FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 33, No 4, December 2020, pp. 631 - 653
https://doi.org/10.2298/FUEE2004631A

A NEW ANOMALOUS TEXT DETECTION APPROACH USING UNSUPERVISED METHODS

Elham Amouee[1], Morteza Mohammadi Zanjireh[1], Mahdi Bahaghighat[1], Mohsen Ghorbani[2]

[1]Computer Engineering Department
Imam Khomeini International University
Qazvin, Iran
[2]Department of Electrical Engineering
Raja University
Qazvin, Iran

Abstract. *Increasing size of text data in databases requires appropriate classification and analysis in order to acquire knowledge and improve the quality of decision-making in organizations. The process of discovering the hidden patterns in the data set, called data mining, requires access to quality data in order to receive a valid response from the system. Detecting and removing anomalous data is one of the pre-processing steps and cleaning data in this process. Methods for anomalous data detection are generally classified into three groups including supervised, semi-supervised, and unsupervised. This research tried to offer an unsupervised approach for spotting the anomalous data in text collections. In the proposed method, a combination of two approaches (i.e., clustering-based and distance-based) is used for detecting anomaly in the text data. In order to evaluate the effciency of the proposed approach, this method is applied on four labeled data sets. The accuracy of Naïve Bayes classification algorithms and decision tree are compared before and after removal of anomalous data with the proposed method and some other methods such as Density-based spatial clustering of applications with noise (DBSCAN). Our proposed method shows that accuracy of more than 92.39% can be achieved. In general, the results revealed that in most cases the proposed method has a good performance.*

Key words: *Anomaly detection, text mining, unsupervised learning, clustering, pre-processing, DBSCAN algorithm.*

Received April 15, 2020; received in revised form August 10, 2020
Corresponding author: Morteza Mohammadi Zanjireh
Computer Engineering Department, Imam Khomeini International University, Qazvin, Iran
E-mail: Zanjireh@eng.ikiu.ac.ir

1 Introduction

The current age is called Information Age. Since the organizations and institutions record a huge amount of data daily, data recovery alone is not enough to make decisions. So automatic classification and analysis of data is very important. Data mining is the process of identifying valid patterns and relationships among the high volume of data which have so far been unknown [1]. Intelligent data exploring helps organizations to discover and predict system behaviors, and patterns to make better and faster decisions. Besides, the Machine learning (ML) is the science that deals with the development of algorithms and statistical models. In machine learning, the goal is to enable computer systems to perform particular tasks without using explicit instruction and merely using patterns and inference instead of being able to perform their functions. Nowadays, this science is wildly used in broad fields such as image processing, machine vision, audio signal processing, natural language processing (NLP), communication networks, financial areas, and so on [2–10].
In many topics of data mining, the data is classified into structured, semi-structured, and unstructured [11]. Data mining and machine learning are strong tools to handle all of these problems. Structured data is that which has a predictable and regular format such as the structure of the tables in relational databases. In contrast, unstructured data is that which does not a specific structure and its analysis is not so easy. The significant growth and diversity of text data can be considered as an example of this data type. Volume and speed of unstructured data are several times more than those of structured type. Therefore, one of the applied areas in data mining is the concept of text mining and natural language processing (NLP). Before starting data mining, some steps should be taken in order to prepare data. The steps for data mining include selecting data, initial cleaning and pre-processing, discovering patterns, and interpreting and displaying them. Diagnosis of anomalous data can be considered as a pre-processing step in the data mining path [12].

Anomaly is a pattern that differs from the other patterns existing in the

data set. Anomaly was first defined by Grubbs (1969): Anomaly is data which dramatically deviates from other available samples in the series [13]. The term 'anomaly in text data' is referred to texts which are abnormal or are significantly different from the other texts in terms of concept. In the text data, anomaly can be investigated in terms of a difference in the text author, the subject, the genre, the style of text, and the emotional tone of the text [14].

The main reason for the development of text mining systems is the increasing volume of textual data in organizations and businesses. One of the challenges in monitoring infectious diseases, such as COVID-19, is that large volumes of textual data are produced continuously. In a pandemic, this value can be far greater than a human being can process [15, 16]. Among the applications of this field, the following can be mentioned [17, 18]

- 1. Diagnosis of anomalies in safety reports sent from space stations
- 2. Tracing the subject of news
- 3. Abnormal in web content
- 4. Identify significant patterns in annual financial reports
- 5. Identify abnormal data in news reports
- 6. Discover knowledge of medical records

There are several anomaly detection systems. These systems are comprised of three parts. The first phase is the pre-processing step which includes removing unwanted words through stemming [19]. In the second phase, text display (e.g., displaying text sentences for vector) is carried out. And the third phase includes text processing for detecting anomalies and comparing between documents.

Anomaly detection is not an easy challenge. So far researchers developed many anomaly detection methods using statistical methods, machine learning, and data mining but the problem is still open and in its progress. Several approaches are shown in the following Figure 1:

The methods of anomaly detection are widespread, and each is used based on input data type and its application. In one approach, the methods are classified based on access to the labelled data. Accordingly, the methods are categorized into three main categories [20]:
1. Supervised anomaly detection

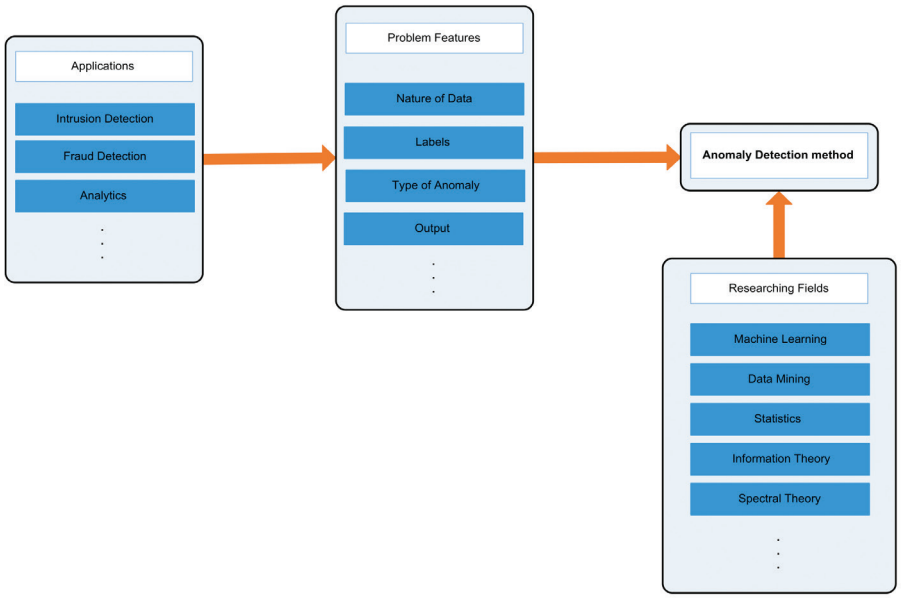

Fig. 1: The key components of anomaly detection methods [12]

