657 | 0.4361 |
| KM + SOM | 0.2755 | 0.4701 | 0.4385 |
| KM + DIANA | 0.2759 | 0.4732 | 0.4415 |
| KMOD + KM++ | 0.2653 | 0.4509 | 0.4195 |
| KMOD + SOM | 0.2687 | 0.4518 | 0.4243 |
| KMOD + DIANA | 0.2688 | 0.4519 | 0.4291 |
| KM++ + SOM | 0.2802 | 0.4738 | 0.4423 |
| KM++ + DIANA | 0.2815 | 0.4743 | 0.4482 |
| SOM + DIANA | 0.2817 | 0.4769 | 0.4518 |
| KM + KMOD + KM++ | 0.2871 | 0.4883 | 0.4662 |
| KM + KMOD + SOM | 0.2879 | 0.4911 | 0.4776 |
| KM + KMOD + DIANA | 0.2883 | 0.4912 | 0.4776 |
| KM + KM++ + SOM | 0.2901 | 0.4933 | 0.4782 |
| KM + KM++ + DIANA | 0.2905 | 0.5000 | 0.4805 |
| KM + SOM + DIANA | 0.2910 | 0.5019 | 0.4831 |
| KMOD + KM++ + SOM | 0.2831 | 0.4782 | 0.4519 |
| KMOD + KM++ + DIANA | 0.2850 | 0.4835 | 0.4613 |
| KMOD + SOM + DIANA | 0.2857 | 0.4842 | 0.4620 |
| KM++ + SOM + DIANA | 0.2929 | 0.5023 | 0.4840 |
| KM + KMOD + KM++ + SOM | 0.2915 | 0.4843 | 0.4882 |
| KM + KMOD + KM++ + DIANA | 0.2917 | 0.4869 | 0.4718 |
| KM + KMOD + SOM + DIANA | 0.2931 | 0.4882 | 0.4819 |
| KM + KM++ + SOM + DIANA | 0.2950 | 0.4935 | 0.4913 |
| KMOD + KM++ + SOM + DIANA | 0.2957 | 0.4942 | 0.4920 |
| Clustering Ensemble | 0.3028 | 0.5114 | 0.5169 |

KM: K-means. KMOD: K-modes. KM++: K-means++. SOM: Self-organizing maps

**TABLE 4.** Ablation test for compared word embedding schemes.

| Model | Jaccard Coefficient | FM Measure | F1-score |
|---|---|---|---|
| KM | 0.2650 | 0.4369 | 0.4231 |
| word2vec (SG) | 0.2814 | 0.4855 | 0.4711 |
| word2vec (CBOW) | 0.3026 | 0.5103 | 0.5123 |
| POS2vec | 0.2909 | 0.5037 | 0.4919 |
| word-position2vec | 0.3129 | 0.5234 | 0.5338 |
| LDA2vec | 0.3438 | 0.5560 | 0.5749 |
| word2vec (SG) + word2vec (CBOW) | 0.3723 | 0.5732 | 0.5793 |
| word2vec (SG) + POS2vec | 0.3551 | 0.5713 | 0.5786 |
| word2vec (SG) + word-position2vec | 0.3851 | 0.5748 | 0.5906 |
| word2vec (SG) + LDA2vec | 0.3867 | 0.5749 | 0.5914 |
| word2vec (CBOW) + POS2vec | 0.3924 | 0.5787 | 0.5942 |
| word2vec (CBOW) + word-position2vec | 0.4041 | 0.5852 | 0.5971 |
| word2vec (CBOW) + LDA2vec | 0.4084 | 0.5861 | 0.6016 |
| POS2vec + word-position2vec | 0.4002 | 0.5826 | 0.5946 |
| POS2vec + LDA2vec | 0.4021 | 0.5838 | 0.5961 |
| word-position2vec + LDA2vec | 0.4092 | 0.5884 | 0.6096 |
| word2vec (SG) + word2vec (CBOW) + POS2vec | 0.4340 | 0.5891 | 0.6113 |
| word2vec (SG) + word2vec (CBOW) + word-position2vec | 0.4405 | 0.5933 | 0.6113 |
| word2vec (SG) + word2vec (CBOW) + LDA2vec | 0.4848 | 0.5993 | 0.6183 |
| word2vec (SG) + POS2vec + LDA2vec | 0.4715 | 0.5991 | 0.6174 |
| word2vec (SG) + word-position2vec + LDA2vec | 0.4870 | 0.6039 | 0.6214 |
| word2vec (CBOW) + POS2vec + word-position2vec | 0.4685 | 0.5971 | 0.6151 |
| word2vec (CBOW) + POS2vec + LDA2vec | 0.4772 | 0.6025 | 0.6214 |
| word2vec (CBOW) + word-position2vec + LDA2vec | 0.4883 | 0.6132 | 0.6301 |
| POS2vec + word-position2vec + LDA2vec | 0.4858 | 0.6075 | 0.6287 |
| word2vec (SG) + word2vec (CBOW) + POS2vec + word-position2vec | 0.4948 | 0.6154 | 0.6313 |
| word2vec (SG) + word2vec (CBOW) + POS2vec + LDA2vec | 0.4960 | 0.6162 | 0.6329 |
| word2vec (SG) + word2vec (CBOW) + word-position2vec + LDA2vec | 0.4962 | 0.6194 | 0.6368 |
| word2vec (SG) + POS2vec + word-position2vec + LDA2vec | 0.5194 | 0.6243 | 0.6374 |
| IWE | 0.5289 | 0.6398 | 0.6499 |

word2vec skip-gram model, word2vec continuous bag-of-word model, POS2vec, word-position2vec and LDA2vec) and their ensemble configurations. In this way, we have examined 30 configurations. The ablation test results on clustering algorithms and word embedding schemes have been presented in Table 3 and Table 4, respectively. For the empirical results listed in Table 4, the predictive performances of word embedding schemes in conjunction with k-means clustering algorithm have been reported.

