I.J. Intelligent Systems and Applications, **2023, 5, 41-56**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2023.05.04

A Hybrid Unsupervised Density-based Approach with Mutual Information for Text Outlier Detection

Ayman H. Tanira*
Computer Science Department, Palestine Technical College, Deir El-Balah, Palestine
E-mail: atanira@ptcdb.edu.ps
ORCID iD: https://orcid.org/0000-0003-4793-9341
*Corresponding author

Wesam M. Ashour
Computer Engineering, Islamic University of Gaza, Gaza, Palestine
E-mail: washour@iugaza.edu.ps

Received: 19 June 2023; Revised: 03 August 2023; Accepted: 27 August 2023; Published: 08 October 2023

Abstract: The detection of outliers in text documents is a highly challenging task, primarily due to the unstructured nature of documents and the curse of dimensionality. Text document outliers refer to text data that deviates from the text found in other documents belonging to the same category. Mining text document outliers has wide applications in various domains, including spam email identification, digital libraries, medical archives, enhancing the performance of web search engines, and cleaning corpora used in document classification. To address the issue of dimensionality, it is crucial to employ feature selection techniques that reduce the large number of features without compromising their representativeness of the domain. In this paper, we propose a hybrid density-based approach that incorporates mutual information for text document outlier detection. The proposed approach utilizes normalized mutual information to identify the most distinct features that characterize the target domain. Subsequently, we customize the well-known density-based local outlier factor algorithm to suit text document datasets. To evaluate the effectiveness of the proposed approach, we conduct experiments on synthetic and real datasets comprising twelve high-dimensional datasets. The results demonstrate that the proposed approach consistently outperforms conventional methods, achieving an average improvement of 5.73% in terms of the AUC metric. These findings highlight the remarkable enhancements achieved by leveraging normalized mutual information in conjunction with a density-based algorithm, particularly in high-dimensional datasets.

Indexed Terms: Text Mining, Text Outliers, Density-based, Mutual Information.

1. Introduction

The outlier detection task is to find a small fraction of data that is different when compared with the rest of the data. Finding outliers from huge data repositories is like finding needles in a haystack. Most existing outlier detection algorithms were designed for mining numeric data which cannot be applied directly to mine outliers from textual documents due to the unstructured content [1]. The dynamic continuous generation of text documents through big systems and current social media applications claims for efficient automated tools to discover and analyze new trends in these data. This information and knowledge can be utilized in many decision support systems. Refining datasets from outlier objects can be very useful to machine learning techniques in increasing their capabilities to capture the emerging trends more accurately.

The vast majority of mining algorithms focus on identifying more repeated objects while discarding the less repeated ones which are candidates to be outliers. In text mining, document outliers contain text data that vary from text resident in other documents belonging to the same domain. Text document outlier detection gains attention due to the availability of a huge amount of text data where recent reports indicate that 95% of the unstructured digital data appears in text form [2]. This availability comes from the greater expansion of social media and Web applications which are rich in text data [3]. Moreover, digitization of news, automation of information centers, and contextualization of user interaction over social media produce larger text collections in their repositories with several topics. We can benefit from text outlier detection methods in these repositories to detect eccentric activities that can be malicious and informative [4]. In addition, data types

on the Web vary from unstructured data (i.e., free text), semi-structured data (e.g., HTML pages), and structured data (e.g., generated tables). But the majority of the Web content is free text making text mining techniques the widely used ones in mining Web content [5].

There are several challenges in the existing applications of detecting text outlier methods. First, the lack of classified data with appropriate labels calls for researchers to focus on developing unsupervised methods. Second, in the free text documents, there are fewer co-occurrences of features between multiple documents and the sparsity of features among documents leads to a small fraction of the features taking non-zero values which prevents traditional document dissimilarity computation from correctly identifying the outliers [3]. Third, short-length posts in social media are characterized by the small number of distinctive common features among posts and the high number of subgroups or topics. These characteristics of social media posts require special kind of methods to identify text outliers that are deviated from all other groups. Fourth, the current big data systems generate larger sizes of text collections which make it a necessity to find efficient but more accurate outlier detection methods.

