Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.1967

Research Article

# Clustering business news for strategic insight: An unsupervised learning approach

**Gaurav V. DAHAKE[1,*], M. S. ALI[1]**

[1]*Department of Computer Science and Engineering, Prof. Ram Meghe College of Engineering and Management, Amravati, 01128, India*

## ARTICLE INFO

*Article history*
Received: 03 April 2024
Revised: 18 July 2024
Accepted: 11 November 2024

**Keywords:**
BERT; Business News; Clustering Evaluation; Information Overload; K-Means; Text Clustering; Unsupervised Learning

## ABSTRACT

Business news assists as an important tool which enables business leaders and marketers to create strategic plans through their decision-making process. The large number of news articles spread across different platforms creates difficulties when trying to find important information through efficient methods. The research presents an approach which employs unsupervised learning methods to cluster business news articles for solving information overload problems. The system uses K-Means and Agglomerative clustering algorithms together with Bag-of-Words and TF-IDF and BERT embeddings to identify articles with similar content. The research evaluates clustering performance through various assessment methods which include Elbow Score and Silhouette Score and Calinski-Harabasz Index and Davies-Bouldin Index to determine which approach works best. The combination of BERT with K-Means produces better results than TF-IDF because it reaches an accuracy level of 84.46%. The preprocessing stage required tokenization and stop word removal and stemming to solve the problems which occurred when dealing with noisy data and unimportant information. The results show that BERT achieves better clustering results because it understands deep semantic meanings in text data. The research investigates both scalability and practical implementation of the system through its proposed studies about real-time system flexibility and data bias ethical problems. The proposed solution improves users' ability to access business news which enables them to make better decisions.

**Cite this article as:** Dahake GV, Ali MS. Clustering business news for strategic insight: An unsupervised learning approach. Sigma J Eng Nat Sci 2026;44(1):55–73.

## INTRODUCTION

The fast expansion of digital news platforms together with the large quantity of news content has created major difficulties for news organization and valuable information retrieval from this extensive information system. The process of grouping similar news articles has proven useful for solving these problems [1]. The system uses information retrieval and natural language processing to organize news articles through their semantic and thematic and contextual connections. The goal is to group articles with everyday topics, themes, or characteristics, enabling users to discover relevant content more efficiently [2,3]. The analysis of news

***Corresponding author.**
*E-mail address: gauravdahake.77@gmail.com

*This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic*

article text content through information retrieval and natural language processing methods in clustering techniques helps researchers detect similar patterns in the data. Users can discover related content and hidden relationships and gain complete knowledge about various subjects through the process of clustering news articles [4]. The process of clustering news articles with similar content helps users organize information while recommending content and identifying trends which results in better news delivery and improved digital news access efficiency [5].

The process of clustering news articles which share similarities requires multiple solutions to achieve successful results. News articles contain language which remains ambiguous and subjective because this creates different ways for readers to understand what makes things similar. The process of article clustering becomes difficult because articles contain both background noise and unimportant data. The system needs to handle increasing news article numbers because its scalability becomes a problem which demands better algorithm performance [6]. The process of feature extraction and similarity measurement becomes more complicated because of different writing styles and language choices which appear in textual variations. News requires immediate processing because of its time-sensitive nature which makes temporal analysis essential for achieving proper clustering results [7]. The evaluation of clustering quality and determination of cluster numbers remains difficult while researchers face challenges when dealing with overlapping topics and multiple cluster assignments for articles [8]. The solution of these problems requires sophisticated methods for data preprocessing and feature extraction and similarity assessment and evaluation which produce better clustering results for news articles with similar content.

The current clustering methods lack sufficient power to detect semantic relationships between words because they use Bag-of-Words and TF-IDF approaches. The methods fail to handle high-dimensional data while their ability to process large datasets including continuously updated news articles remains limited. The current clustering methods fail to consider how news topics change over time which results in clusters that become irrelevant. The current methods fail to process news data which contains noisy information together with missing or unimportant content that appears in actual news datasets. Research studies depend on restricted evaluation criteria which produce restricted assessments of clustering performance. The current methods lack practical use in real-world settings which creates challenges for evaluating their performance in active operational environments. The research fills these knowledge gaps through its implementation of BERT embedding methods and its use of strong text processing techniques and its detailed assessment protocol which enhances its ability to work in practical scenarios.

The Paper contains its primary contribution which follows.

The system provides an organized framework to process large news datasets which enables users to find relevant information and discover important knowledge from their data. Clustering facilitates effective content organization and trend analysis by grouping similar articles together. It assists in identifying hidden patterns, emerging topics, and relationships within news articles, allowing users to navigate the news landscape more effectively.

- In order to develop clustering algorithms the following can be used: K-mean and Agglomerative, these integrated with Bag-Of-Words, Word2Vec, TF-IDF and BERT to generate data embeddings.
- The choice of evaluation metric such as Silhouette Score, Davies-Bouldin Index, Calinski Harabasz Index and Elbow method plays a crucial role in refining and improving clustering techniques for news articles that are similar.

The organization of the Paper is as follows: Section 2 presents the different approaches and techniques for clustering similar news articles. Section 3 presented the complete methodology for clustering identical news articles. The final section of this study presents both result analysis and establishes the research findings and future research possibilities.

**Related Work**

The field of text clustering involves various algorithms for grouping similar texts together. Among the most commonly used algorithms are the vector space model, k-means clustering plus its derivatives, techniques based on generation and on spectra, methods to reduce dimensionality and lastly, methods based on phrases [1]. This traditional approach, vector space model, is well suited to problems that have similar themes however the model assumes the number of clusters beforehand [2]. K-means clustering can be used for hierarchical clustering and partitioned clustering [3]. However, this kind of clustering has a few significant limitations. These include a high computational demand for large data sets, requirement for random initialisation and also the sensitivity to outliers present in the data. Another approach for clustering fake news based on the k-mean algorithm [9]. Advanced deep learning techniques such as CNN and BERT are being utilised for clustering based fake news detection [10]. The performance of generative algorithms decreases when working with different types of data and they need to know the number of clusters in advance [5]. The algorithm of spectral clustering achieves its best results when data exists as a bipartite graph because it lacks the need for cluster number specification [4-6]. The original computer vision methods for dimensionality reduction achieve high performance but they need random starting points which produce different results when analyzing the same data [7,8]. Phrase-based methods encode word order but don't guarantee higher accuracy than other clustering methods [9,10]. The Agglomerative were proposed for clustering short texts and news articles [11]. For instance,

discriminant topic models and social network analysis on Twitter [12,13]. Special kernel functions for semantic similarity and collective clustering methods have also shown promise in improving clustering accuracy for short texts and news [14,15]. RELEVANTNews, a web feed reader by Sonia Bergamaschi et al.[16], organizes news on similar topics from different newspapers based on syntactic and lexical similarities. However, many aggregators with similar functionalities are commercial products with proprietary systems [17]. Text clustering has become a crucial approach used to manage the enormous quantity of textual data that exists in the healthcare sector which holds important health-related information. In recent years, there has been a growing interest in the fields of natural language processing to the area of healthcare due to significant advancements in emerging deep learning approaches [18]. A combined optimal strategy utilizing K-means has been proposed that help doctors properly aggregate healthcare information connected to heart problems and discover the optimal remedy [18]. A cloud-edge computation architecture is the foundation for the C-means method established, allowing for the gathering of health information from several institutions [19]. Numerous techniques have been put out thus far for optimizing the cluster design [20]. These fall within the categories of neutral and biased techniques. Most approaches used to address the atomic framework optimization involve variations, namely, some methodologies, such as static lattice search [21]. These can make cluster optimization easier. However, their usefulness is severely limited because they work only with certain clusters and are hard to apply to other optimized feature situations. The Rock algorithm was used to optimizing the data cluster based on k-means [22,23]. In evaluating performance, hyperparameter optimization (HPO) acts as a valid process in and of itself. In contrast to the very substantial scientific work on the suitable assessment of supervised using, however, there is a lesser understanding regarding the appropriate assessment of cluster techniques in combination with HPO [24].

These clustering techniques fall into three broad categories, each with specific characteristics and applications.

1. Straightforward per user gives a graphical interface to imagining and gathering RSS channels from various papers. Essential capacities supporting the client in perusing are shown (for example, web crawler, distinctive requesting, a relationship of information to a guide);
2. News classifiers show the news arranged based on standards now and then chosen by the client. Straightforward characterizations might take advantage of the classes or potentially the watchwords given by the sites;
3. Progressed aggregators give extra elements to support the client in perusing, grouping, arranging, and putting away news.

Various studies have explored different approaches to organizing and presenting news. Velthune, introduced in [25], is a news search engine that uses a simple classifier to categorize information. In contrast, RELEVANTNews focuses on clustering similar information based on their content. This clustering method can create large newsgroups in the same category, making navigating harder for readers. RCS (RSS Clusgator device) proposed in [26] updates clusters of news over time, aiming to maintain the relevance of information. NewsInEssence [27], another aggregator, uses the TF*IDF clustering algorithm to group similar news and provide summarized content for readers. However, RELEVANTNews doesn't offer summaries but utilizes a customizable clustering algorithm based on syntax, words, and primary relationships to improve clustering accuracy.

The authors describe a system based on Matrix-based News Aggregation (MNA) which operates through a five-stage processing pipeline in their research [28]. The system starts by gathering data from web sources which gets stored before it moves to article classification and content reduction and visual presentation for user consumption. The main MNA methodology uses a structured matrix format to organize information where news entities serve as matrix rows and their attributes function as matrix columns. The system enables users to personalize their search preferences which become the starting point for the matrix. The process of summarization comprises several tasks such as summarization based on a topic and TF-IDF summarization of the content within each cell of the matrix.

These works illustrate various techniques used in clustering similar news articles, utilizing different techniques such as topic modeling, entity-based attributes, graph ranking, embedding techniques, and personal clustering. These works help in improving clustering techniques used in the task of clustering similar news articles, as they provide effective means of dealing with challenges of clustering news.