The empirical results listed in Table 3 indicate that clustering ensembles (configurations incorporating several baseline clustering algorithms) yield higher predictive performance compared to the baseline clustering algorithms. The highest performance among the baseline clustering algorithms has been achieved by DIANA algorithm and the second highest performance has been achieved by self-organizing

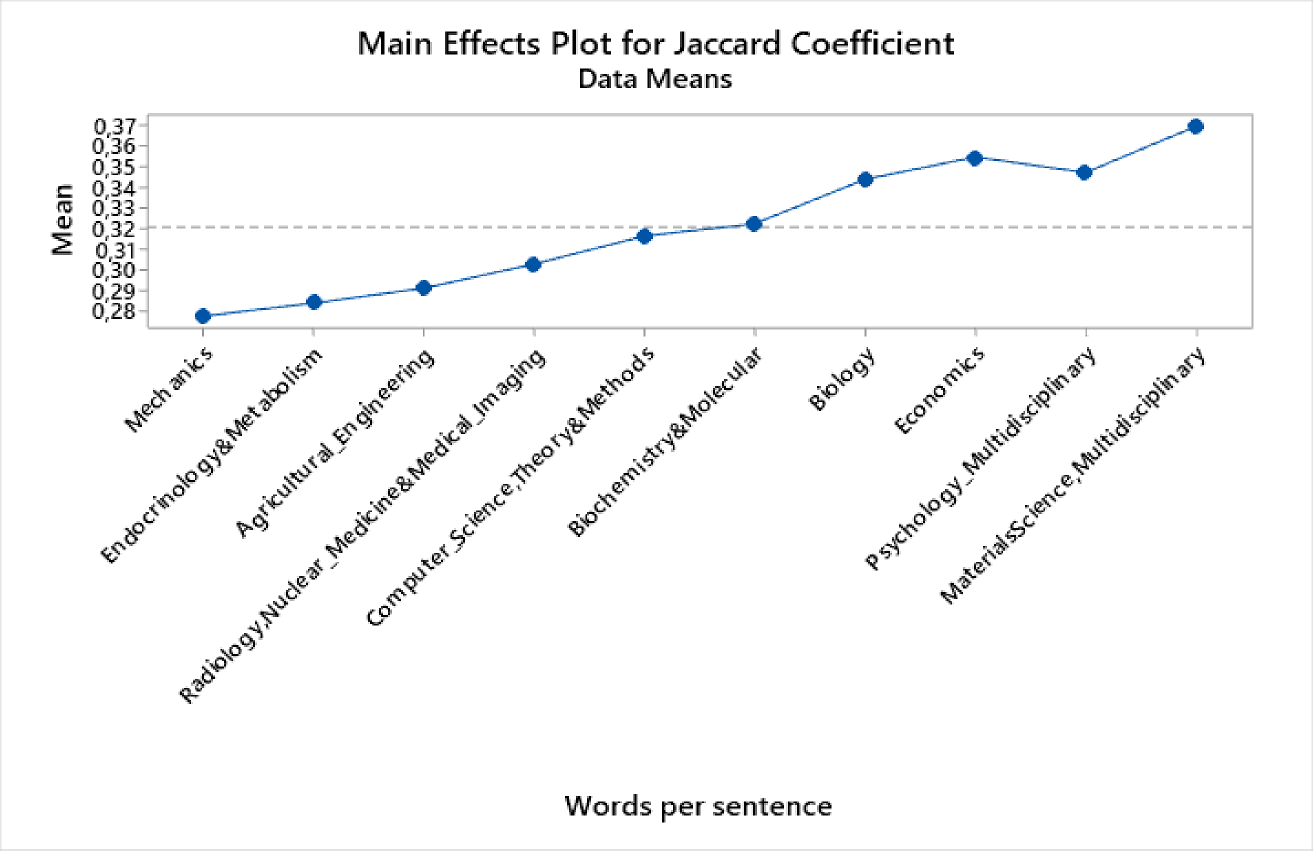

**FIGURE 11.** Main effects plot for Jaccard coefficient values based on words per sentence for different categories.

maps algorithm. Among the results obtained by combining two clustering algorithms, the highest performance has been obtained by combining self-organizing maps and DIANA algorithm. Regarding the results obtained by combining three clustering algorithms, the highest performance has been achieved by ensemble of k-means++, self-organizing maps and DIANA algorithm. Regarding the results obtained by combining four clustering algorithms, the highest performance has been achieve by ensemble of k-modes algorithm, k-means++, self-organizing maps and DIANA algorithm. The proposed clustering ensemble framework contains k-means, k-modes, k-means++, self-organizing maps and DIANA algorithm. The highest predictive performance in terms of cluster validation measures has been achieve by the proposed clustering ensemble framework.

The empirical results listed in Table 4 indicate that the incorporation of several word embedding schemes outperform the baseline word embedding schemes. The highest performance among the baseline word embedding schemes has been achieve by LDA2vec scheme and the second highest predictive performance has been achieved by word-position2vec scheme. In a similar way, the configurations which utilize LDA2vec and word-position2vec scheme in their ensembles generally yield higher predictive performance results. The empirical results indicate word2vec continuous bag-of-word model based configurations generally outperform the configurations which contain continuous bag-of-word model. The proposed improved word embedding scheme yields the highest performance among all the configurations taken into consideration.

To analyze the effects of different number of topics, vector sizes and dimensions of the projection layer on topic extraction framework, we have conducted parameter analysis. The performance metric values have been evaluated on 16 different number of topics (ranging from 5 to 20). The main effects plot for number of topics based on Jaccard coefficient values has been presented in Figure 7. As it can be observed from the results presented in Figure 7, the compared methods reach their peak predictive performance at 14 topics. From 5 to 14 topics, the mean predictive performance values (in terms of Jaccard coefficient) obtained by the compared methods have been increasing. However, the mean predictive performance has been steadily decreasing after reaching its peak at 14 topics. The same pattern is also valid for other compared evaluation measures and the base methods.