In text mining, many terms may occur in the document, but they are not related to the subject of the domain. These terms may cause errors in distance computation, and they can be considered noisy data. But when the structure of the data within the context of a certain data position is considered, the importance of a term can be discovered and interpreted. Even more challenging is choosing a subset of features through the high dimensionality of features which required using effective feature selection techniques to produce a representative set of features for a given category. Consequently, text documents need special treatment before applying outlier detection algorithms such as stop word removal, feature generation (e.g., stemming or N-grams feature generation), and dimensionality reduction. The IR systems can manage the text data successfully [6]. For example, search engines, use special efficient methods in determining and retrieving relevant documents to a given query. They are designed with the capability of dealing with huge text document collections [7]. Accordingly, we adopt Vector Space Model (VSM), a well-known scheme in IR, as a representation scheme for outlier detection algorithms.

The significant problem in text document datasets is choosing the most discriminative features effectively and efficiently to define the domain under mining due to the curse of dimensionality. Therefore, it is important to resolve this challenge by addressing the way in which the relevance between extracted features and outliers is measured in an unsupervised way. Therefore, this study aims to improve the effectiveness of text document outlier detection by extracting the most discriminative features for text documents which are considered as high dimensional data. Traditional methods depend on term frequency to identify features that appear more frequently in the corpus, considering them as discriminative features. But term frequency alone does not consider the contextual information of the terms within the documents. It treats all terms equally based on their frequency, without taking into account their importance or relevance to the overall document. Our hypothesis is that features in text documents belong to the same domain are more related one another and consequently they have higher correlation. Therefore, when an appropriate feature selection method is used, we can extract more informative features and ignore the irrelevant features (i.e., noisy terms) which results in improving the effectiveness of the outlier detection algorithm. Accordingly, this work proposes the use of mutual information as a feature selector to determine the related documents and then apply a powerful density-based algorithm, LOF, for calculating the outlier factor and detecting the document outliers based on the domain under mining.

The main contributions of this paper are as follows:

- Presenting a comprehensive survey of outlier detection techniques including the advantages and disadvantages of each category.
- Introducing a theoretical background of density-based technique and customizing a local outlier factor algorithm for mining text document outliers.
- Proposing a hybrid unsupervised density-based technique with mutual information to cope with the dimensionality problem of the generated features and detect text document outliers. The detected document outliers contain text data varying from text found in other documents taken from the same domain.
- Conducting an extensive set of experiments using twelve datasets to measure the performance of the proposed system and compare the proposed technique with the traditional method. Analysis of the experimental results indicates that the proposed method outperforms the traditional method in terms of the ROC AUC metric.

The rest of the paper is organized as follows. Section 2 presents previous work on outlier detection techniques. Section 3 introduces the proposed framework and methodology used in this study. The experimental results and discussions are reported in Section 4. Finally, Section 5 concludes the paper.

2. Related Works

In recent years, most users of the Web interact with others using free text through emails, blogs, and social media like Facebook and Twitter. Thus, it is useful for tracking to find and analyze parts of the text that may be interesting or suspicious inside these collections containing different topics [4]. A sparse representation usually results due to the high dimensionality of terms contained in text collections [6]. Different text representation schemes based on the well-known method, term-frequency inverse document-frequency (tf_idf), from IR have been used with the VSM [8].

Based on VSM representation, many researchers employ data mining outlier detection techniques which are supervised and unsupervised approaches. Recently supervised and semi-supervised methods are proposed. These methods use training datasets to learn a classifier directly or indirectly to predict outliers such as Neural Network (NN)-based methods that use deep feature extraction, and Generative Adversarial Network (GAN)-based active learning methods in detecting outliers [9, 10]. The performance of supervised methods depends on the training dataset quality and the capabilities of the classifier.

In the unsupervised category, there are three main types: distance-based, density-based, and cluster-based. Most of these methods utilize proximity methods in identifying outliers. The distance-based methods compute the distance between each data point and other ones and identify as outliers those data points with larger distances [11]. A well-known approach, *k*-nearest neighbor, calculates the distance of each object's nearest neighbors and marks objects with the highest distances as outliers [12]. The main advantages of the distance-based techniques are that: firstly, they do not require a priori knowledge of the data distribution to determine outliers. Secondly, the parameters of the algorithms have clear meaning and are easy to choose [13]. Moreover, the results of the detected outliers are easy to interpret. In addition, they are applicable for multidimensional datasets [1]. Finally, the distance-based techniques can be applied to any feature space for which a distance metric can be defined [13]. On the other hand, they suffer from some difficulties: firstly, the majority of distance-based algorithms have quadratic complexity, so they are time-consuming. The performance can be improved by using assistant techniques such as the pruning rule to eliminate the non-outlier data objects [12]. The second difficulty is associated with the fact that the majority of modern systems contain heterogeneous data of complex structures where the definition of distance on sets of such data is a nontrivial problem [13]. Finally, they depend on a priori given parameters such as the distance, D, or the number of neighbors, k. If these parameters changed, the model is needed to be reconstructed.