**Proposed Methodology**

This study employs a methodology to cluster similar news articles by combining text embedding techniques with K-Means and Agglomerative clustering. Initially, textual data is converted into numerical feature vectors using four representations: Bag-of-Words (BoW), which counts word frequencies; TF-IDF, which weights word importance; Word2Vec, which captures semantic relationships; and BERT, which generates context-aware embeddings. The embeddings provide a vector representation that captures the semantic meaning of each article. During the next stage the algorithm makes clusters by successively placing each story into the cluster whose centroid is closest according to a distance function such as Euclidean distance, thereby aiming to reduce the within cluster sum of squares. The iterations are continued until a stopping criterion is met. The hierarchical clustering algorithm employed is agglomerative clustering, where the procedure starts with each item of information as a separate cluster and then merges those clusters which are closest to each other according

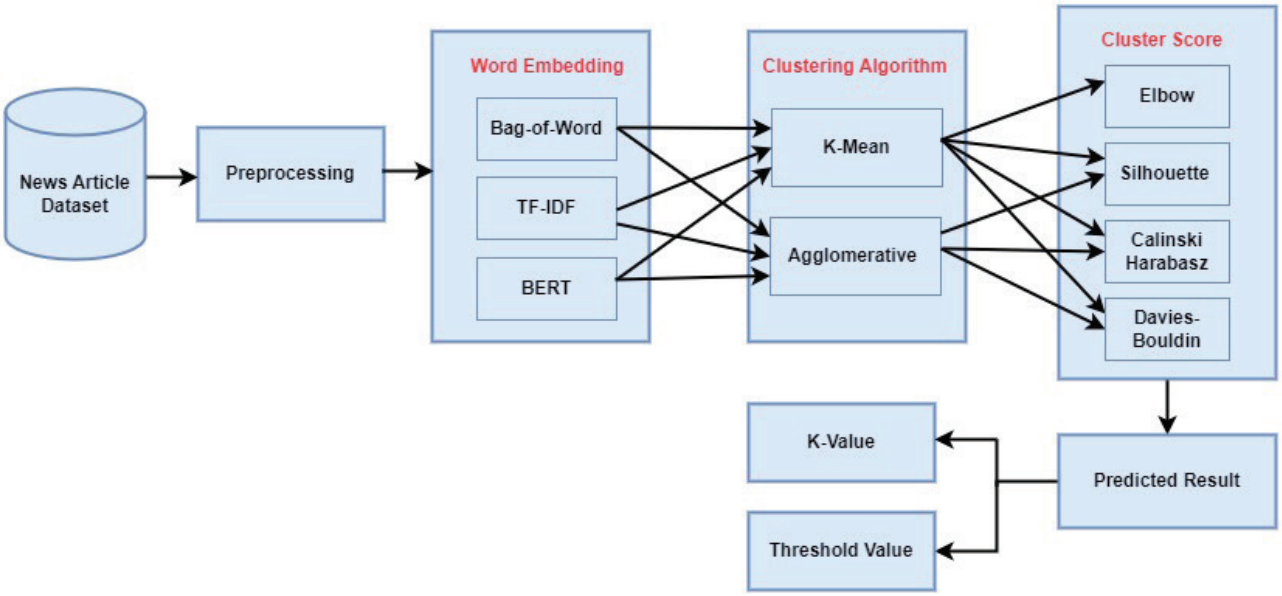


**Figure 1.** Proposed Architecture of clustering similar news articles.

to a defined similarity measure. The choice of appropriate clustering validation tool depends on data type. Techniques like silhouette, Calinski Harabasz, Davies-Bouldin index and elbow method can be used to decide on the optimal number of clusters. Through the use of these clustering algorithms and the appropriate application of embeddings, the proposed methodology seeks to efficiently cluster news articles based on content and facilitate an understanding of the relevant groupings of information.

**Overview of Dataset**

The dataset used in this study is the News Aggregator Dataset from Kaggle, a well-established resource for text analysis and machine learning research. This dataset encompasses many articles sourced from diverse publishers and covers a wide range of topics, including politics, technology, entertainment, and sports. The dataset contains 422,937 news stories or articles. Each item within the database includes detailed information, for instance publication date, news agency, the category the news story belongs to and title. This combination of large scale, topical variety, and rich metadata provides a robust foundation for developing and evaluating our clustering methodology, enabling rigorous testing of the algorithms' ability to organize and categorize real-world news content.

**Data Preprocessing**

At the beginning of the processing stage of a program which gathered news articles a number of problems were encountered. The presence of these elements, though, led to a loss of semantic coherence and a lessening of the semantic importance of the words used in the text, which resulted in a clustering accuracy that was lower than expected. In order to tackle these problems, various natural language processing techniques were utilised. The text was initially split into individual words by means of tokenisation, which at the same time removed punctuation from the text. By excluding articles such as 'the', 'a' and 'in', word filtering removes words which generally do not add much to the meaning of the document during the process of clustering. Stemming further reduced words to their root forms, minimizing variations of the same word, and Case Transformation standardized the text by converting all characters to lowercase. The NLTK tools were used to apply a series of text pre-processing steps which cleaned the data, improved its relevance and increased the accuracy of the clustering algorithms.

**Embedding Techniques**

This study employed three-word embedding techniques: Bag-of-Words, TF-IDF, and BERT. Each method was integrated with two clustering algorithms, namely K-means and Agglomerative, to cluster similar news articles.

**Bag of words**

A bag-of-words model is a method for extracting the relevant features from a similar news article dataset for modeling, such as with proposed classifiers. The approach is very simple and flexible and can be used in a myriad of ways [29].

To create a mathematical model using the Bag of Words (BoW) approach for clustering similar news articles,

Create a vocabulary $V$ by collecting all unique terms from the preprocessed articles.

$$V = [n(w_{1,d})\, n(w_{2,d})\; n(w_{3,d})\; \ldots . n(w_{T,d})] \quad (1)$$

$N(.)$ is the number of news articles occurring in datasets

The similarity between the two articles is measure using distance based on cosine similarity

$$d(p_i, q_j) = \cos\theta \tag{2}$$

$$= \frac{p_i \cdot q_j}{||p_i||\,||q_j||} \tag{3}$$

**Term Frequency-Inverse Document Frequency (TF-IDF)**

Term frequency-ınverse document frequency (TF-IDF) is a commonly used statistical technique in natural language processing and information retrieval. It assesses the similarity of an article compared to a set of news articles. TF-IDF turns an article into a vector by multiplying its TF with the Inverse Document Frequency (IDF). This calculation helps determine how relevant a news article is to a dataset.

To establish a mathematical model using TF-IDF for clustering similar news articles

$$V = [n(w_{1,d})\ n(w_{2,d})\ n(w_{3,d})\ \ldots . n(w_{T,d})] \tag{4}$$

all the articles are not equally created

$$V_d = [(w_{1,d}) \propto_1 n(w_{2,d}) \propto_2 n(w_{3,d}) \propto_3 \ldots . n(w_{T,d}) \propto_T] \tag{5}$$

$$n(w_{i,d}) \propto_i = n(w_{i,d}) \log \left\{\frac{D}{\sum_{d'} 1[w_i \in d']}\right\} \tag{6}$$

**Word2Vec**

Word2Vec creates vectors of the article that are distributed numerical representations of news article features; these features could comprise articles that represent the context of the individual news articles present in datasets [30]. Word embeddings eventually help establish the association of news articles with other similar-meaning news articles through the created vectors.

To create a mathematical model using Word2Vec for clustering similar news articles, Measure the position x of each article and define the center news article at that position as c and the context article *o*. To recognize article, define window size *w*, denotes the proposed model will used as articles in position *t* - *w to t* + *w* as the context. Once the measure of all the context article at position *t*, to maximize the context article given the centroid article, measure the probability of the proposed model predicting the context news articlesbased on center article. The following equation shows to maximize the probability.

$$L(\emptyset) = \prod_{t=1}^{T} \prod_{\substack{-w \le j \le w \\ j \neq 0}} P(M_{t+j} | M_t; \emptyset) \tag{6}$$

Convert equation 6 into derivations and make it maximization problem, take the log and multiply by -1

$$J(\emptyset) = -\frac{1}{T} log L(\emptyset) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-w \le j \le w \\ j \neq 0}} \log P(M_{t+j} | M_t; \emptyset) \tag{7}$$

Measure the probability of context article. Denote each article by two sets of vectors, *Pm* and *Qm*. If m is context article, then use *Pm* and if *m* is centre article, then use *Qm*. according to this vector, probability equation for center article o and context article *c* as shown in equation 8

$$P\ (O = o | C = c) = \frac{\exp\ (p_o{}^T q_c)}{\sum_{m \in Dataset D} \exp\ (p_m{}^T q_c)} \tag{8}$$

Equation 8 retrieves the similarity of two vectors o and c. More similarity, more probability.

$$\emptyset = \begin{bmatrix} Qm1 \\ Qm2 \\ Qm3 \\ . \\ . \\ Qmn \\ Pm1 \\ Pm2 \\ Pm3 \\ . \\ . \\ Pmn \end{bmatrix} \in R^{2dV} \tag{9}$$

Equation 9 represents the vector that consists of both p and q vectors of d length for the articles.

**BERT**

This study uses BERT for word embedding to cluster similar articles. BERT embeddings capture contextual and semantic information of words and phrases, which highly benefits clustering similar news articles. By representing articles as vectors of BERT embeddings, measure the semantic similarity between articles. This enables the clustering of articles based on their content and context [31].

Figure 2 shows the representation of the BERT model to generate the context-based data embedding. This model is trained on a massive news article dataset of the English language. BERT's depth (12 layers), hidden size (512), and multi-head self-attention (12 heads) are vital components that allow it to capture rich contextual information from similar article data. In this research, information is taken from the output of the second-to-last layer of a transformer model, specifically the 11th layer. This information is turned into a form that looks like a tensor with measurements (n, 25, 511). Where 'n' stands for the number of articles in the dataset, '25' is how many words or parts are in each article, and '511' shows the size of the hidden information. Later, methods for extracting features and making the data consistent will be used on this information, leading to a matrix with measurements of (n, 511).