In Figures 8-10, main effects plots for cluster evaluation metrics based on different vector sizes and dimensions of projection layer have been presented. As it can be seen from the figures, the highest predictive performances have been

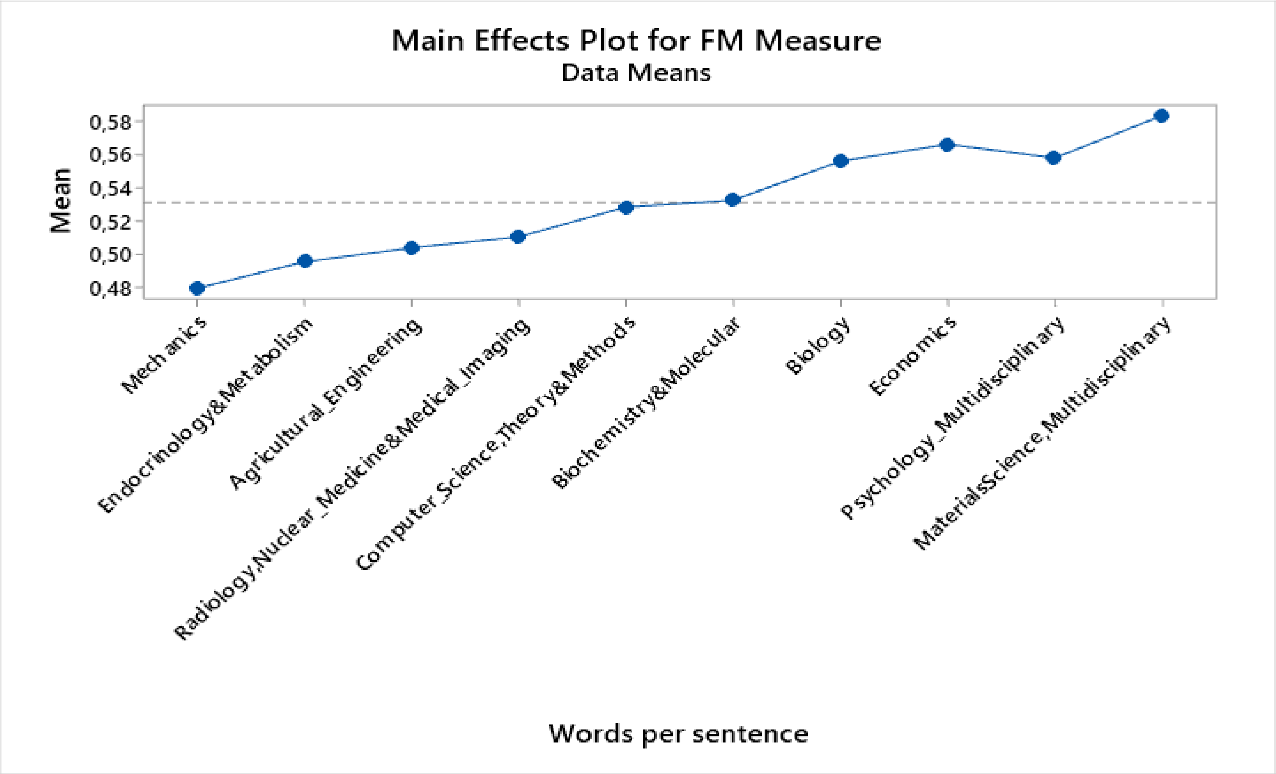

**FIGURE 12.** Main effects plot for FM measure values based on words per sentence for different categories.

achieved by the improved word embedding scheme, the second highest predictive performances have been achieved by LDA2vec word embedding scheme. Regarding the effect of vector size on topic extraction, seven different vector sizes (ranging from 100 to 400) have been considered. As it can be seen from the empirical results, the highest predictive performances have been achieved for vector size of 250 for the word embedding schemes. Regarding the effect of dimension of projection layer on topic extraction, five different dimension values (ranging from 100 to 300) have been taken into account. As it can be observed from Figures 8-10, the highest predictive performances have been achieved for dimension of 200 for word embedding schemes.

In Figures 11-13, main effect plots for cluster evaluation metrics based on words per sentence for different categories have been presented. As outlined in Table 1, the lowest average words per sentence has been encountered in mechanics category, the second lowest average words per sentence has been encountered in endocrinology & metabolism category. The highest average words per sentence has been encountered in mathematics applied category and the second highest average words per sentence has been encountered in materials science multidisciplinary category. As it can be observed from the results depicted in Figures 11-13, the highest average cluster evaluation metric values in terms of Jaccard coefficient, FM values and F1-scores have been achieved for materials science multidisciplinary category. The lowest average performance values have been achieved by mechanics category. Regarding the performance of topic extraction schemes, the empirical results indicate that the predictive performance enhances as the average words per sentence increases for a particular category.

To further evaluate the statistical significance of the empirical results, we have performed one-way ANOVA test in Minitab statistical analysis software. The results for the one-way ANOVA test of overall results obtained by the conventional clustering methods, word embedding based schemes and the proposed scheme have been presented in Table 5, where DF, SS, MS, F and P denote degrees of freedom,

**TABLE 5.** One-way Anova test results.

| Jaccard Coefficient Values | | | | | |
|---|---|---|---|---|---|
| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
| Model | 6 | 0.339635 | 0.056606 | 283.78 | 0.000 |
| Error | 36 | 0.007181 | 0.000199 | | |
| Total | 42 | 0.346816 | | | |
| FM Measure Values | | | | | |
| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
| Model | 6 | 0.14621 | 0.024369 | 85.75 | 0.000 |
| Error | 36 | 0.01023 | 0.000284 | | |
| Total | 42 | 0.15644 | | | |
| F1-Score Values | | | | | |
| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
| Model | 6 | 0.19586 | 0.032644 | 73.94 | 0.000 |
| Error | 36 | 0.01589 | 0.000442 | | |
| Total | 42 | 0.21176 | | | |