2. Semi-supervised anomaly detection
3. Unsupervised anomaly detection

In the supervised method, both normal and abnormal data are labeled in the training dataset. Typical approach in such a method is to create a predictive model for both normal and abnormal data. After comparing each test data sample with the model, it is determined to which class this data belongs [12]. In the semi-supervised method, it is supposed that only normal samples are labelled. Since this method does not require anomalous data labelling, this method is more applicable than the supervised method [11]. In comparison, the methods which are run based on unsupervised method do not require the training data; so, they are more applicable than the two previous approaches. The most important advantage of this method is that it does not need to access the labelled data. Usually, this group of methods is known as clustering solutions [11]. In the Figure 2, a summary of a set of supervised and unsupervised anomaly detection methods is shown.

Text clustering refers to the process of dividing a text group into similar subgroups based on content. Semantic clustering refers to cluster texts based

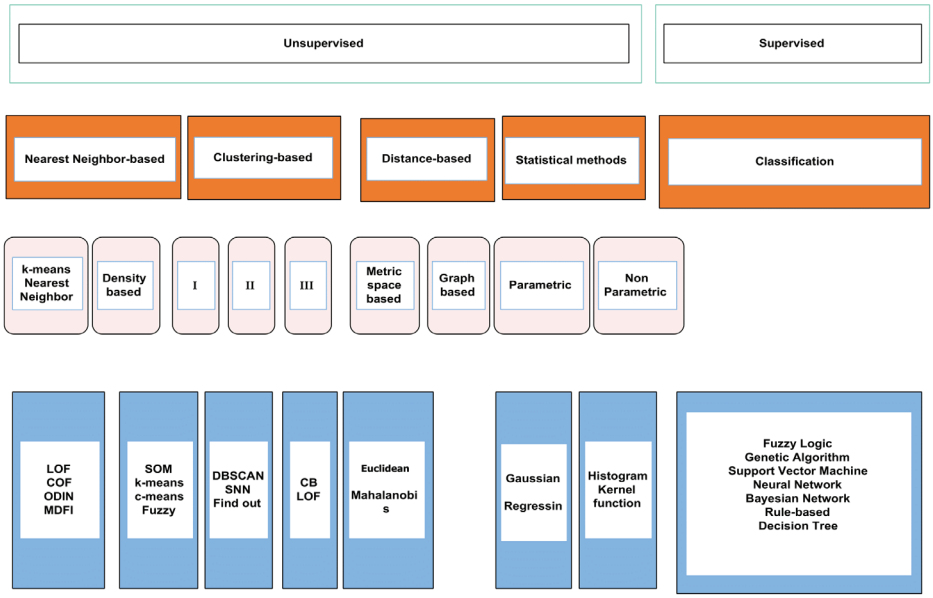

Fig. 2: Various categories of anomalous data detection methods: The supervised and unsupervised approaches [20], [11].
I: The assumption of these methods is that the normal data belong to at least one cluster while the anomalous data do not belong to any cluster [11].
II: In this method, the normal instance locates near the centroid of its nearest cluster, while the anomalous sample is in a long distance of the nearest cluster gravity center.
III: Normal data belongs to high density clusters while anomalous data is distributed in low density clusters.

on their contents or meaning [21, 22].

The remainder of this paper is organized as follows. In Section 2, we review some related works. In Section 3, we present the methodology of our proposed method. The simulations and experimental results of the proposed algorithm are presented in Section 4. Finally, in Section 5, we conclude the paper.

2 Related works

There are many studies in the research literature that try to address the anomaly detection issues. Consequently, different algorithms are provided to diagnose anomaly in multidimensional data sets. The key methods used in this area include the distance-based approach, the density-based approach, and their subset methods [23]. Many researchers have worked to detect anomalies in textual documents. Hence, different aspects of text attributes are considered. A number of researchers changed the text to numbers and used algorithms that are suitable for numerical data. Others chose a limited number of documents, such as document titles, to detect anomalies and find a pattern for the dataset [24].

In [25], the authors used the conceptual graph method in order to identify anomalous data in the text. Two-way graph that includes two different types of nodes (i.e., concepts and relations). This method differs from classical statistical approaches and distance-based approaches. In this approach, a data deviation is identified based on the concept of regularity. Using a conceptual graph and the relationships between the entities (concepts), the template pattern is identified, and patterns differing from those that are rare are deemed as an anomaly [25]. Sumithiradevi et al. used clustering methods for anomalous data detection. Initially using the greedy method, they improved the k-means algorithm and clustered the data set. Then, all records were read, and a flag with the initial value of zero was attributed to them. Later on, one sample was considered as outlier and removed them from the data set. In the next step, the amount of entropy changes in the remaining set was calculated. If the entropy of the remaining set increases by removing data, the deleted data is anomaly, and the value of its flag is changed to one [26].

Juntao Wang et al. made use of the density-based approach in order to remove anomalous data. First, by clustering approach of Fast k-means, they classified data sets. Then, for each data in the cluster, the degree of anomaly was calculated, and any data whose anomalous degree is much larger than one is removed from the data set. In the next step, the average of the remaining set of a cluster is selected as the new center gravity of the cluster. This trend continues until converging the clusters so that all of the anomalous data is removed from the data set [27]. In [28], the authors applied a similar method to the Local Outlier Factor (LOF) algorithm in order to determine the degree of anomaly based on distance from centroid. In the first step, by improving the K-means algorithm using genetic algorithm, data set