Utilizing BERT to cluster similar news articles involves the following mathematical model:

For the Articles *Ai* shows the sequence of words *w* = [*w*1, *w*2, *w*3.... *wn*], the article embedding *E*(*A*) can be computed as:

$$E(Ai) = BERT([w1, w2, w3 \ldots . wn]) \tag{10}$$

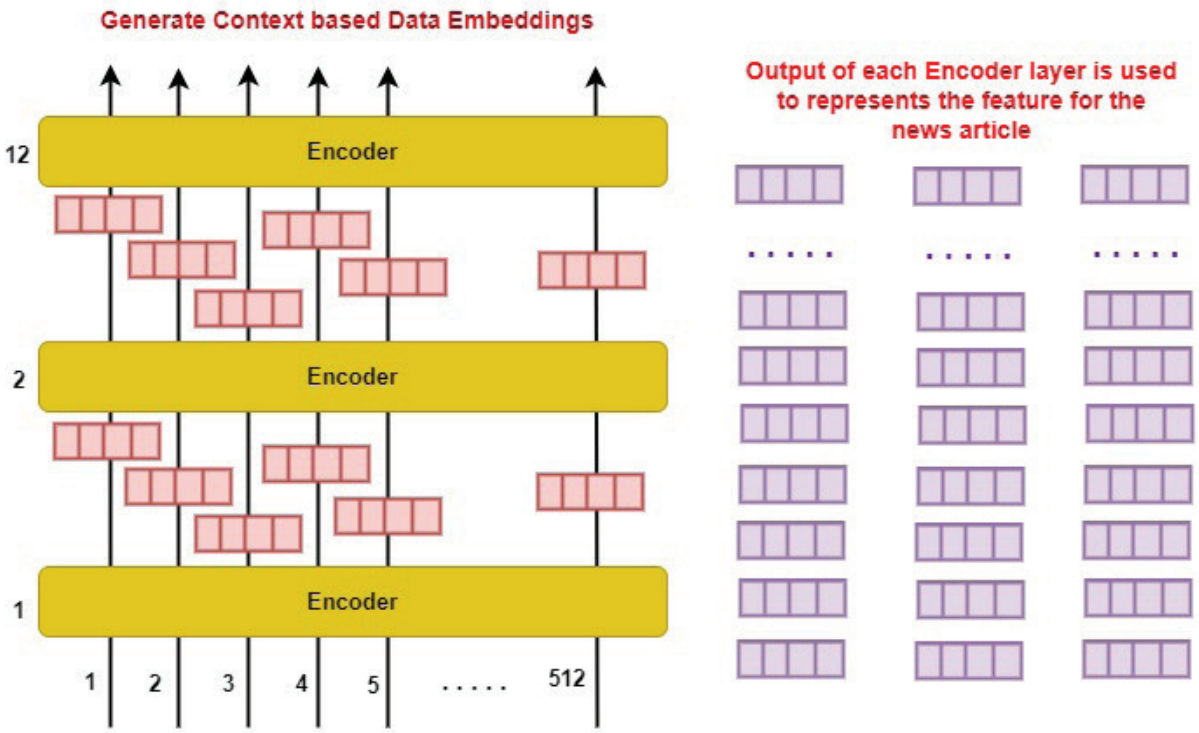


**Figure 2.** Representation of BERT for context-based data embedding.

The Max pooling layer cab be showed by further equation 11

$$a[k] = \max_{l=1,2....,n} w_{ik} \tag{11}$$

where, $w[k]$ is the *k-th* entry of similar articles data of vector w. The complete articles can represent all the contextual information vectors, and taking the mean of these vectors removes noise in the data, which is why mean pooling is considered. Equation 12 illustrates the mean pooling approach.

$$w[k] = \frac{\Sigma_{i=1}^{n} w_{ik}}{n} \tag{12}$$

$f(w) = w$ represents the identity normalization, which is used as an identity function. For the vector ai, standard normalization involves applying the function as described in the equation 13.

$$\overline{w} = \frac{w_i}{\|w_i\|} \tag{13}$$

Where, $\overline{w}$ represents the normalized vector. it converts vector into norm of 1. The layer normalization approach can minimize the covariate shift issue, during the training the model [31], equation shows the function in layer.

$$\overline{w} = \frac{w_i - \theta_i}{\sigma_i} \tag{14}$$

Where, $\theta_i$ and $\sigma_i$ represents the mean and SD of the vector $w_i$. min-max normalization approach is used to maintain the primary distribution of the similar article data vectors. It utilizes the function described in the equation 15.

$$\overline{w} = \frac{w_i - min_d(w_{id})}{max_d(w_{id}) - min_d(w_{id})} \tag{15}$$

BERT processes the entire sequence and retrieves the contextual data and semantics of the articles. With article embedding obtained from BERT, similarity between articles is computed using various similarity measures based on Cosine similarity techniques. Example two articles embedding $E(Ai)$ and $E(Aj)$, the cosine similarity $S(Ai,Aj)$ is calculated as:

$$S(Ai, Aj) = \frac{E(Ai).E(Aj)}{||E(Ai)||.||E(Aj)||} \tag{16}$$

The dot (.) is used to measure the similarity the vectors, and the Euclidean norm (magnitude) is used to normalize the vectors.

**Functioning of Clustering Techniques**

Clustering methods used include K-Means and Agglomerative hierarchical clustering, due to their suitability for handling large numbers of features in each data point, such as those found in text vector representations. The K-means algorithm works well with very large collections and also performs well with techniques such as BERT and TF-IDF feature

extraction when the number of groups is known. In contrast, agglomerative clustering enables hierarchical cluster structuring which facilitates the identification of connections between documents that exhibit varying degrees of similarity. In such instances where these methods may not be optimal is when dealing with non-spherical cluster forms.

**K-Means Clustering**

Clustering analysis using the k-means method groups together a collection of data points, without prior knowledge, into K clusters. Randomly select k cluster centers from the data, which are often referred to as the centroids of the cluster. The centroids used are actually randomly chosen points from the given set of data. They are used to train a classifier with these centroids. This classifier then assigns data points to clusters, initially creating random clusters [33]. Afterward, each centroid is updated to the average position of the cluster it represents. This classification and centroid update process is repeated until the centroids' positions stop changing. As a result, the vectors lying in this space finally converge to stable centroids, which in turn classify the data points into groups or classes. This classification gives anonymous data points a particular identity and assigns them to specific clusters. In order to establish a mathematical model for K-means clustering for clustering similar news articles, we can define the equation 17 (Table 1):

$$A = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \quad (17)$$

$r_{nk}$ is belongs to (0,1) represents the membership value of data $x_n$ in cluster K. To minimizes the objective function A need to measure the specific value for $r_{nk}$ and $\mu_k$. The objective function A can be minimized by set the values of $r_{nk}$ and $\mu_k$ based on the equation 18

$$r_{nk} = \begin{cases} 1 & k = arg\ min_k ||x_n - \mu_k||^2 \\ 0, Others \end{cases} \quad (18)$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}} \quad (19)$$

**Agglomerative Clustering**

News articles are clustered based on their similarity in vector space, using agglomerative hierarchical clustering. This method uses a measure of distance such as the Euclidean distance. The program provides a comprehensive view of the way clusters are formed. Through this, it is possible to establish a detailed level of granularity in the clustering which the algorithm produces. This makes it easier to discover which news items are grouped together in the hierarchy created by the Table 2.

To create a mathematical model for Agglomerative Clustering to cluster similar news articles and define the following:

The agglomerative clustering techniques applied to non-vector data

Consider $A = x1, x2, ..... xn$

$A$ is the article dataset

Now updated the predefined threshold of distance. There are two cluster $p$ and $q$ are combined to each other and create new cluster $r$. Consider $k$ represents the other existing cluster. If measured the distance between cluster $q$ and cluster $k$. $p$ and $q$ are the existed, the distance between $p$ and $k$ and the distance between $q$ and $k$. Represents the distance by $D(p, k)$ and $D(q, k)$

The cluster is the single link cluster

$$D(r, k) = min(D(r, k), D(q, k)) \quad (20)$$

$D(r, k)$ represents the minimum distance between the articles in cluster r and k resp. It can be denoted the equation 20 update the distance is similar to minimum distance between two articles over the two cluster.

Complete cluster is

$$D(r, k) = max(D(r, k), D(s, k)) \quad (21)$$

where, $D(r, k)$ represents the max distance between two articles in cluster $r$ and $k$

The two distances, $D(p, k)$ and $D(q, k)$ are aggregated with the weighted sum

$$D(p, k) = \frac{1}{2} D(p, k) + \frac{1}{2} D(q, k) \quad (22)$$

The centroid measure for each cluster article and the distance between article clusters is represented as the distance between their centroids.

$$D(p, k) = \frac{1}{2} D(p, k) + \frac{1}{2} D(q, k) - \frac{1}{4} D(p, q) \quad (23)$$

**Table 1.** K-means clustering

**Step 1:** Initialize $K$ cluster centroids, $\mu = \{\mu 1, \mu 2, .., \mu K\}$, randomly.

**Step 2:** Assign each news article $xi$ in $X$ to the nearest centroid:

$zi = argmin(||xn - \mu k||)$, for $k = 1$ to $K$, using a distance metric such as cosine similarity.

**Step 3:** Update the centroids by computing the mean of the vectors in each cluster:

$\mu k = (1/|Ck|) * \Sigma xi$, for $xi$ in $Ck$.

**Step 4:** Repeat the assignment and update steps until convergence or a predefined stopping criterion is met. Convergence is achieved when the centroids no longer change significantly or when the maximum number of iterations is reached.

**Table 2.** Agglomerative clustering

**Step 1:** Initialize each news article *xi* as a separate cluster.

**Step 2:** While the number of clusters is greater than 1

- Find the pair of clusters with the minimum distance, i.e., the closest pair of clusters in the distance
- Matrix *D*.
- Combine the two neighbour clusters into a single cluster.
- Modify the distance *D* by recalculating the distances between the new cluster and the remaining clusters using a linkage criterion such as complete linkage.

**Step 3:** Assign each news article to its corresponding cluster based on the final clustering structure.

The following equation 24 can update the distance

$$D(r,k) = \frac{n_p+n_k}{n_p+n_q+n_k} D(p,k) + \frac{n_q+n_k}{n_p+n_q+n_k} D(q,k) - \frac{n_k}{n_p+n_q+n_k} D(p,q) \quad (24)$$

**Hyperparameter Setting**

When performing cluster analysis on similar news stories using K-means and agglomerative clustering, parameter tuning is required. The choice of the number of clusters, the way of initializing cluster centroids and the maximum number of iterations relate to the k-means clustering algorithm. Three metrics can be used to determine the number of clusters for k-means: the Silhouette method, the Davies-Bouldin method and the elbow method. The k-means algorithm can be enhanced by using the 'k-means++' algorithm for initialization to improve the clustering's stability and efficiency of convergence. The number of iterations (max_iter) should be set to 300, which ensures a sufficient number of iterations for convergence. Agglomerative clustering can be optimised by adjusting three parameters: the number of clusters, the average linkage method and the metric used to calculate the distance between clusters. In general, grid search with cross-validation was employed to find the best parameters of the algorithm so as to improve the cluster's accuracy and quality.