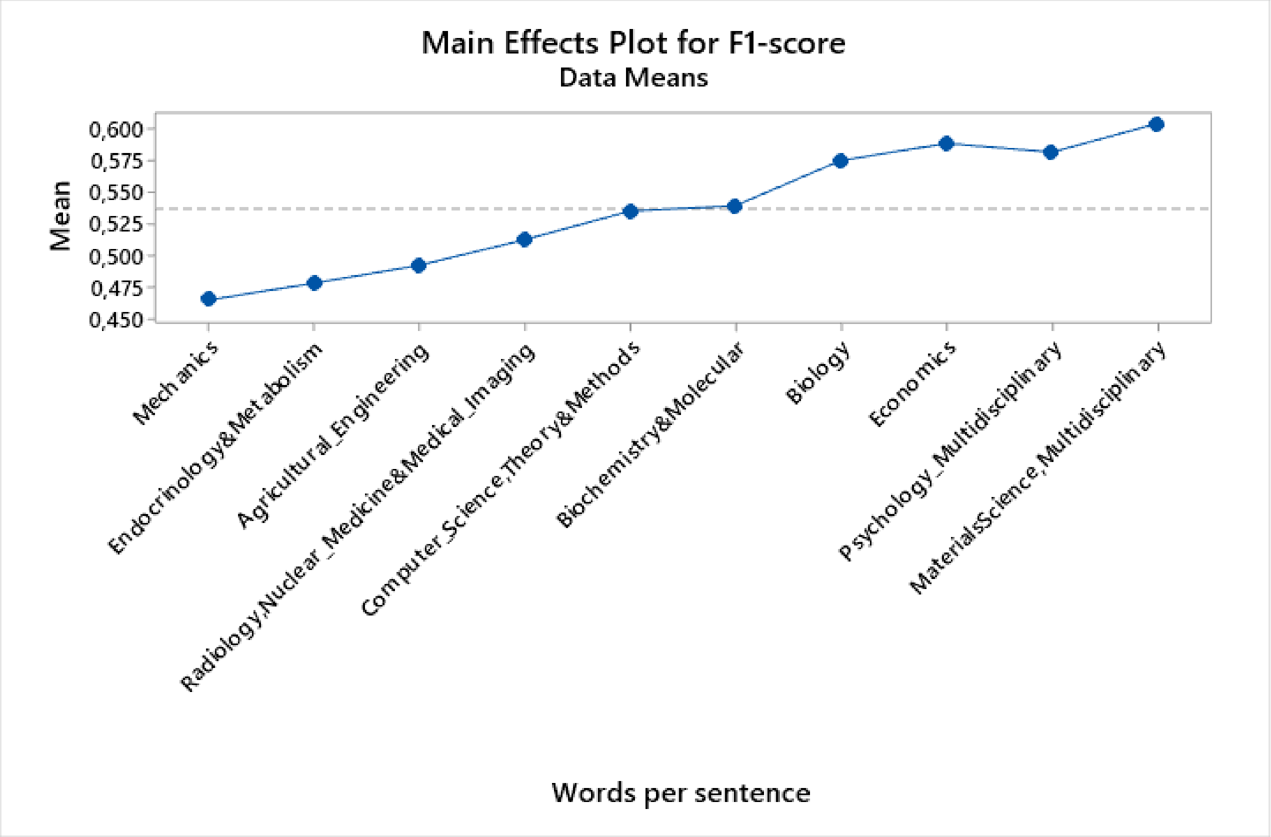

**FIGURE 13.** Main effects plot for F1-score values based on words per sentence for different categories.

adjusted sum of squares, adjusted mean square, F-Value and probability value, respectively.

According to the one-way ANOVA test results presented in Table 5, there are statistically meaningful differences between the clustering quality values obtained by compared groups ($p<0.0001$). To further indicate the statistical meaningfulness of higher predictive performance obtained by the proposed scheme, confidence intervals for the mean values of compared algorithms in terms of three clustering evaluation measures for a confidence level of 95% have been presented in Figures 14-16. Based on the statistical significances between the results of compared models, Figures 14-16 have been divided into two regions denoted by red dashed lines. For the results presented in Figures 14-16, confidence intervals for baseline models, word2vec models (SG), word2vec models (CBOW), POS2vec models and word-position2vec model overlap. However, confidence interval obtained by the improved word embedding scheme has been deployed in another region of interval plots. Since the intervals for the baseline models and the improved word embedding scheme do not overlap, the population means are statistically meaningful. As it can be observed from interval plots, predictive performance differences obtained by the proposed scheme is statistically significant.

In addition, Figure 17 presents t-Distributed stochastic neighbor embedding (t-SNE) plot for topics extracted for different categories. t-SNE is an effective dimensionality reduction method for the visualization of high-dimensional datasets [60]. Two-dimensional map determined by t-SNE indicates that there are several overlapping topics between different subject categories. There are overlapping topics between subject categories computer science and economics. In addition, there are overlapping topics between mechanics and mathematics (applied). In addition, there are overlapping topics between materials science (multidisciplinary) and mechanics.

## V. DISCUSSION

Based on the empirical analysis on clustering methods, topic modelling and word embedding schemes, several insights follow:

The utilization of word embedding schemes in conjunction with clustering methods enhance the predictive performance of clustering approaches for topic extraction. The utilization of word embedding schemes outperform the baseline clustering results. Word embedding schemes can be utilized to extract syntactic and semantic relations among the words and phrases. The experimental analysis indicates that word embedding schemes are viable tools to capture latent features on topic extraction.

Topic modelling approach (the latent Dirichlet allocation method) yields higher predictive performance for topic