is clustered. In the next step, the data set is filtered, and by defining the threshold limit, the data items whose degree of anomaly is more than a specific value is considered as anomalous data. In this method, for each vector a degree of anomaly is determined based on the distance from the centroid of the cluster. Lei et al. using the subtractive clustering algorithm estimated the potential of each data to be as the initial seeds according to the neighborhood's radius of samples. In the next step, by combining the Silhouette index with the K-means algorithm, they improved the estimated number of clusters. the Silhouette index is a criterion for measuring the amount of the desirability of the data assignment to the cluster. This means that each data is closer to the samples in its cluster or to data from other clusters. If the number obtained is closer to one, the assignment of data to the cluster is desirable but if the number is closer to 0.5, this means that it is likely that data belongs to another cluster. Finally, the improved Cluster-Based Local-Outlier Factor (CBLOF) algorithm is used to identify anomalous data [29]. In [30], the authors put their work on the basis of the improvement of the K-means algorithm clustering and established their method in parallel. Firstly, using Principal Component Analysis (PCA), they decreased dimensions of the problem. Then, by applying the DD Algorithm [31], they improved K-means' performance. This algorithm selects initial seeds according to the method of distribution of data and improves their choice quality. Also, instead of using a certain point as the initial seeds, it uses the average value of some points as the centroid of the cluster. Through a number of tests, Yin et al. were able to define a threshold in order to determine the number of clusters in the K-means algorithm clustering so that there is no need for it to be determined by the user [30].
Identifying items or events that do not match the expected patterns or other items in the data set is called an anomaly detection. These anomalies items cause problems such as structural defects, errors, credit card fraud, and a cyber-attack and etc. The ability to detect anomaly behavior can provide very useful insights into various industries and be an important key to solving these problems.

Machine learning algorithms make processing faster and more efficient for detecting anomalies. These algorithms can learn from data and predictions based on that data. In [32], they examined the issue of discovering emerging relations from news using machine learning. These relations can help with news-related tasks, such as retrieving the news, discovering events, ranking, and more, which is a challenging task. In this research, a novel Heterogeneous graph Embedding framework for Emerging Relation detection (HEER)

and a global graph perspective was presented. HEER can embed words and entities by learning from the heterogeneous textual graph and the knowledge graph and predicts the emerging relations via a positive and unlabeled learning (PU) classifier. In [33], the authors presented a kernel-based ensemble clustering approach and used a prototype reduction scheme to decrease the time required to generate the ensemble members. They showed that the reduction method could improve the results. The method they used was a learning process for documents clustering that correspondence-based aggregation in conjunction with kernel clustering on a matrix constructed using density-biased prototype selection.

3 Proposed Method

Similar to other existing studies such as [32–35], we deal with numerical data in this research. A combination of clustering-based and distance-based methods was used by us. At the first, it is required to convert text data into an understandable format for the system. To this end, text documents were converted into vectors. anomalous data is detected in two phases. In the first phase, the K-means algorithm is used for clustering of data items in the k clusters. In the second phase, anomalous data in each cluster is detected based on the similarity comparison of each data item with the document of the centroid of the cluster. In clustering phase, probably some clusters of empty values might be created, and/or one of the anomalous data items is selected as the initial seeds and forming a cluster. Therefore, the clustering stage is carried out several times in order to get a more desirable result. In the step after clustering, centroid of each cluster is considered as the representative of that cluster and since the text data is displayed in the vector space, using the Cosine Similarity (CS) formula, the angle between the data within the cluster is compared to its center. It should be compared with the threshold limit. If the similarity rate is less than the threshold, the data is considered as an anomaly. It should be noted that the number of abnormal data is negligible in comparison with the total data in the set. In this approach, the method of K-means algorithm clustering was used to divide data set to a few smaller parts based on the criterion of Cosine Similarity which will result in a decrease in the number of comparison of documents. In other words, instead of calculating distance (similarity) of each and every document in the whole set of data, we first divide the set according to the most similar documents in the K-means algorithm clustering so that the number of comparison between documents within each cluster with the cluster centroid

will be fewer. To cluster text data and to determine their similarity rate, the bag of words approach and Cosine Similarity were used, respectively. In order to identify anomalous data in each cluster, we are looking for data items that differ from the behavioral pattern of other members or their differences from the cluster centroid is much. After the clustering phase, the weight of each cluster will be calculated using the following formula:

$$W_{(k)} = \sum_{x=1}^{n} S_x \tag{1}$$

where, k is the cluster number ($K = 1, 2, \ldots, n$) and given that the text attributes are moved to the vector, S_x is the level of similarity of each data to the cluster centroid. In the next step, the average similarity of documents in each cluster is calculated by the following formula:

$$M_{(k)} = \sum_{x=1}^{n} S_x / T_n \tag{2}$$

Thus, the total amount of document similarity relative to the centroid in each cluster is divided by the total number of documents T_n, and a numeric value is obtained as the average similarity of documents in each cluster. In the next step, the maximum and minimum amount of similarity in each cluster is calculated based on the following formulas:

$$S_{(k-max)} = Max(S_1, S_2, ..., S_n) \tag{3}$$

$$S_{(k-min)} = Min(S_1, S_2, ..., S_n) \tag{4}$$

$$Avg_{(k)} = (S_{k-max} + S_{k-min})/2 \tag{5}$$

$$Diff_{(k)} = M_{(k)} - Avg_{(k)} \tag{6}$$

Finally, the threshold limit of similarity is obtained for each cluster by this formula:

$$Threshold_{(k)} = \left|Diff_{(k)}\right| \tag{7}$$

The similarity of each data item in the cluster is compared to the centroid with the threshold value. If the similarity rate of the document to the

centroid is less than the threshold, it is considered as outlier in the cluster. As it was determined, to calculate the threshold limit, the difference between the median values of similarity and the average of similarities was used. Since the average of each cluster is obtained by dividing the sum of the values by the number of them, the existence of anomaly in the cluster leads to exceeding of standard deviation from the average value and increases the variance. As a result, mean similarity value of documents alone cannot be a good indicator of threshold limit. Therefor, the median amount of similarities, that is, the average of the similarity values of the most similar and most different documents are also entered into the threshold formula and its difference from the average values of the similarity will result in modulating the threshold limit.