**Parameter Evaluation**

To obtain the most accurate cluster results in grouping similar news articles together, parameter settings need to be optimised. Some key parameters to consider when evaluating the effectiveness of clustering techniques:

**Elbow Score:** A heuristic to determine the optimal number of clusters. Generally, the cluster radius is obtained from the within cluster sum of squares, calculated as follows:

$$WCSS = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - c_j)^2 \quad (25)$$

Where, *K* represents the number of clusters to measure, n represents the number similar articles, $x_{ij}$ denotes the *jth* similar articles in the ith cluster, $c_j$ is the centroid of the *ith* cluster

**Silhouette Score:** The Silhouette Score is a tool that checks how healthy clusters created by a clustering method work. It determines how much an article matches its cluster compared to the other clusters. To find the Silhouette Score for a particular article, a specific formula is used:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i),b(i)\}} \quad (26)$$

Where, $S(i)$ is the Silhouette Score for similar news articles point $i$, $a(i)$ represents the average distance of similar news articles $i$ to all other articles in the similar cluster, $b(i)$ is the smallest average distance of similar news articles $i$ to all news articles in other cluster

**Calinski-Harabasz (CH):** The Calinski-Harabasz (CH) score, also called the Variance Ratio Criterion, is used to judge how good clusters are when created by a clustering method. It calculates the ratio between how different clusters are from each other compared to how similar data points are within the clusters. A higher CH score means more apparent and more distinct clusters. The formula for the Calinski-Harabasz score is used to calculate this:

$$CH = \frac{B(K)}{W(K)} \times \frac{N-K}{K-1} \quad (27)$$

Where, *CH* is the Calinski-Harabasz score, *K* is the number of clusters, *N* is the total number of similar news articles. $B(K)$ is the between-cluster variance, which is calculated as the sum of squared distances between cluster centroids and the overall mean of the similar news articles, weighted by the number of articles in each cluster. $W(K)$ is the within-cluster variance, which is the sum of squared distances between similar news articles and their respective cluster centroids.

**Davies-Bouldin:** The Davies-Bouldin Index is a tool used to judge the quality of clusters made by a clustering method. It looks at how similar each cluster is to its most similar one while also considering the size of the clusters. The formula for calculating the Davies-Bouldin Index is utilized for this purpose:

$$DB = \frac{1}{K} \sum_{i=1}^{K} max_{j \neq i} \left( \frac{S_i+S_j}{a(c_i,c_j)} \right) \quad (28)$$

Each evaluation metric provides a different perspective on clustering quality. The Elbow method for determining the optimal number of clusters, while the Silhouette Score evaluates how well-separated the clusters are. Calinski-Harabasz and Davies-Bouldin indexes measure the internal structure of clusters, with the former focusing on the variance ratio and the latter assessing cluster compactness.

## RESULT AND DISCUSSION

### Clustering Result of K-Mean

In this phase, we employ various combinations of embedding techniques and clustering methods to group news articles. To assess these combinations, we used a collection of 3811 news articles. Techniques such as Bag-Of-Words, TF-IDF, Word2Vec, and BERT embeddings were applied to these articles. Subsequently, we employed K-means and Agglomerative clustering methods on the embedded news articles. The elbow method was utilized to determine the optimal value for K in K-means clustering.

Figure 3(a) shows the inertia of clusters from 15 to 30. The Bag-Of-Words technique was used to embed text into vectors. It can be observed from Figure 3 that there are considerable variations in inertia as try clusters from 15 to 25,

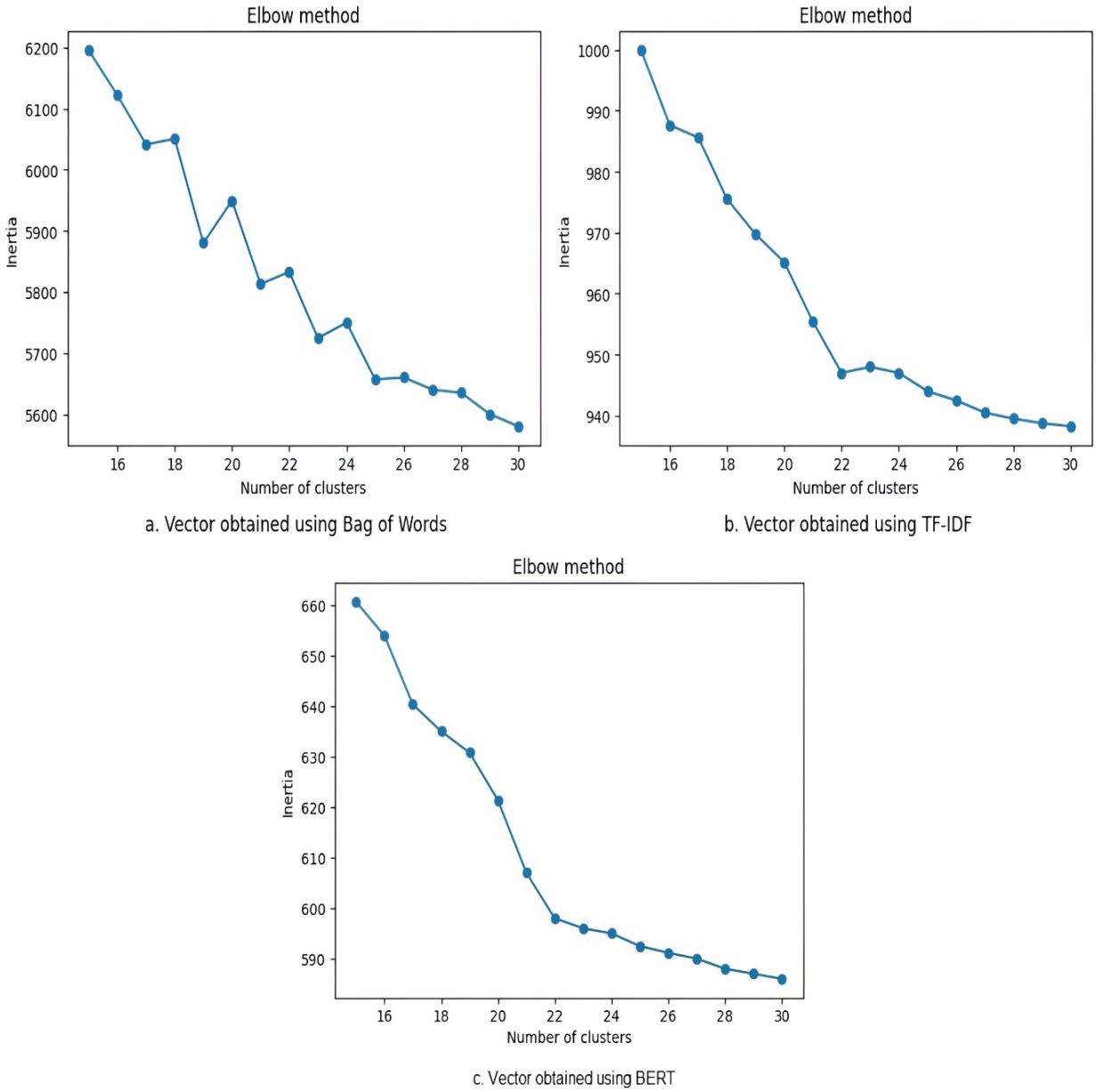


**Figure 3.** Vector obtained using Elbow method (a) Bag-Of-Words (b) TF-IDF (c) BERT of K-Mean.

and after 25 shows that inertia does not change much. So, according to Figure 3(a), the value of K can be taken as 25.

Figure 3(b) shows the inertia of clusters from 15 to 30. The TF-IDF technique was used to embed text into vectors. From Figure 3(b), the elbow shape is formed at 22 clusters. So, according to Figure 3(b), the value of K is 22.

Figure 3(c) shows the inertia of clusters from 15 to 30. The BERT technique was used to embed text into vectors. From Figure 3(c), it can be seen clearly that the elbow shape is formed at 22 clusters. According to Figure 3(c), the value of K is 22.

Figure 4 (a) shows the Silhouette Score of clusters from 15 to 30. The Bag-Of-Words technique was used to embed text into vectors. According to Figure 4 (a), the value of K is 22.

Figure 4 (b) shows the Silhouette Score of clusters from 15 to 30. The TF-IDF technique was used to embed text into vectors. According to Figure 4 (b), the value of K is 22.

Figure 4 (c) shows the Silhouette Score of clusters from 15 to 30. The BERT technique was used to embed text into vectors. According to Figure 4 (c), the value of K is 23.