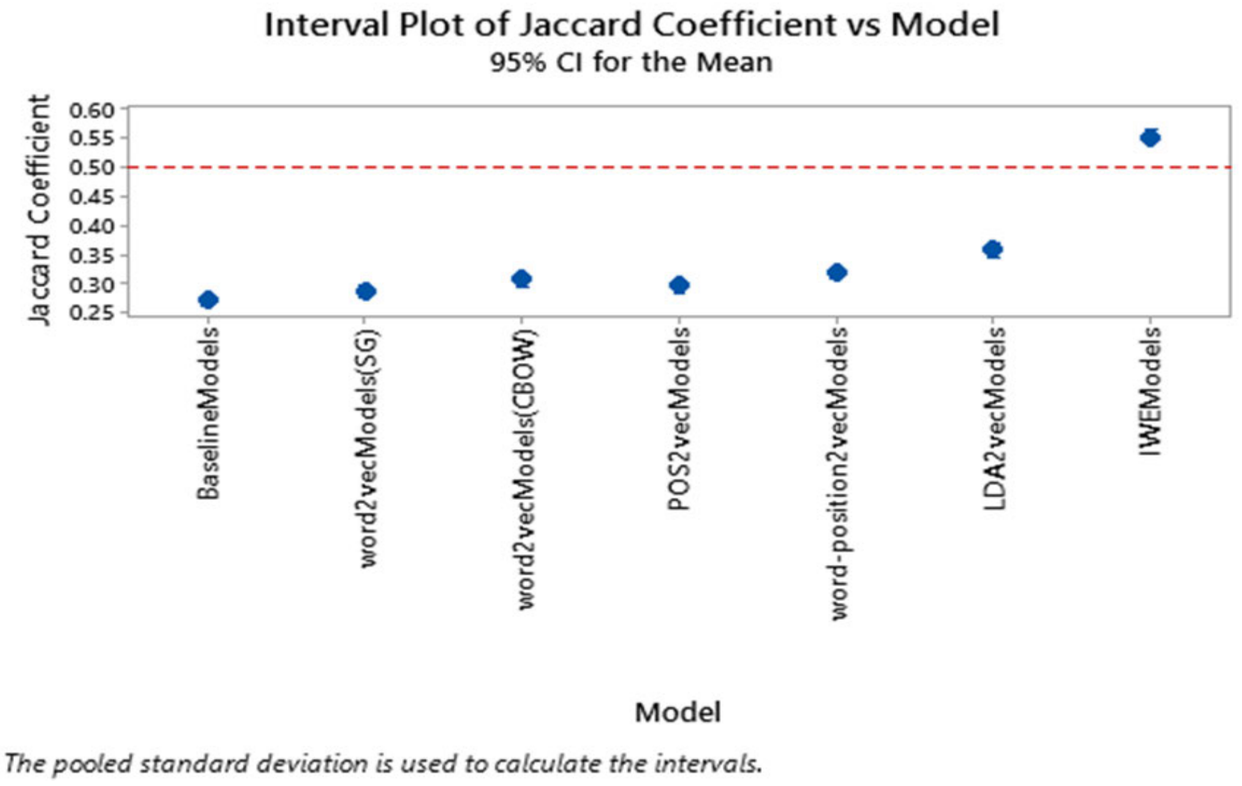

**FIGURE 14.** Interval plot of Jaccard coefficient values for compared models.

extraction compared to the conventional clustering algorithms. These empirical results are in line with recent empirical results on topic extraction [11, 52].

In the empirical analysis, five conventional word embedding schemes (word2vec skip-gram model, word2vec continuous bag-of-word model, POS2vec, word-position2vec and LDA2vec) have been considered. The empirical analysis indicates that LDA2vec word embedding scheme provides more enhancement in performance of clustering methods than the other word embedding schemes. The LDA2vec scheme is based on the word2vec skip-gram model and the latent Dirichlet allocation model. The empirical analysis validates that obtaining a context vector by adding the pivot word vector and a document vector can enhance the predictive performance for topic extraction.

In the empirical analysis, five conventional clustering algorithms (namely, k-means, k-modes, k-means++, self-organizing maps and DIANA algorithm) have been considered. The empirical results indicate that DIANA algorithm outperforms the other clustering baselines.

The proposed two-stage topic extraction model utilizes the improved word embedding scheme and clustering ensemble framework. The empirical analysis indicates that the ensemble word embedding scheme, which incorporates word vectors obtained by word2vec, POS2vec, word-position2vec and LDA2vec schemes, can yield higher predictive performance in terms of cluster validation measures for topic extraction. The ensemble word vectors have been researched for other tasks in natural language processing with promising results [7], [30]. The empirical analysis indicates that the ensemble word vectors can yield also promising results for topic extraction.

Clustering ensemble framework outperforms the conventional clustering baselines. Cluster ensemble is the process of integrating the several clustering algorithms to obtain a more robust clustering result with higher clustering quality. Clustering ensembles yield promising results on several tasks in information retrieval. These empirical results are in line with recent empirical results on clustering ensembles [53], [54]. The experimental analysis indicates that clustering ensembles can be utilized for topic-extraction related tasks.

Regarding the effect of vector size on topic extraction, the highest predictive performances have been achieved for vector size of 250. Regarding the effect of dimension of projection layer on topic extraction, five different dimension values (ranging from 100 to 300) have been taken into account. As it can be observed from Figures 8-10, the highest predictive performances have been achieved for dimension of 200 for word embedding schemes.

In the empirical analysis, the effect of average words per sentence for different categories have been examined.