The pseudo-code of proposed algorithm is shown as below:

Algorithm 1 The proposed algorithm

Require: Input ⇒
Data set $D = \{d1, d2.....dn\}$, where n is the number of Documents (k the number of clusters)
Ensure: output⇒
A set of K-clusters without outliers
Require: Choose k objects from D as initial cluster centers
repeat
1. Calculate distance of each data instances to centroid using CS
2. Reassign objects to the cluster with the most similarity
3. Update the cluster centroid due to the CS

until Until no changes
Calculate the weight based center W_k
repeat
1. Calculate the Mean CS as $M_{(k)}$
2. Calculate the $Max(S_1, S_2, ..., S_n)$ and $Min(S_1, S_2, ..., S_n)$
3. Calculate the threshold limit of similarity for each cluster

if $d_n < d_{(k)}$ **then**
Delete d_n from Cluster k
end if
until End of Clusters

3.1 Data Set

In this study, two data sets (i.e., BBC and BBC sports News) were used [35]. BBC News contains 2225 documents from news articles on the BBC website in five news groups between 2004 and 2005. These five news groups were labeled under the title of Business, Entertainment, Politics, Sport, and Tech. The BBC sports Collection also contains 737 documents from the BBC sports website articles in five sports areas between 2004 and 2005, labeled as Athletics, Cricket, Football, Rugby, and Tennis. Each news data set contains a large number of text files from the broadcasted news text in several newsgroups with different topics on the BBC website. Since news topics differ in these texts, the words used in the text will also vary according to the type of news. According to the news genre, similar words are used in political news which are not used in sports news. Consequently, the words used in the sports news genre is similar, but it differs from the words of the business news genre. As a result, the similarity or difference of documents is characterized after conversion to the vector space. Documents in the sports news genre is placed in the same category at the clustering time. If business news documents are placed in this cluster, they are considered as outliers. The purpose of this approach is to find irrelevant documents that should be placed in a different cluster with regard to their subject. To identify anomaly, the algorithm is performed on the text data set with and without pre-processing. The goal of pre-processing of texts is stemming, removing of stop words, and weighing by term frequency–inverse document frequency (TF-IDF) method. Table 1 shows the details of the BBC data sets.

Table 1: Summarization of the data sets

Dataset	Descriptions	Documents	Class Labels	Classes
BBC News	News articles from BBC	2225	5	Business, Entertainment, Politics, Sport, Tech
BBC Sports	Sports news articles from BBC	737	5	Athletics, Cricket, Football, Rugby, Tennis

3.2 Text pre-processing

The implementation of various operations on the text, including classification and clustering, requires the conversion of it into an understandable format for the system. As mentioned earlier, text documents are of an unstructured data type, and to perform the calculations it is necessary to convert them to a structured way.

Stemming: The stemming process will convert words to their root form. For

example, the words 'apply,' 'applied,' and 'application' have the same root and they are all converted to the word 'apply' [36].
Stop words removal: In this step, a batch of worthless words like conjunctions and prepositions that are repeated alternately and do not have certain semantic meaning are deleted [37].
Bag-of-words model: it is a simple demonstration of text documents that are used in Natural Language Processing. In this model, each document is displayed regardless of grammar and how words are shown, but the number of words' repetition matters. The result obtained from this model, will be a word-document matrix in which every row represents each document and every column represents each word. If there is a word in the document, in the corresponding column in the matrix 1 will be inserted otherwise 0 will be inserted. The first reference to "bag-of-words model" in a linguistic context can be found in Zellig Harris's article on Distributional Structure [38]. In this study, to display the texts of the bag-of-words model and conversion to the vector space model were used. Vector space model is an algebraic model for displaying text documents in the vector space.
Weighing words by TF-IDF: In this method, words are assigned a weight based on its frequency in the text relative to their frequency in other texts. This weighing system shows how important a word is for a document. The first form of term weighting is due to Hans Peter Luhn (1957) which may be summarized as [39]:

The weight of a term that occurs in a document is simply proportional to the term frequency. IDF was introduced by Karen Spärck Jones as "term specificity" in [40, 41]. Although it has worked well as a heuristic, many researchers trying to find information theoretic justifications for it [41].

This criterion is made up of two functions of the TF (Term Frequency function) and IDF (Inverse Document Frequency function), and that means that if the number of repetitions of a specific word in the document is more and in other documents under investigation is less, this word is very important. This criterion is derived from the multiplication of two values (i.e., TF * IDF). TF equals the number of word repeats divided by the total number of words contained in the document [42, 43].

$$tf - idf_{t.d} = tf_{t.d} \times idf_t \tag{8}$$

$$tf_{t,d} = \begin{cases} \log(1 + f_{t,d}), & if f_{t,d} > 0 \\ 0, & otherwise \end{cases} \tag{9}$$

where t is the term, d is a bag of words (a document in IR terms), and $f_{t,d}$ is a frequency of the term in a bag.

$$tf_{t.d} = \frac{f_{t.d}}{Max(f_{t'.d} : t' \in d)} \tag{10}$$

IDF is the logarithm of the total number of documents divided by the total number of documents containing the target word [44].

$$idf_{t,D} = \log \frac{|D|}{|d \in D : t \in d|} = \log \frac{N}{df_t} \tag{11}$$

where N is the cardinality of a corpus D (the total number of classes) and the denominator df_t is a number of bags where the term t appears.

$$idf_t = \log \frac{N}{df_t} \tag{12}$$

Then, tf $*$ idf weight value for a term t in the bag d of a corpus D is defined as:

$$tf * idf(t, d, D) = tf_{t,d} \times idf_{t,D} \tag{13}$$

$$tf * idf(t, d, D) = \log(1 + f_{t,d}) \times \log \frac{N}{df_t}, for f_{t,d} > 0 \tag{14}$$

for all cases where $f_{t,d} > 0$ and $df_t > 0$, or zero otherwise.
Once all frequency values are computed, term frequency matrix becomes the term weight matrix, whose columns used as class' term weight vectors that facilitate the classification using Cosine Similarity.

In accordance with the equation, the less the number of word repetition in the documents containing the target word, the more important.

3.3 Criterion for Assessing the Similarity of Two Documents

The Cosine Similarity is the similarity criterion between the two vectors that calculates the cosine of the angle between the two vectors. A zero cosine is equal to 1, as a result, if two vectors coincide each other, their similarity is equal to one. It is obvious that this amount will show the highest possible similarity between vectors [45]. after preparing the words bag, the document will be displayed in the vector space. Then, the angle between the two vectors (the similarity of two documents) is calculated from the following formula:

For two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ Cosine Similarity is based on their inner product and defined as:

$$similarity(\boldsymbol{a}, \boldsymbol{b}) = cos(\theta) \tag{15}$$

$$cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{||a|| \cdot ||b||} \tag{16}$$

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i \tag{17}$$

$$||a||, ||b|| = \sqrt{\sum_{i=1}^{n} a_i^2}, \sqrt{\sum_{i=1}^{n} b_i^2} \tag{18}$$

$$similarity(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sum\limits_{i=1}^{n} a_i b_i}{\sqrt{\sum\limits_{i=1}^{n} a_i^2} \sqrt{\sum\limits_{i=1}^{n} b_i^2}} \tag{19}$$