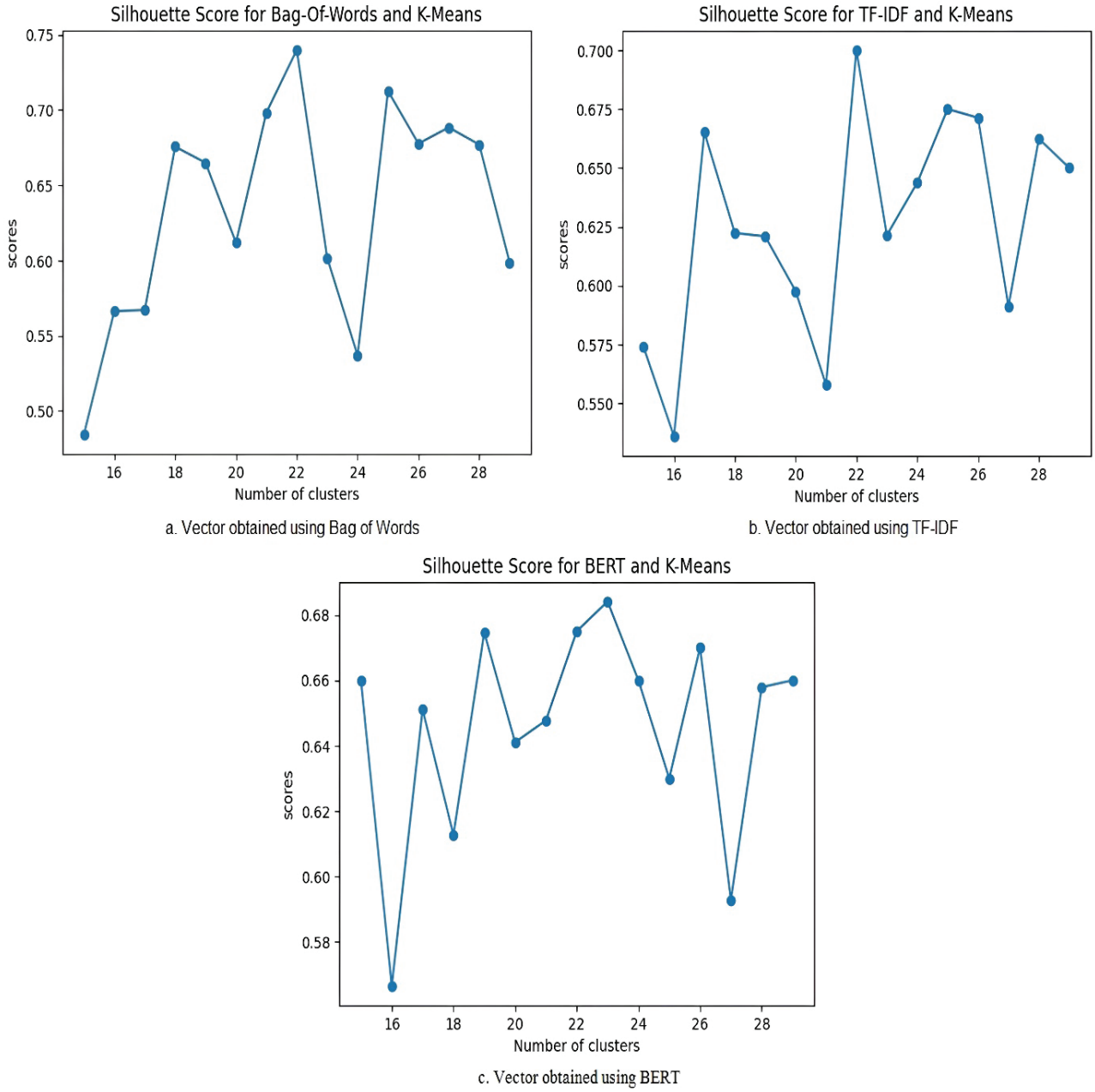


**Figure 4.** Vector obtained using Silhouette Score (a) Bag-Of-Words (b) TF_IDF (c) BERT of K-mean.

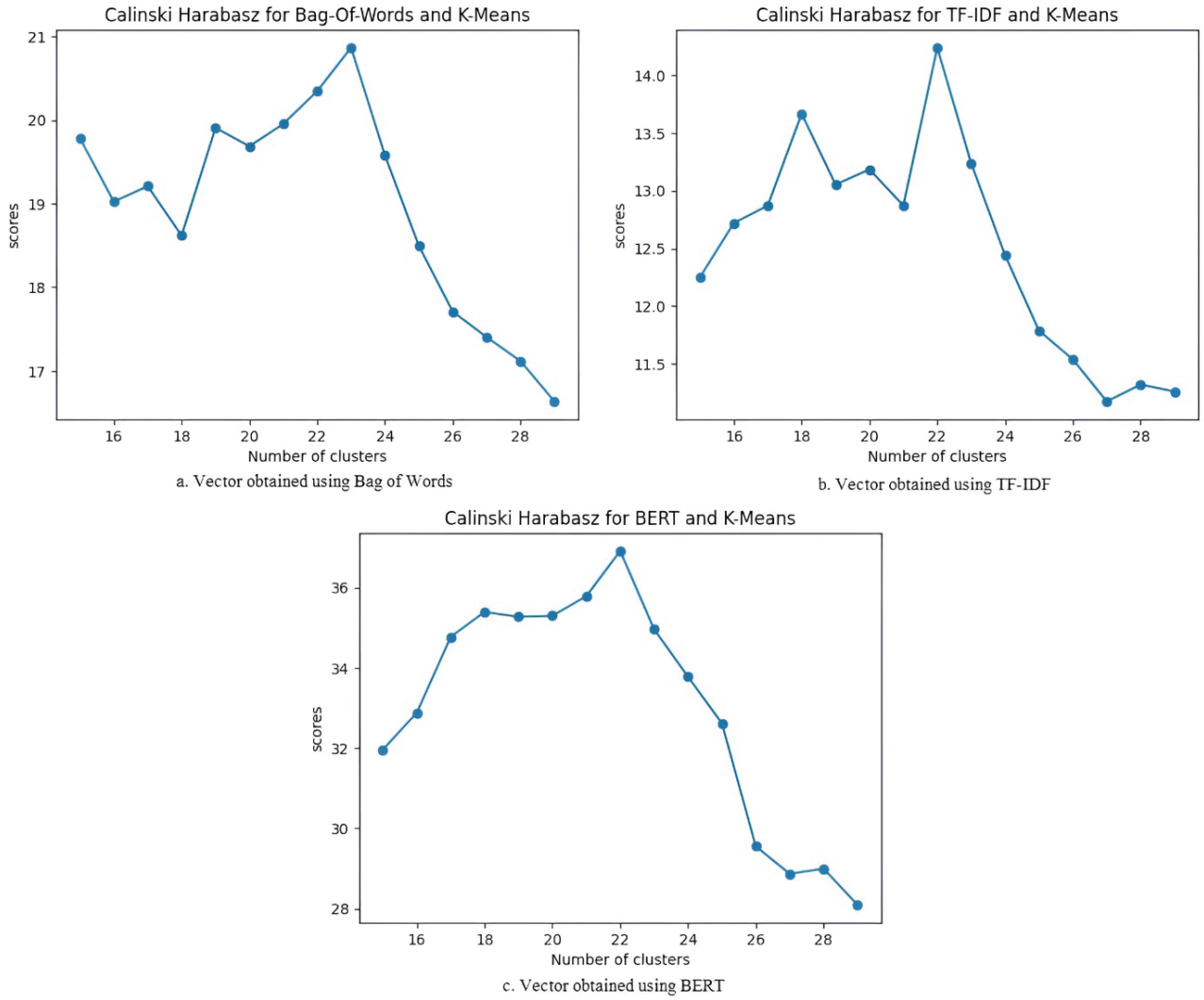


**Figure 5.** Vector obtained using Calinski Harabasz (a) Bag-Of-Words (b) TF-IDF (c) BERT of K-mean.

Figure 5 (a) shows the Calinski Harabasz Score of clusters from 15 to 30. The Bag-Of-Words technique was used to embed text into vectors. According to Figure 5 (a), the value of K is 23.

Figure 5 (b) shows the Calinski Harabasz of clusters from 15 to 30. The TF-IDF technique was used to embed text into vectors. According to Figure 5 (b), the value of K is 22.

Figure 5 (c) shows the Calinski Harabasz of clusters from 15 to 30. The BERT technique was used to embed text into vectors. According to Figure 5 (c), the value of K is 22.

Figure 6(a) shows the Davies-Bouldin Score of clusters from 15 to 30. The Bag-Of-Words technique was used to embed text into vectors. According to Figure 6(b), the value of K is 24.

Figure 6(b) shows the Davies-Bouldin of clusters from 15 to 30. The TF-IDF technique was used to embed text into vectors. According Figure 13, the value of K is 22.

Figure 6(c) shows the Davies-Bouldin of clusters from 15 to 30. The BERT technique was used to embed text into vectors. According to Figure 6(c), the value of K is 22.

**Clustering Result of Agglomerative**

A for loop is used to find the optimum threshold value, which starts from 0.75 to 0.85. In each iteration, this value is incremented by 0.0125. So, various scores, such as Silhouette, Calinski-Harabasz, Davies-Bouldin, etc., are evaluated for eight threshold values.

Figure 7(a) shows the Silhouette Score for threshold 0.75 to 0.85. The Bag-Of-Words technique was used to embed text into vectors. According to Figure 7(a), the value of threshold is 0.8.

Figure 7(b) shows the Silhouette Score for threshold 0.75 to 0.85. The TF-IDF technique was used to embed text into vectors. According to Figure 7(b), the value of threshold is 0.7875.

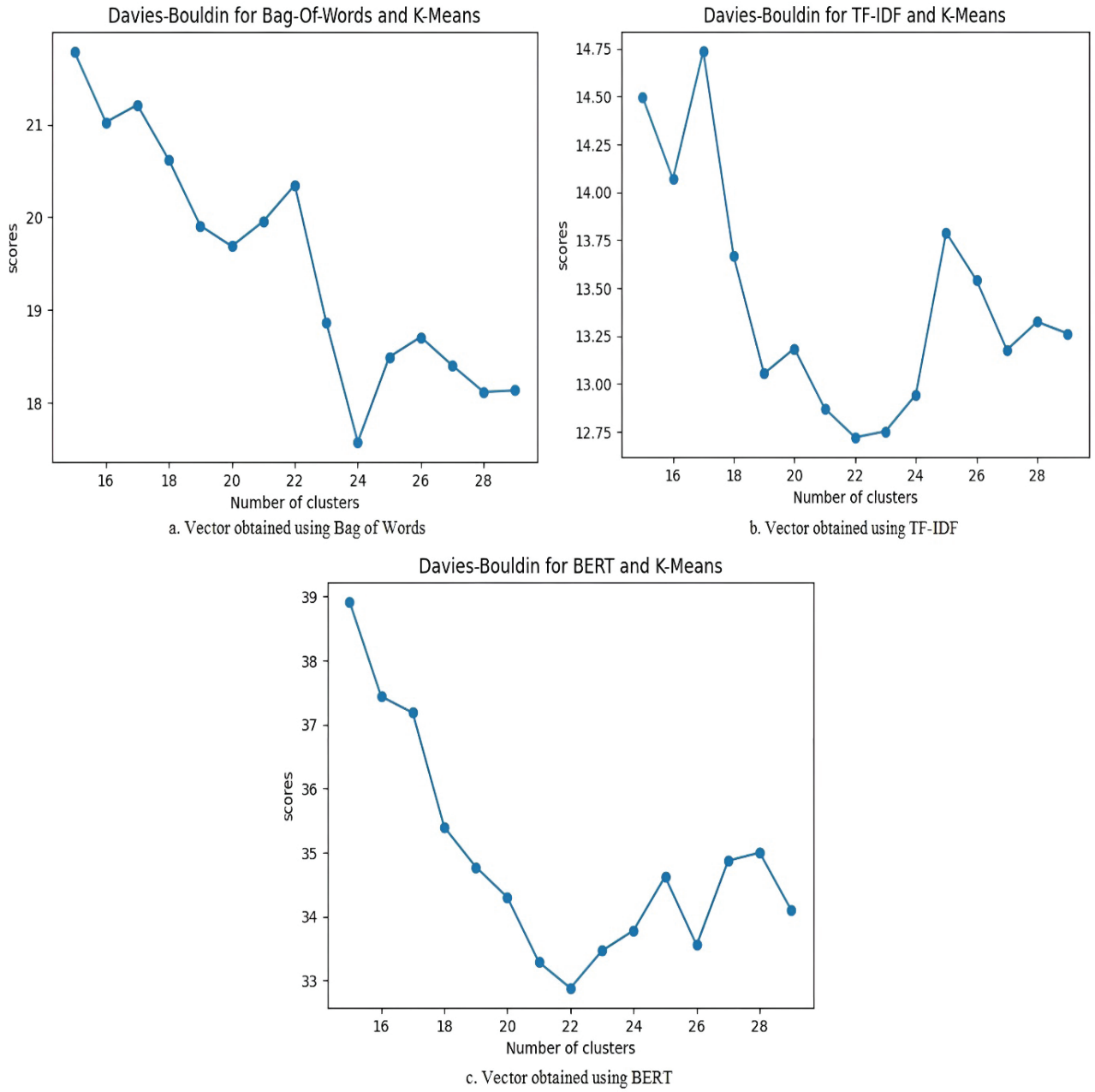


**Figure 6.** Vector obtained using Davies-Bouldin (a) Bag-Of-Words (b) TF-IDF (c) BERT of K-mean

Figure 7(c) shows the Silhouette Score for threshold 0.75 to 0.85. The BERT technique was used to embed text into vectors. According to Figure 7(c), the value of threshold is 0.7875.