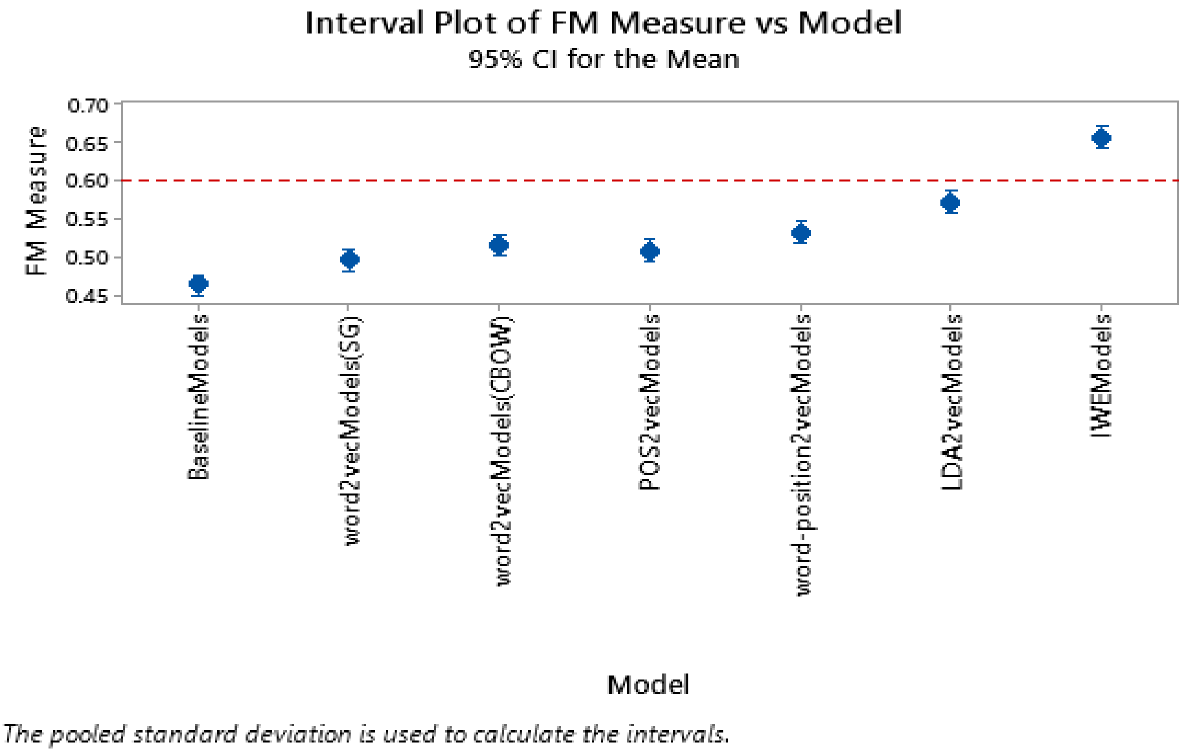

**FIGURE 15.** Interval plot of FM values for compared models.

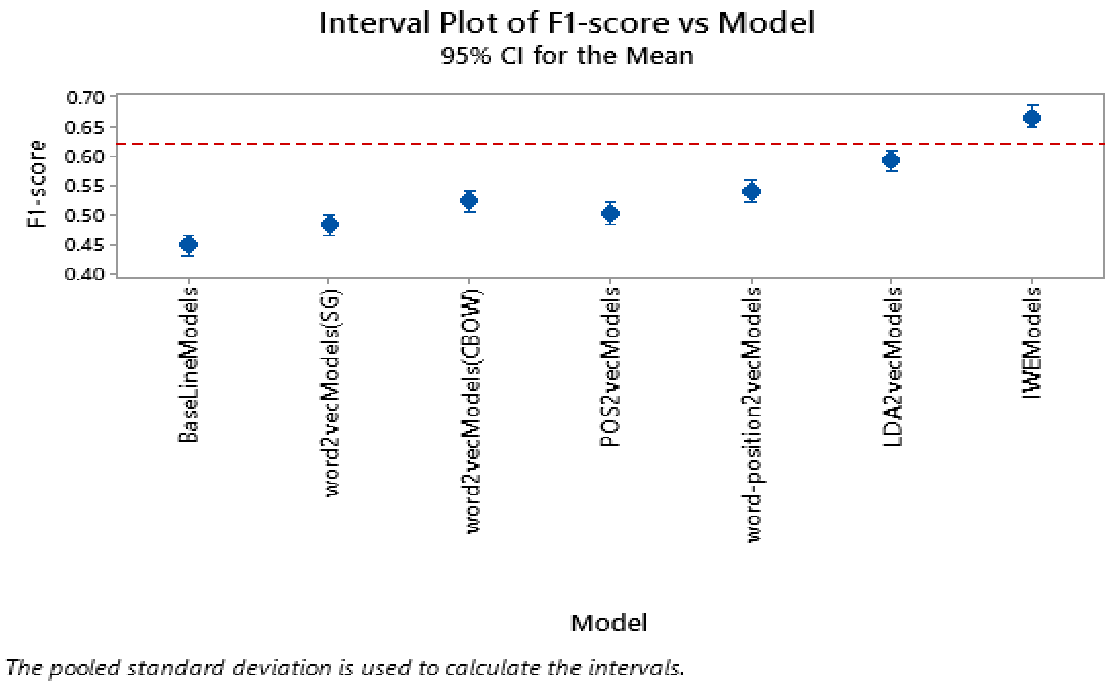

**FIGURE 16.** Interval plot of F1-score values for compared models.

Regarding the performance of topic extraction schemes, the empirical results indicate that the predictive performance enhances as the average words per sentence increases for a particular category.

The clustering ensemble framework enhances the clustering quality of conventional clustering algorithms, the utilization of word embedding schemes enhance the clustering quality of clustering algorithms and the integration of sev-

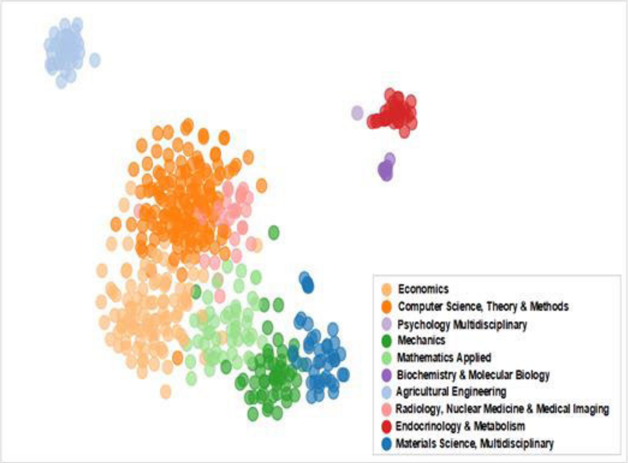

**FIGURE 17.** t-SNE plot for topics obtained by different subject categories.

eral word embedding schemes can yield higher predictive performance.

In the proposed improved word embedding scheme, we have incorporated word vectors obtained by word2vec, POS2vec, word-position2vec and LDA2vec. In this scheme, several word embedding schemes have been integrated for an improved word embedding. The empirical results on machine learning tasks indicate that weighted voting schemes can enhance the predictive performance of ensemble systems [55], [56]. The proposed scheme does not assign any weights to define the contributions of the word embedding methods on the final embedding. It should be beneficial to extend this research by adjusting optimal weight values to word embedding schemes. In addition, recent proposals on text classification tasks indicate that weighting schemes (such as, inverse document frequency, smooth inverse frequency and subsampling function) can be integrated into word embedding schemes, such as word2vec or GloVe, to compute vector embedding for a text document [57]–[59]. In this regard, it should be beneficial to extend this research by taking these weighting schemes into account.