4 Experimental Results

In the following, the results of using the proposed approach in order to identify the abnormal data in the text will be investigated. To implement and evaluate the proposed approach, the following procedures are implemented: To evaluate the accuracy of the proposed approach, the data set is first divided into two parts. A part of data is considered as training data and other part as test data to measure the accuracy of the proposed algorithm. Also, in order to weigh the text keywords, the TF-IDF coefficient was used. Since five news genres exists in the data set, the number of clusters in the K-means algorithm is pre-determined and is equal to five. Since the used data set is labeled, the training data is used in order to learn K-means algorithm. Then, the number of documents placed in each cluster will be counted. By appointing a two-dimensional array, the index of each document with a Cosine Similarity relative to the centroid of each cluster (distance criterion) is stored in the array. According to the proposed formula in order to determine the threshold limit, the lowest and highest similarity values in each cluster relative to the centroid as well as the average spacing values are calculated.

In the following, the results of the accuracy of decision tree classification

algorithms and Naïve Bayes method on two data sets before and after pre-processing by removing the anomalous data by the proposed method and the DBSCAN method was presented with different neighborhood distances.

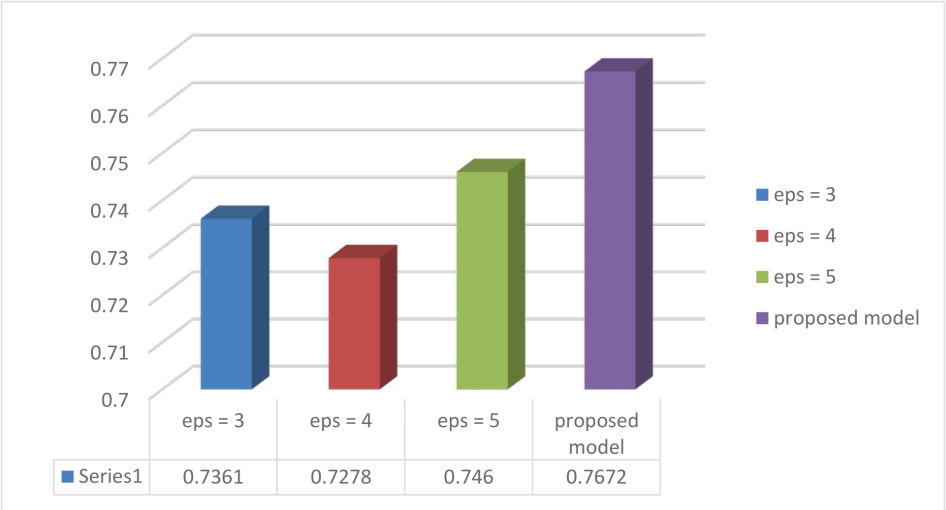

Fig. 3: Accuracy diagram of the decision tree after removing the anomalous data using the proposed method and DBSCAN with different neighborhood distance on the BBC data set

The results in figures 3, and 4 showed that the proposed method on the BBC data set before the pre-processing phase was not improved in comparison with the accuracy of Naïve Bayes method, but the accuracy of the decision tree using the proposed method increased in comparison with the DBSCAN method.

The results in figures 5, and 6 proved that the accuracy of the proposed method like previous results on a pre-processed set of BBC has been improved compared to the non-pre-processed set so that the accuracy of the Naïve Bayes method after eliminating the anomalous data by the proposed method increased compared to the DBSCAN method. Also, the accuracy of decision tree after eliminating the anomalous data by the proposed method has increased on the pre-processed data set.

The results in figures 7, and 8 disclosed that the proposed method on the BBC sports data set before the pre-processing phase in comparison with the accuracy of the Naïve Bayes method was not improved, but the accuracy of decision tree after removal of the anomalous data using the proposed method increased in comparison with the DBSCAN method.

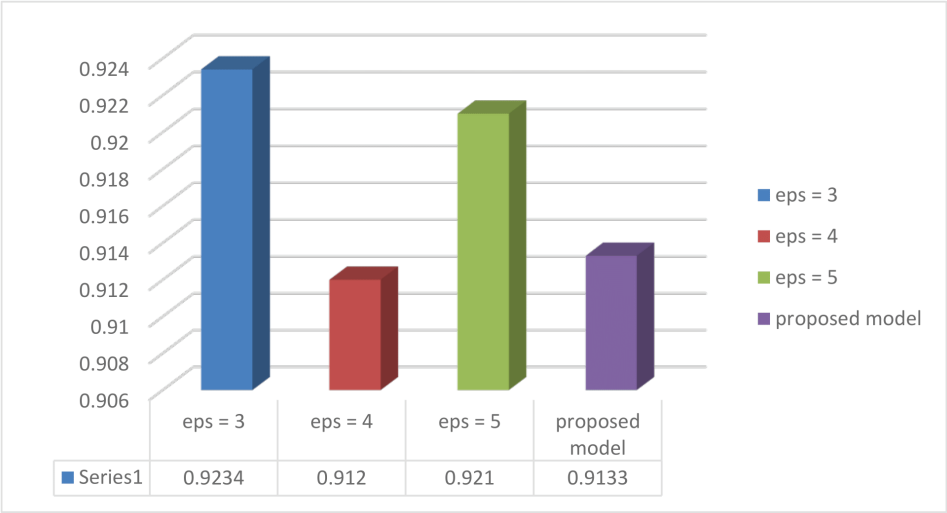

Fig. 4: Accuracy diagram of the Naïve Bayes algorithm after removing the anomalous data using the proposed method and DBSCAN with different neighborhood distance on the BBC data set