Figure 8(a) shows the Calinski-Harabasz Score for threshold 0.75 to 0.85. The Bag-Of-Words technique was used to embed text into vectors. According to Figure 8(a), the value of threshold is 0.7875.

Figure 8(b) shows the Calinski-Harabasz Score for threshold 0.75 to 0.85. The TF-IDF technique was used to embed text into vectors. According to Figure 8(b), the value of threshold is 0.7875.

Figure 8(c) shows the Calinski-Harabasz Score for threshold 0.75 to 0.85. The BERT technique was used to embed text into vectors. According to Figure 8(c), the value of threshold is 0.7875.

Figure 9(a) shows the Davies-Bouldin score for thresholds 0.75 to 0.85. The Bag-Of-Words technique was used to embed text into vectors. According to Figure 9(a), the value of the threshold is 0.8.

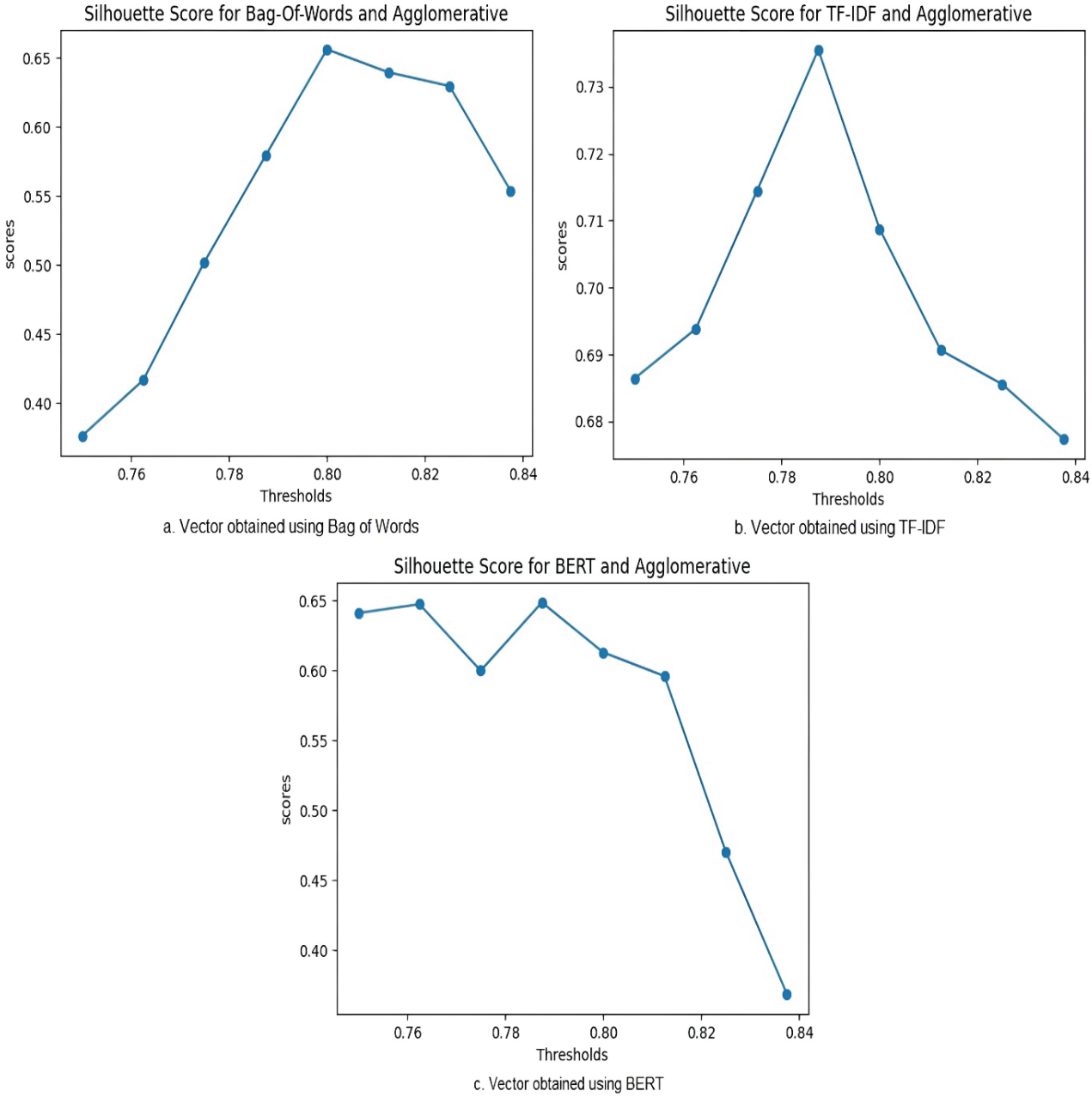


**Figure 7.** Vector obtained using Silhouette Score (a) Bag-Of-Words (b) TF-IDF (c) BERT of agglomerative.

Figure 9(b) shows the Davies-Bouldin score for threshold 0.75 to 0.85. The TF-IDF technique was used to embed text into vectors. According to Figure 9(b), the value of threshold is 0.7875.

Figure 9(c) shows the Davies-Bouldin score for thresholds 0.75 to 0.85. The BERT technique was used to embed text into vectors. The threshold value is 0.7875.

So, considering Figure 9(c), the value of K was set to 22. whereas for Agglomerative the threshold was set to 0.7875.

Once the clustering was done, the important task was to check the quality of the clusters, i.e., whether similar news was clustered in the same group or not. The proposed algorithm finds out the number of similar news articles correctly clustered in the same group. The accuracy was obtained by dividing this number by the total number of articles obtained, and finally, this ratio was multiplied by 100.

$$Accuracy = \frac{Total\ number\ of\ articles\ obtained}{Total\ Number\ of\ Articles} \times 100 \quad (29)$$

Tables 3 and 4 show the optimal and threshold values for tailoring the clustering process to the specific characteristics

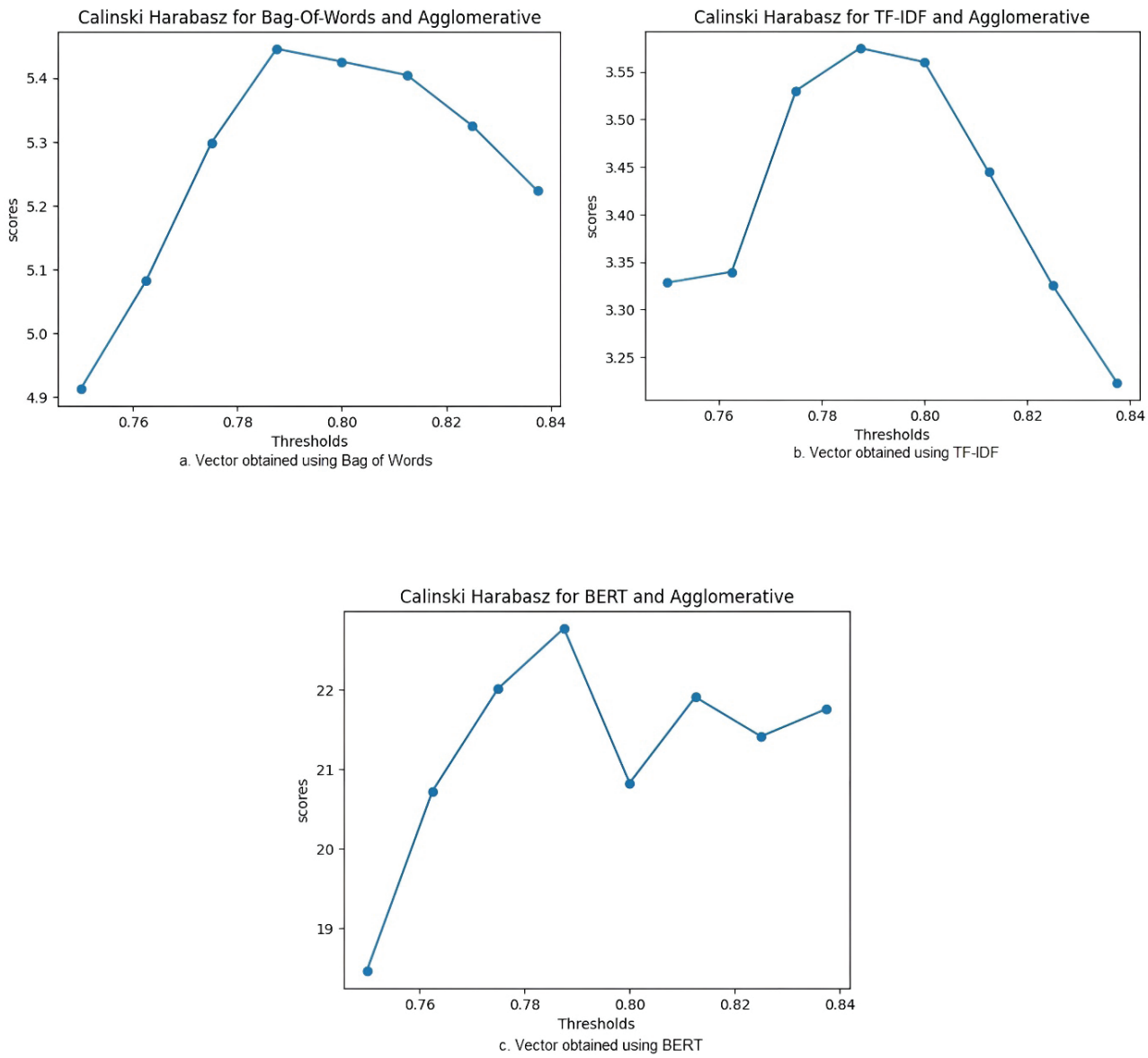


**Figure 8.** Vector obtained using Calinski-Harabasz (a) Bag-Of-Words (b) TF-IDF (c) BERT of agglomerative.

of the similar news article dataset and to ensure effective clustering of similar news articles.