## VI. CONCLUSION

Topic extraction is an essential task in bibliometric data analysis, data mining and information retrieval. The conventional topic extraction schemes require comprehensive pre-processing tasks to represent text collections in an appropriate way. Word embedding schemes are applications of deep learning in natural language processing, in which words or phrases are represented in low-dimensional vectors of a continuous space. Word embedding schemes, such as, word2vec and global vectors (GloVe) have been successfully employed for natural language processing tasks. In this contribution, we present a two-staged framework for topic extraction from scientific literature. The presented scheme employs a two-staged procedure, where word embedding schemes have been utilized in conjunction with cluster analysis. The empirical analysis on a corpus containing 160,424 abstracts of articles from various disciplines, including economics, engineering and computer science, indicate that the utilization of word embedding schemes significantly enhance the performance of clustering algorithms.

## REFERENCES

[1] K. Hofmann, M. Tsagkias, E. Meij, and M. D. Rijke, "A comparative study of features for keyphrase extraction in scientific literature," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, Hong Kong, 2009, pp. 1725–1728.

[2] A. Bagheri, M. Saraee, and F. de Jong, "ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences," *J. Inf. Sci.*, vol. 40, no. 5, pp. 621–636, 2014.

[3] A. Bougouin, F. Boudin, and B. Daille, "TopicRank: Graph-based topic ranking for Keyphrase extraction," in *Proc. Int. Joint Conf. Natural Lang. Process.*, Nagoya, Japan, 2013, pp. 543–551.

[4] A. Ferrara and S. Salini, "Ten challenges in modeling bibliographic data for bibliometric analysis," *Scientometrics*, vol. 93, no. 3, pp. 765–785, Dec. 2012.

[5] Y. Zhang, A. L. Porter, Z. Hu, Y. Guo, and N. C. Newman, "'Term clumping' for technical intelligence: A case study on dye-sensitized solar cells," *Technol. Forecasting Social Change*, vol. 85, pp. 26–39, Jun. 2014.

[6] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[7] M. Kamkarhaghighi and M. Makrehchi, "Content tree word embedding for document representation," *Expert Syst. With Appl.*, vol. 90, pp. 241–249, Dec. 2017.

[8] S. M. Rezaeinia, R. Rahmani, H. Veisi, and A. Ghodsi, "Sentiment analysis based on improved pre-trained word embedd ings," *Expert Syst. Appl.*, vol. 117, pp. 139–147, Mar. 2019.

[9] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," in *Proc. IEEE 14th Int. Conf. Cogn. Inform. Cogn. Comput. (ICCICC)*, New York, NY, USA, Jul. 2015, pp. 136–140.

[10] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic tweets using deep convolutional neural networks," 2016, *arXiv:1610.08815*. [Online]. Available: https://arxiv.org/abs/1610.08815

[11] Y. Zhang, J. Lu, F. Liu, Q. Liu, A. Porter, H. Chen, G. Zhang, "Does deep learning help topic extraction? A Kernel K-means clustering method with word embedding," *J. Informetrics*, vol. 12, no. 4, pp. 1099–1117, Nov. 2018.

[12] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 1334, pp. 183–186, 2017.

[13] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann, 2006, pp. 600–611.

[14] J. Ghosh and A. Acharya, "Cluster ensembles," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 1, no. 4, pp. 305–315, 2011.

[15] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Inf. Process. Manage.*, vol. 53, no. 4, pp. 814–833, Jul. 2017.

[16] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. BÃűrner, "Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches," *PLoS ONE*, vol. 6, no. 3, pp. 1–11, Apr. 2011.

[17] Q. Lu, J. G. Conrad, K. Al-Kofahi, and W. Keenan, "Legal document clustering with built-in topic segmentation," in *Proc. 20th Int. Conf. Inf. Knowl. Manage.*, New York, NY, USA, 2011, pp. 383–392.

[18] W. Ding and C. Chen, "Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 10, pp. 2084–2097, Oct. 2014.

[19] A. Suominen and H. Toivanen, "Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 10, pp. 2464–2476, Oct. 2016.

[20] Y. Zhang, G. Zhang, H. Chen, A. L. Porter, D. Zhu, and J. Lu, "Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research," *Technol. Forecasting Social Change*, vol. 105, pp. 179–191, Apr. 2016.

[21] J. Yang, Z. Liu, and Z. Qu, ''Text representation based on key terms of document for text categorization,'' *Int. J. Database Theory Appl.*, vol. 9, no. 4, pp. 1–22, Apr. 2016.

[22] A. Onan, S. Korukoğlu, and H. Bulut, ''Ensemble of keyword extraction methods and classifiers in text classification,'' *Expert Syst. With Appl.*, vol. 57, pp. 232–247, Sep. 2016.

[23] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, ''Hierarchical attention networks for document classification,'' in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, New York, NY, USA, 2016, pp. 1480–1489.

[24] T. Velden, K. W. Boyack, J. Gläser, R. Koopman, A. Scharnhorst, and S. Wang, ''Comparison of topic extraction approaches and their results,'' *Scientometrics*, vol. 111, no. 2, pp. 1169–1221, May 2017.

[25] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, ''Learning word vectors for sentiment analysis,'' in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics Hum. Lang. Technol. Volume*, New York, NY, USA, 2011, pp. 142–150.

[26] R. Socher, D. Chen, C. D. Manning, and A. Ng, *Reasoning With Neural Tensor Networks for Knowledge Base Completion*. New York, NY, USA: MIT Press, 2013, pp. 926–934.

[27] Q. Le and T. Mikolov, ''Distributed representations of sentences and documents,'' in *Proc. Int. Conf. Mach. Learn.*, New York, NY, USA, 2014, pp. 1188–1196.

[28] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, ''Learning sentiment-specific word embedding for twitter sentiment classification,'' in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, New York, NY, USA, 2014, pp. 1555–1565.