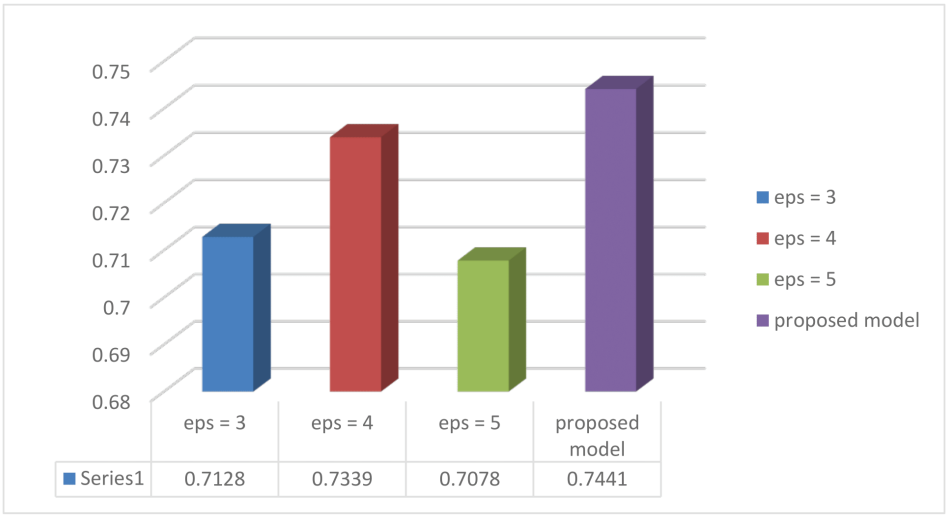

Fig. 5: Accuracy diagram of the decision tree after removing the anomalous data using the proposed method and DBSCAN with different neighborhood distance on the pre-processed BBC data set

The results in figures 9, and 10 revealed that the accuracy of the proposed method like previous results on a pre-processed set of BBC sports

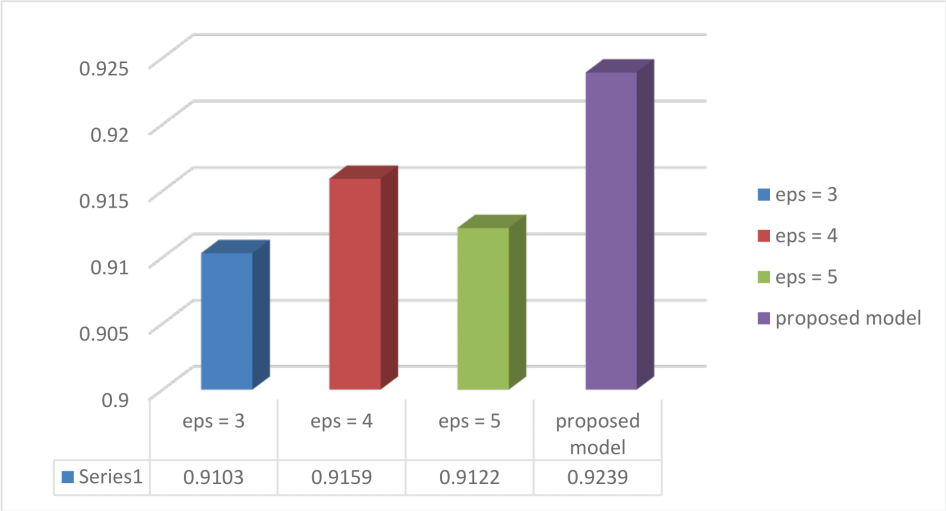

Fig. 6: Accuracy diagram of the Naïve Bayes algorithm after removing the anomalous data using the proposed method and the DBSCAN method with different neighborhood distance on the pre-processed BBC data set

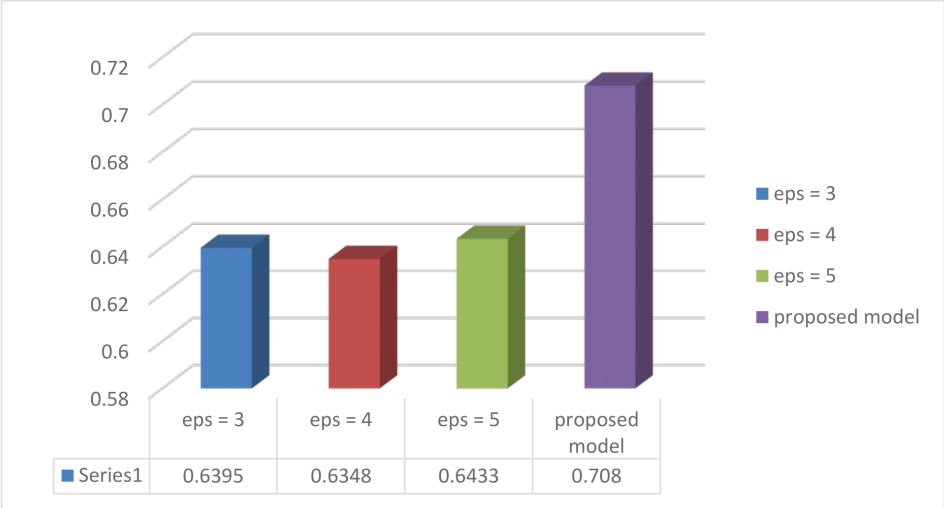

Fig. 7: Accuracy diagram of the decision tree after removing the anomalous data using the proposed method and DBSCAN method with different neighborhood distance on the sports data set

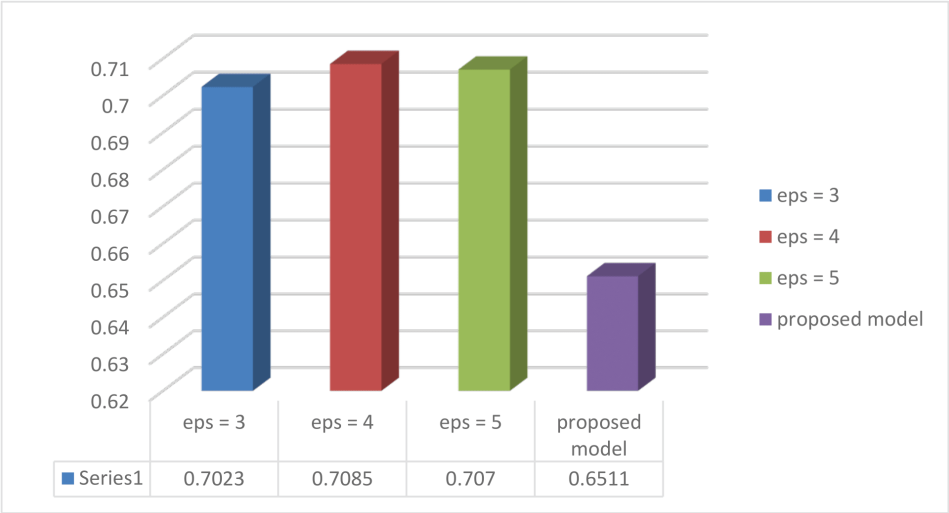

Fig. 8: Accuracy diagram of the Naïve Bayes algorithm after removing the anomalous data using the proposed method and DBSCAN method with different neighborhood distances on the BBC Sport data set