Table 5 shows the accuracy of each embedding technique and clustering technique combination. It concludes that the combination of BERT and K-means algorithms was better than all other combinations.

A comparative analysis of the proposed BERT based clustering model is presented in Table 4 to the existing clustering algorithms that are used for clustering news articles. Malo et al. cite{malo} achieved an accuracy of 71% using the LPS approach, while Krishnamoorthy et al. cite{krishnamoorthy2} achieved 89% accuracy. The team's approach also reached an accuracy of 71% in the HSC assessment. Researchers achieved a score of 80.31% with a model they called "Ctrl-BERT". By combining K-means clustering with a pre-trained BERT language model, the study achieved a clustering accuracy of 84.46%, outperforming previously reported results in text-based classification tasks. This better performance is because BERT can understand the full context of a news article, including the relationships between the different pieces of information in it. This means it groups similar news stories more accurately. The proposed approach exhibits a notable edge over traditional methodologies as reflected in the data (Table 6).

When a technique that uses BERT to derive word embeddings and K-Means clustering was tested, it achieved higher accuracy than methods which use Bag-of-Words and Term

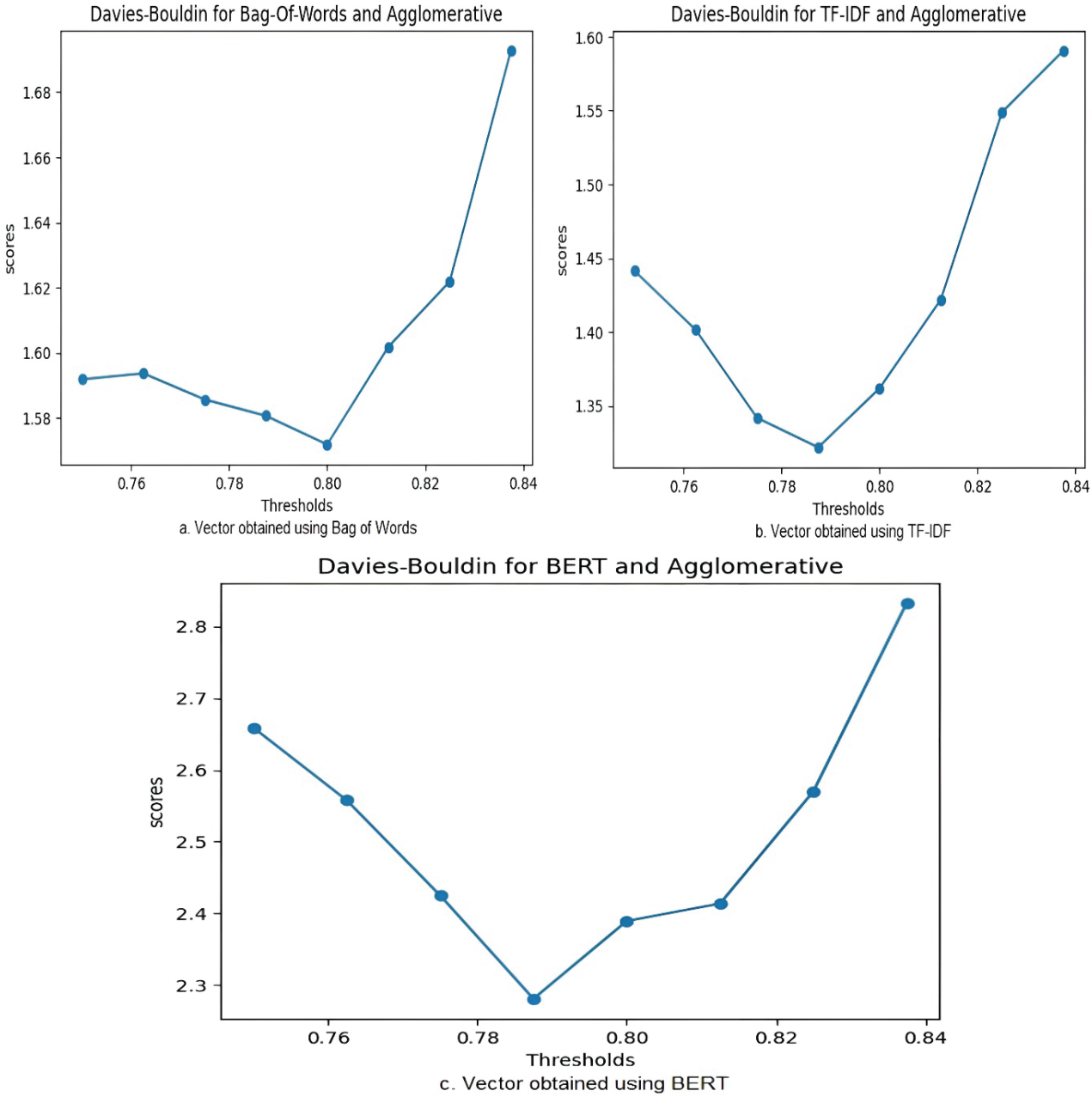


**Figure 9.** Vector obtained using Davies-Bouldin (a) Bag-Of-Words (b) TF-IDF (c) BERT of agglomerative.

Frequency-Inverse Document Frequency in experiments, 84.46% clustering accuracy was achieved. One of the main reasons BERT can improve upon traditional clustering methods is its ability to identify the deep relationships that exist between words. This relationship identification helps to improve the accuracy of the clustering results. In addition, several overall performance metrics were found to be useful. This includes the Davies-Bouldin Index, Calini ski - Harabasz statistic and the Silhouette Statistic, which are three methods that give cluster evaluation from multiple angles. Despite the effectiveness of BERT, one major drawback of BERT is the significant increase in computational power required for its operation, leading to difficulties when working with datasets of a considerable size. In order to minimize the effect of computational overhead, use of batch operations and the distribution of computational task were adopted. While BERT embeddings are effective to retain contextual and semantic information within text, which offers higher computational resources compared to Word2Vec and TF-IDF. This limitation becomes more

**Table 3.** Optimum value of 'K' using K-Mean algorithm

| Methods | Embedding techniques | K-value |
|---|---|---|
| Elbow | Bag of Words | 25 |
| | TF-IDF | 22 |
| | BERT | 22 |
| Silhouette Score | Bag of Words | 22 |
| | TF-IDF | 22 |
| | BERT | 23 |
| Calinski-Harabsz Score | Bag of Words | 23 |
| | TF-IDF | 22 |
| | BERT | 22 |
| Davies-Bouldin Score | Bag of Words | 24 |
| | TF-IDF | 22 |
| | BERT | 22 |

**Table 4.** Optimum value of threshold using agglomerative algorithm

| Methods | Embedding Techniques | Threshold Value |
|---|---|---|
| Silhouette Score | Bag of Words | 0.8 |
| | TF-IDF | 0.7875 |
| | BERT | 0.7875 |
| Calinski-Harabsz Score | Bag of Words | 0.7875 |
| | TF-IDF | 0.7875 |
| | BERT | 0.8 |
| Davies-Bouldin Score | Bag of Words | 0.7875 |
| | TF-IDF | 0.7875 |
| | BERT | 0.8 |

**Table 5.** Comparative result analysis of each embedding technique and clustering algorithm

| Embedding Technique | Size of Each Vector | Clustering Algorithm | Accuracy (%) |
|---|---|---|---|
| Bag of words | 1567 | K-means clustering | 63.89 |
| | | Agglomerative clustering | 60.16 |
| TF-IDF | 1567 | K-means clustering | 74.96 |
| | | Agglomerative clustering | 68.85 |
| BERT | 512 | K-means clustering | 84.46 |
| | | Agglomerative clustering | 80.71 |

pronounced when working with large-scale datasets. In contrast, to ensure that the proposed method can be efficiently used with large datasets, the K-Means method was optimised. The linear computational complexity of K-Means makes it properly matched for handling large volume of datasets. An additional efficiency was achieved by employing BERT embeddings with the use of batch operations and distributed computing methodologies.

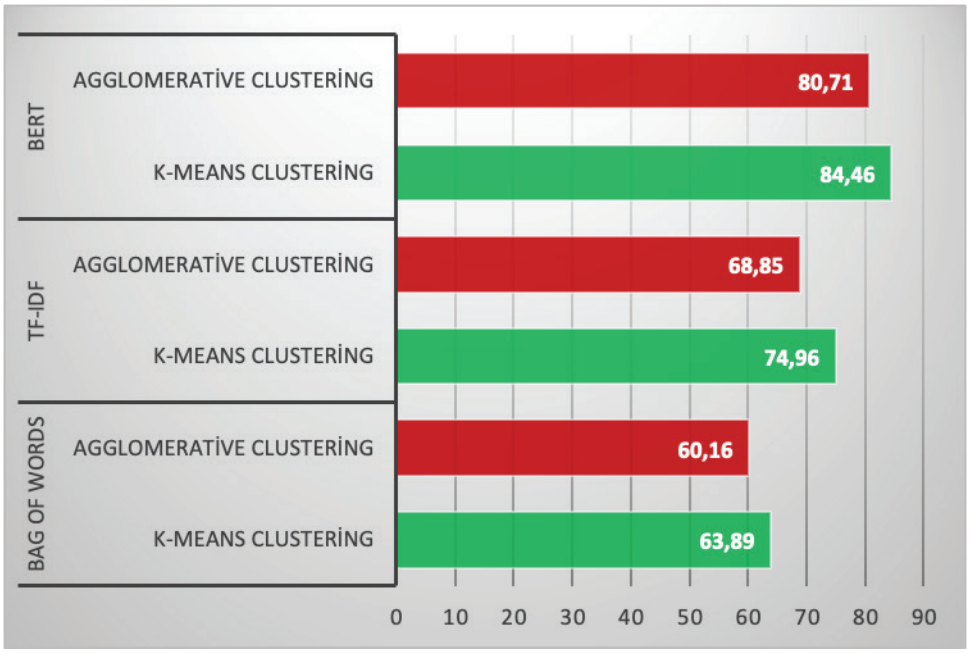


**Figure 10.** Bar represents the result analysis of each embedding technique and clustering algorithm.

**Table 6.** Comparative result analysis of cluster with existing methods

| Authors | Methods | Accuracy in % |
|---|---|---|
| Malo et. al [35] | LPS | 71.00 |
| Krishnamoorthy et al. [36] | HSC | 71.00 |
| Raaci et al. [37] | ULMFit | 83.00 |
| **Proposed** | **BERT** | **84.46** |

## CONCLUSION

Large archives of similar news stories can be better understood and their key information identified through grouping news stories of a similar type together. News articles can be grouped based on their statistical properties by applying clustering algorithms such as K-means, hierarchical agglomerative clustering, Hierarchical or other techniques. The advantages of clustering data include the simplification of data retrieval, better structured data, the ability to more easily identify trends, the generation of tailored suggestions and the improvement of the process of exploring data. By facilitating browsing through news articles, it is possible to more easily discover trends and links. This also allows users to learn about topics which are becoming popular. This allows readers to gain greater insight into news and trends which are developing.