[29] Y. Hong and T. Zhao, ''Automatic hilghter of lengthy legal documents,'' to be published.

[30] Y. Zhang and B. Wallace, ''A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification,'' 2016, *arXiv:1510.03820*. [Online]. Available: https://arxiv.org/abs/1510.03820

[31] C. Wei, S. Luo, X. Ma, H. Ren, J. Zhang, and L. Pan, ''Locally embedding autoencoders: A semi-supervised manifold learning approach of document representation,'' *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0146672.

[32] P. Lauren, G. Qu, F. Zhang, and A. Lendasse, ''Discriminant document embeddings with an extreme learning machine for classifying clinical narratives,'' *Neurocomputing*, vol. 277, pp. 129–138, Feb. 2018.

[33] A. M. Butnaru and R. T. Ionescu, ''From image to text classification: A novel approach based on clustering word embeddings,'' *Proc. Comput. Sci.*, vol. 112, no. 2017, pp. 1783–1792, Sep. 2017.

[34] R. T. Ionescu and A. Butnaru, ''Vector of locally-aggregated word embeddings (VLAWE): A novel document-level representation,'' in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 363–369.

[35] H. P. F. Peters and A. F. van Raan, ''Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling,'' *Res. Policy*, vol. 22, no. 1, pp. 23–45, Feb. 1993.

[36] A. Rip, *Handbook of Quantitative Studies of Science and Technology*. San Francisco, CA, USA: Elsevier, 1988, pp. 253–273.

[37] (2019). *Web of Knowledge*. [Online]. Available: https://images.webofknowledge.com/WOKRS56B5/help/WOS/hp_subject_category_terms_tasca.html

[38] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ''Efficient estimation of word representations in vector space,'' 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781

[39] J. Perkins, *Python Text Processing with NLTK 2.0 Cookbook*. New York, NY, USA: Packt Publishing, 2010, pp. 15–23.

[40] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, ''Convolutional sequence to sequence learning,'' in *Proc. 34th Int. Conf. Mach. Learn.*, New York, NY, USA, 2017, pp. 1243–1252.

[41] C. E. Moody, R. Johnson, and T. Zhang, ''Mixing Dirichlet topic models and word embeddings to make lda2vec,'' to be published.

[42] T. Boongoen and N. Iam-On, ''Cluster ensembles: A survey of approaches with recent extensions and applications,'' *Comput. Sci. Rev.*, vol. 28, pp. 1–25, May 2018.

[43] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, ''On combining classifiers,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, 1998.

[44] A. K. Jain, ''Data clustering: 50 years beyond K-means,'' *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[45] Z. Huang, ''A fast clustering algorithm to cluster very large categorical data sets in data mining,'' *DMKD*, vol. 3, no. 8, pp. 34–39, 1997.

[46] D. Arthur and S. Vassilvitskii, ''K-means++: The advantages if careful seeding,'' in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, New York, NY, USA, 2007, pp. 1027–1035.

[47] T. Kohonen, *Self-Organizing Maps*. Berlin, Germany: Springer, 2001.

[48] J. Vesanto and E. Alhoniemi, ''Clustering of the self-organizing map,'' *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, May 2000.

[49] E. Glaab, ''Analysing functional genomics data using novel ensemble, consensus and data fusion techiques,'' Ph.D. dissertation, Depart. Comput. Sci., Nottingham University, Nottingham, U.K., 2011.

[50] H. Chipman and R. Tibshirani, ''Hybrid hierarchical clustering with applications to microarray data,'' *Biostatistics*, vol. 7, no. 2, pp. 286–301, Apr. 2006.

[51] N. Nguyen and R. Caruana, ''Consensus clusterings,'' in *Proc. 7th IEEE Int. Conf. Data Mining (ICDM)*, New York, NY, USA, Oct. 2007, pp. 607–612.

[52] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, ''An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit,'' *Inf. Process. Manage.*, to be published.

[53] A. Onan, ''Consensus clustering-based undersampling approach to imbalanced learning,'' *Sci. Program.*, vol. 2019, Mar. 2019, Art. no. 5901087.

[54] R. Ünlü and P. Xanthopoulos, ''Estimating the number of clusters in a dataset via consensus clustering,'' *Expert Syst. With Appl.*, vol. 125, pp. 33–39, Jul. 2019.

[55] S. Bashir, U. Qamar, and F. H. Khan, ''Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble,'' *Qual. Quantity*, vol. 49, no. 5, pp. 2061–2076, Sep. 2015.

[56] A. Ekbal and S. Saha, ''Weighted vote-based classifier ensemble for named entity recognition: A genetic algorithm-based approach,'' *Trans. Asian Lang. Inf. Process.*, vol. 10, no. 2, p. 9, Jun. 2011.

[57] C. W. Schmidt, ''Improving a tf-idf weighted document vector embedding,'' 2019, *arXiv:1902.09875*. [Online]. Available: https://arxiv.org/abs/1902.09875

[58] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt, ''Representation learning for very short texts using weighted word embedding aggregation,'' *Pattern Recognit. Lett.*, vol. 80, pp. 150–156, Sep. 2016.

[59] H. Ren, Z. Zeng, Y. Cai, Q. Du, Q. Li, and H. Xie, ''A weighted word embedding model for text classification,'' in *Proc. Int. Conf. Database Syst. Adv. Appl.*, Germany, Cham, 2019, pp. 419–434.

[60] L. van der Maaten and G. Hinton, ''Visualizing high-dimensional data using t-SNE,'' *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

**AYTUĞ ONAN** was born in İzmir, Turkey, in 1987. He received the B.S. degree in computer engineering from the Izmir University of Economics, Turkey, in 2010, and the M.S. degree in computer engineering and the Ph.D. degree in computer engineering from Ege University, Turkey, in 2013 and 2016, respectively. He has been an Associate Professor with the Department of Computer Engineering, Izmir Katip Celebi University, Turkey, since April 2019. He has published several journal articles on machine learning and computational linguistics. Dr. Onan has been reviewing for several international journals, including *Expert Systems with Applications*, *Plos One*, the *International Journal of Machine Learning and Cybernetics*, the *Journal of Information Science*.