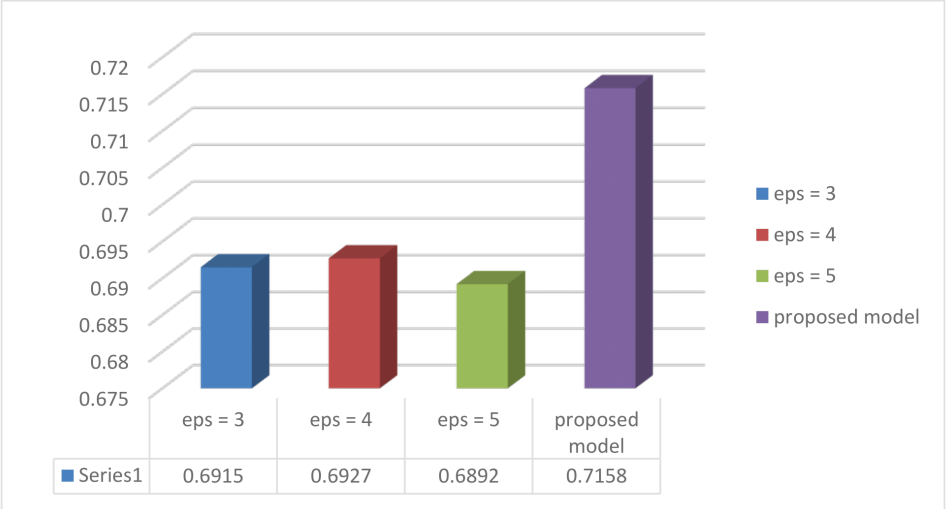

Fig. 9: Accuracy diagram of the decision tree after removing the anomalous data using the proposed method and DBSCAN with different neighborhood distance on the pre-processed BBC Sport data set

news has been improved compared to the non-pre-processed set so that the accuracy of the Naïve Bayes method after eliminating the anomalous data by

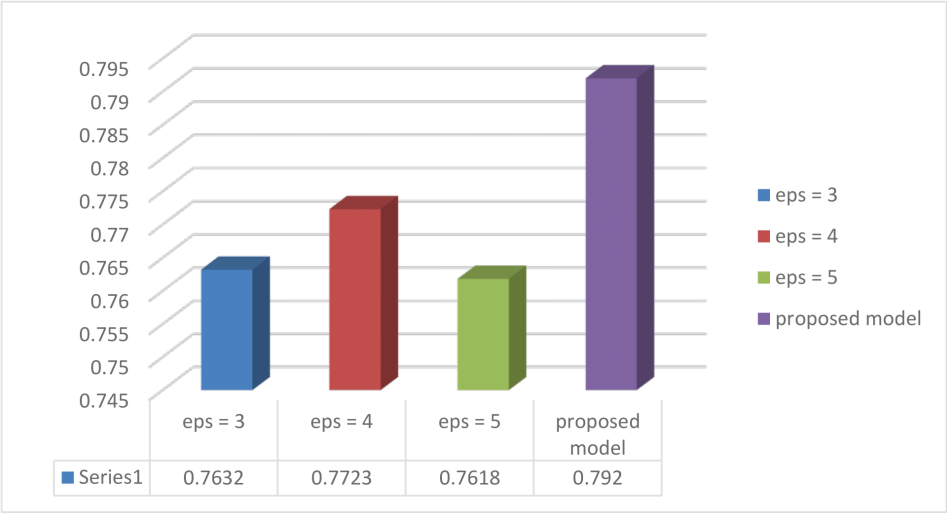

Fig. 10: Accuracy diagram of the Naïve Bayes algorithm after removing the anomalous data using the proposed method and DBSCAN with different neighborhood distance on the pre-processed BBC data set

the proposed method increased compared to the DBSCAN method. Also, the accuracy of decision tree after removal of the anomalous data by the proposed method increased on the pre-processed data set.
Table 2, compares two similar works with our proposed method on the same data set.

Table 2: Results on the BBC News data set

Authors	Descriptions	Accuracy
Zhang et al.	Heterogeneous graph embedding for emerging relation detection from news	64.4%
Greene et al.	Kernel-based ensemble clustering approach	88.0%
Our Proposed Method	A combination of two approaches (clustering-based and distance-based)	92.39%

5 Conclusion and Future Works

In this research, a novel approach for identifying the anomalous text data using unsupervised methods was proposed well. The advantage of using our proposed model as an unsupervised method is that there is no need for prior knowledge and training data. In this research, we assumed that the number

of anomalous data is negligible compared to normal data. The documents studied were also in English and the Cosine Similarity (CS) criterion was used to compare the distance between the documents. Therefore, a given document which is the least similar to others is considered as an anomalous document. In the proposed method, a combination of two approaches clustering-based and distance-based is used for detecting anomaly in the text data. In order to evaluate the efficiency of the proposed approach, this method is applied on four labeled data sets. In general, the obtained results show that the classification accuracy of the documents after applying the anomalous outlier detection algorithm and removing them from the pre-processed data set is always improved and performs well in non-pre-processed data sets.

In order to determine the threshold, our model iteratively runs some algorithms with high mathematical calculation. Consequently computational complexity increases in our approach. Besides, it should be noted that the user must specify k (the number of clusters) at the beginning. So an improved k-means would be used as a solution. We have to repeat the k-means algorithm several times to fix the best clustering and prevent the selection of outliers as initial seeds. Furthermore, we should point out that based on the achieved results the pre-processing step can affect the accuracy. In the future, we are going to use improved k-means algorithm, evaluate other types of clustering algorithms, apply the model to other languages, investigate different distance thresholds (similarity) to tackle these issues.

References

[1] Z. A. Bakar, R. Mohemad, A. Ahmad, and M. M. Deris, "A comparative study for outlier detection techniques in data mining," in *2006 IEEE conference on cybernetics and intelligent systems*. IEEE, 2006, pp. 1–6.

[2] A. Esmaeili Kelishomi, A. Garmabaki, M. Bahaghighat, and J. Dong, "Mobile user indoor-outdoor detection through physical daily activities," *Sensors*, vol. 19, no. 3, p. 511, 2019.

[3] M. Ghorbani, M. Bahaghighat, Q. Xin, and F. Özen, "Convlstmconv network: a deep learning approach for sentiment analysis in cloud computing," *Journal of Cloud Computing*, vol. 9, no. 1, pp. 1–12, 2020.