The number of clusters, K, which best categorizes news articles of a similar nature can differ based on the specific collection of articles and the degree of similarity desired. Since there is no single universally applicable value for the number of clusters (K) that works for all data sets, it is crucial to determine the ideal number of clusters by taking into account several factors. These include the nature of the data, the amount of data available, and the goals which the cluster analysis aims to achieve. Upon examination of all these algorithms and methods, it appears that the best outcome for this grouping of like news stories is obtained by setting k to 22 and a threshold of 0.7875. Presently the integration of real-time adaptability and deep learning into a news aggregator poses difficulties such as the maintenance of low computational costs, the processing of continuous streams of data and the adaptation to changing news topics. In future work, we anticipate the challenges identified in this research to be addressed by various methods including online learning and incremental clustering. Texts such as news websites can also use clustering to manage their content and offer readers with a news feed that is more tailored to their interests. Research into the grouping of similar newspaper articles is likely to be mainly based on deep learning techniques, the ability to process data in real time, clustering methods which use elements of other techniques, the incorporation of data in multiple formats,

understanding how decisions are made, tailoring the results to individual users, the detection of false information and large-scale applications. While the temporal effects of news articles were not studied here, their consideration is essential for further research. The evolving nature of news stories can be captured by the incorporation of techniques such as trend analysis or time-based clustering. In the rapidly changing news media environment, a goal of this research is to improve both the usability and the accuracy of cluster analysis tools.

## AUTHOR CONTRIBUTION

Conceptualization, Gaurav V. Dahake and Dr. M. S. Ali; methodology, software, validation, formal analysis, Gaurav V. Dahake; investigation, Gaurav V. Dahake; resources, Gaurav V. Dahake; data curation, Gaurav V. Dahake; writing-original draft preparation, Gaurav V. Dahake; writing-review and editing, Gaurav V. Dahake; visualization, Gaurav V. Dahake; supervision, Dr. M. S. Ali; project administration, Dr. M. S. Ali.

## DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

## CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. ETHICS There are no ethical issues with the publication of this manuscript.

## STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

## REFERENCES

[1] Disayiram N, Rupasingha RA. A comparative study of clustering English news articles using clustering algorithms. In: Proc Int Res Conf Smart Comput Syst Eng; 2022; Colombo, Sri Lanka. p. 108–113. [CrossRef]

[2] Singh R, Singh S. Text similarity measures in news articles by vector space model using NLP. J Inst Eng India Ser B 2020;102. [CrossRef]

[3] Zhang Z, Liu X, Wang L. Spectral clustering algorithm based on improved Gaussian kernel function and beetle antennae search with damping factor. Comput Intell Neurosci 2020;2020. [CrossRef]

[4] Bianchi F, Grattarola D, Alippi C. Spectral clustering with graph neural networks for graph pooling. In: Proc Int Conf Mach Learn; 2020. p. 874–883.

[5] Hu Z, Nie F, Wang R, Li X. Multi-view spectral clustering via integrating nonnegative embedding and spectral embedding. Inf Fusion 2020;55:251–259. [CrossRef]

[6] Mirzal A. Statistical analysis of microarray data clustering using NMF spectral clustering k-means and GMM. 2020 2nd International Conference on Computer and Information Sciences (ICCIS) 13-15 Oct. 2020. [CrossRef]

[7] Viale L, Daga P, Fasana A, Garibaldi L. Dimensionality reduction methods of a clustered dataset for the diagnosis of a SCADA-equipped complex machine. Machines 2022;11:36. [CrossRef]

[8] Yujia S, Jan P. High-dimensional text clustering by dimensionality reduction and improved density peak. Wirel Commun Mob Comput 2020;2020:8881112.

[9] Zhang Z, Liu X, Wang L. Spectral clustering algorithm based on improved Gaussian kernel function and beetle antennae search with damping factor. Comput Intell Neurosci 2020;2020:1648573. [CrossRef]

[10] Izhar T, Mohammad U, Said MM, Ishak N, Dita A, Elly A. Clustering fake news with K-means and agglomerative clustering based on Word2Vec. Int J Math Comput Res 2024;12:3999–4007. [CrossRef]

[11] Hu Z, Nie F, Wang R, Li X. Multi-view spectral clustering via integrating nonnegative embedding and spectral embedding. Inf Fusion 2020;55:251–259. [CrossRef]

[12] Ravi J, Kulkarni S. Text embedding techniques for efficient clustering of Twitter data. Evol Intell 2023;16:1667–1677. [CrossRef]

[13] Shevendrakumar, Dubey. Clustering and retrieval of news articles using natural language processing. Int J Sci Res Eng Manag 2023;7:1–5. [CrossRef]

[14] Bhatnagar V, Ahuja S, Kaur S. Discriminant analysis-based cluster ensemble. Int J Data Min Model Manag 2015;7:83. [CrossRef]

[15] Bisandu B, Prasad R, Liman M. Clustering news articles using efficient similarity measure and N-grams. Int J Knowl Eng Data Min 2018;5:333–348. [CrossRef]

[16] Bergamaschi S, Guerra F, Orsini M, Sartori C, Vincini M. RELEVANTNews: A semantic news feed aggregator. 2007. Available at: https://ceur-ws.org/Vol-314/45.pdf Accessed on Jan 19, 2026.

[17] Gionis A, Mannila H, Tsaparas P. Clustering aggregation. ACM Trans Knowl Discov Data 2007;1:1217303. [CrossRef]

[18] Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural natural language processing for unstructured data in electronic health records: A review. Comput Sci Rev 2022;46:100511. [CrossRef]

[19] Bu F, Hu C, Zhang Q, Bai C, Yang LT, Baker T. A cloud-edge-aided incremental high-order possibilistic c-means algorithm for medical data clustering. IEEE Trans Fuzzy Syst 2021;29:148–155. [CrossRef]

[20] Jaya Mabel Rani A, Pravin A. Clustering by hybrid k-means-based rider sunflower optimization algorithm for medical data. Adv Fuzzy Syst 2022;2022:7783196. [CrossRef]

[21] Lynn N, Suganthan PNS. Ensemble particle swarm optimizer. Appl Soft Comput 2017;55:533–548. [CrossRef]

[22] Li HR, Tian XX, Luo XM, Yan M, Mu Y, Lu HG, Li SD. Heteroborospherene clusters Nin ∈ B40 (n = 1–4) and heteroborophene monolayers Ni2 ∈ B14 with planar heptacoordinate transition-metal centers in η7-B7 heptagons. Sci Rep 2017;7. [CrossRef]

[23] Ullmann T, Beer A, Hünemörder M, et al. Over-optimistic evaluation and reporting of novel cluster algorithms: An illustrative study. Adv Data Anal Classif 2023;17:211–238. [CrossRef]

[24] Bischl B, Binder M, Lang M, Pielok T, Richter J, Coors S, et al. Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. arXiv 2021. Preprint. doi: 10.48550/arXiv.2107.05847

[25] Karwa R, Honmane V. Building search engine using machine learning technique. In: Proc Int Conf Comput Commun Syst; 2019. p. 1061–1064. [CrossRef]

[26] Messina A, Montagnuolo M. Content-based RSS and broadcast news streams aggregation and retrieval. In: Proc Int Conf Digit Inf Manag; 2008. p. 93–98. [CrossRef]

[27] Radev D, Blair-Goldensohn S, Zhang Z, Raghavan R. NewsInEssence: A system for domain-independent, real-time news clustering and multi-document summarization. In: Proc Int Conf Hum Lang Technol Res; 2001. [CrossRef]

[28] Felix H, Norman M, Bela G. Bias-aware news analysis using matrix-based news aggregation. Int J Digit Libr 2020;21:129–147. [CrossRef]

[29] Zhou B, Gao T. Automatic method for determining cluster number based on silhouette coefficient. Adv Mater Res 2014;951:227–230. [CrossRef]

[30] Mehta V, Bawa S, Singh J. WEClustering: Word embeddings-based text clustering technique for large datasets. Complex Intell Syst 2021;7:3211–3224. [CrossRef]

[31] Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv 2016. Preprint. doi: 10.48550/arXiv.1607.06450

[32] Wang H, Zhou C, Li L. Design and application of a text clustering algorithm based on parallelized k-means clustering. Rev Intell Artif 2019;33:453–460. [CrossRef]

[33] Thakare Y, Bagal S. Performance evaluation of k-means clustering algorithm with various distance metrics. Int J Comput Appl 2015;110:12–16. [CrossRef]

[34] Dubey D. Clustering and retrieval of news articles using natural language processing. Procedia Comput Sci 2022;207:3449–3458. [CrossRef]

[35] Malo P, Sinha A, Korhonen PJ, Wallenius J, Takala P. Good debt or bad debt: Detecting semantic orientations in economic texts. J Assoc Inf Sci Technol 2014;65:782–796. [CrossRef]

[36] Krishnamoorthy S. Sentiment analysis of financial news articles using performance indicators. Knowl Inf Syst 2018;56:373–394. [CrossRef]

[37] Raaci D. FinBERT: Financial sentiment analysis with pre-trained language models. arXiv 2019. Preprint. doi: 10.48550/arXiv.1908.10063