[4] M. Bahaghighat, L. Akbari, and Q. Xin, "A machine learning-based approach for counting blister cards within drug packages," *IEEE Access*, vol. 7, pp. 83 785–83 796, 2019.

[5] M. Bahaghighat, S. A. Motamedi, and Q. Xin, "Image transmission over cognitive radio networks for smart grid applications," *Applied Sciences*, vol. 9, no. 24, p. 5498, 2019.

[6] F. Abedini, M. Bahaghighat, and M. S'hoyan, "Wind turbine tower detection using feature descriptors and deep learning," *Facta Universitatis, Series: Electronics and Energetics*, vol. 33, no. 1, pp. 133–153, 2019.

[7] M. Bahaghighat, F. Abedini, M. S'hoyan, and A.-J. Molnar, "Vision inspection of bottle caps in drink factories using convolutional neural networks," in *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2019, pp. 381–385.

[8] S. Hasani, M. Bahaghighat, and M. Mirfatahia, "The mediating effect of the brand on the relationship between social network marketing and consumer behavior," *Acta Technica Napocensis*, vol. 60, no. 2, pp. 1–6, 2019.

[9] M. Bahaghighat, Q. Xin, S. A. Motamedi, M. M. Zanjireh, and A. Vacavant, "Estimation of wind turbine angular velocity remotely found on video mining and convolutional neural network," *Applied Sciences*, vol. 10, no. 10, p. 3544, 2020.

[10] M. Bahaghighat and S. A. Motamedi, "Vision inspection and monitoring of wind turbine farms in emerging smart grids," *Facta universitatis-series: Electronics and Energetics*, vol. 31, no. 2, pp. 287–301, 2018.

[11] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[12] J. D. Parmar and J. T. Patel, "Anomaly detection in data mining: A review," *International Journal*, vol. 7, no. 4, 2017.

[13] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," *Egyptian informatics journal*, vol. 17, no. 2, pp. 199–216, 2016.

[14] D. Guthrie, "Unsupervised detection of anomalous text," Ph.D. dissertation, Citeseer, 2008.

[15] J. Samuel, G. Ali, M. Rahman, E. Esawi, Y. Samuel *et al.*, "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, 2020.

[16] S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller *et al.*, "Leveraging data science to combat covid-19: A comprehensive review," 2020.

[17] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining." in *Ldv Forum*, vol. 20, no. 1. Citeseer, 2005, pp. 19–62.

[18] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653–7670, 2014.

[19] A. Mahapatra, N. Srivastava, and J. Srivastava, "Contextual anomaly detection in text data," *Algorithms*, vol. 5, no. 4, pp. 469–489, 2012.

[20] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PloS one*, vol. 11, no. 4, p. e0152173, 2016.

[21] T.-E. Lin, H. Xu, and H. Zhang, "Discovering new intents via constrained deep adaptive clustering with cluster refinement." in *AAAI*, 2020, pp. 8360–8367.

[22] I. Aalto *et al.*, "Discovering topics in slack message streams," 2020.

[23] R. Kannan, H. Woo, C. C. Aggarwal, and H. Park, "Outlier detection for text data," in *Proceedings of the 2017 siam international conference on data mining*. SIAM, 2017, pp. 489–497.

[24] M. T. Sereshki and M. M. Zanjireh, "Outlier detection in text data: An unsupervised method based on text similarity and density peak," 2020.

[25] M. Montes-y Gómez, A. Gelbukh, and A. López-López, "Detecting deviations in text collections: An approach using conceptual graphs," in *Mexican International Conference on Artificial Intelligence*. Springer, 2002, pp. 176–184.

[26] S. Chellamuthu and M. Punithavalli, "Enhanced k-means with greedy algorithm for outlier detection," *International Journal of Advanced Research in Computer Science*, vol. 3, no. 3, 2012.

[27] J. Wang and X. Su, "An improved k-means clustering algorithm," in *2011 IEEE 3rd International Conference on Communication Software and Networks*. IEEE, 2011, pp. 44–46.

[28] M. Marghny and A. I. Taloba, "Outlier detection using improved genetic k-means," *arXiv preprint arXiv:1402.6859*, 2014.

[29] D. Lei, Q. Zhu, J. Chen, H. Lin, and P. Yang, "Automatic k-means clustering algorithm for outlier detection," in *Information engineering and applications*. Springer, 2012, pp. 363–372.

[30] C. Yin and S. Zhang, "Parallel implementing improved k-means applied for image retrieval and anomaly detection," *Multimedia Tools and Applications*, vol. 76, no. 16, pp. 16 911–16 927, 2017.

[31] X.-j. Tong, F.-R. Meng, and Z.-x. Wang, "Optimization to k-means initial cluster centers," *Computer Engineering and Design*, vol. 32, no. 8, pp. 2721–2723, 2011.

[32] J. Zhang, C.-T. Lu, M. Zhou, S. Xie, Y. Chang, and S. Y. Philip, "Heer: Heterogeneous graph embedding for emerging relation detection from news," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 803–812.

[33] D. Greene and P. Cunningham, "Efficient ensemble methods for document clustering," Department of Computer Science, Trinity College Dublin, Tech. Rep., 2006.

[34] J. Manoharan, S. H. Ganesh, and J. Sathiaseelan, "Outlier detection using enhanced k-means clustering algorithm and weight-based center approach," *Int. J. Comput. Sci. Mobile Comput.*, vol. 5, no. 4, pp. 453–464, 2016.

[35] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proc. 23rd International Conference on Machine learning (ICML'06)*. ACM Press, 2006, pp. 377–384.

[36] F. N. Flores and V. P. Moreira, "Assessing the impact of stemming accuracy on information retrieval–a multilingual perspective," *Information Processing & Management*, vol. 52, no. 5, pp. 840–854, 2016.

[37] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of information science*, vol. 18, no. 1, pp. 45–55, 1992.

[38] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[39] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of research and development*, vol. 1, no. 4, pp. 309–317, 1957.

[40] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, 1972.

[41] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," *Journal of documentation*, 2004.

[42] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[43] C. D. Manning, P. Raghavan, and H. Schütze, "Scoring, term weighting and the vector space model," *Introduction to information retrieval*, vol. 100, pp. 2–4, 2008.

[44] K. Church and W. Gale, "Inverse document frequency (idf): A measure of deviations from poisson," in *Natural language processing using very large corpora*. Springer, 1999, pp. 283–295.

[45] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, vol. 4, 2008, pp. 9–56.