The authors in [14] introduce a new proximity-based technique, *kj*-Nearest Neighbors (*kj*-NN). The proposed technique estimates the original labels of the text data objects by adopting semantic similarities and a self-supervision method in detecting text outliers. The researchers in [15] offer an algorithm for outlier detection in categorical data and define each object by a set of feature vectors namely a matrix-object. They propose their idea by using the well-known notions in software engineering; cohesion and coupling. The coupling of a matrix-object is the computed average distance with other matrix-objects, and the cohesion of a matrix-object is defined based on information entropy. The method in [16] computes the similarity of every text document to the remaining documents in the collection and identifies insufficiently similar documents as outliers. The authors recommend that for multiple similarity measures and based on authorship verification techniques, it is more efficient to use two second-order measure than to use the corresponding original first-order measure.

Density-based techniques take into account the density of the neighbors surrounding each data object. These methods compute a local outlier factor (LOF) for each object which describes the degree of the anomaly with respect to the neighbors The main advantage of density-based techniques is that they take into account the density of an object with respect to local neighborhoods. Most of these techniques rely on a distance metric to compute the distance between each object and its surrounding neighbors. Thus, they are applicable to any feature space for which a distance metric can be defined. Finally, the density-based techniques can identify meaningful local outliers that other approaches cannot find [17]. The main disadvantage of density-based techniques is that they are considered time-consuming where most of these techniques compute outlying factor for each object based on previously computed equations. Another problem regards parameter settings such as the number of neighbors, k, which tend to affect their performance if chosen wrongly. Finally, we need a meaningful concept of distance for sparse high-dimensional data. If this does not exist, then the detected outliers are unlikely to be very useful. Although, LOF algorithm has some limitations, it is considered a powerful outlier detection algorithm in terms of accuracy. Therefore, this work adopts LOF algorithm to compute the outlier factor for each document.

The authors in [18] propose a density-based technique for detecting text data outliers from the text corpus. They use the *n*-grams technique to generate the representative features, calculate the similarity between all pairs of documents and use the LOF algorithm to detect the top outlier documents. The proposed method suffers from the high dimensionality of the generated features due to the use of *n*-grams. The researchers in [19] present a new technique that estimates a local kernel density for each object. They identify the degree of the anomaly by considering three types of neighbors: *k*-nearest, reverse nearest, and shared nearest, for the purpose of estimating local kernel density. They introduce an effective density-based method named Relative Density-based Outlier Score (RDOS) which calculates the score of anomaly of each object. The authors in [20] concentrate on the concept of relative neighborhood space which considers the structure of the local neighborhood of an object. Accordingly, they propose a novel algorithm that is capable of detecting local and global outlier objects at the same time. It can identify objects with low-density as outliers where most other algorithms cannot. The researchers in [4] investigate text outliers based on ranking concepts in IR. They utilize the idea of ensemble algorithms to introduce a novel ensemble algorithm. The proposed algorithm computes the frequency and then ranks the text objects concerning NNs and the local sub-density of the neighbors to identify the text outliers.

Some researchers tend to use cluster-based methods in outlier detection. These methods are particularly designed to group objects with shared characteristics into clusters however, they can detect outlier clusters. In addition, most of these methods depend on distance threshold parameters [21]. A key advantage of the clustering-based techniques is that they generate clusters addition to the outliers they produce as a by-product, so the miner can benefit from the produced clusters. Another important advantage is that some clustering algorithms scale to large real datasets. On the other hand, the notions of outliers in the clustering-based techniques are essentially binary, and there is no qualification as to how outlying an object is [17]. They do not provide any means of ranking the detected outliers. Also, the clustering algorithms do not

distinguish between noise and outliers. Finally, the clustering algorithms identify any data objects falling outside the produced clusters as outliers, which may not necessarily be true in a data mining context. In the case of text outliers, more preprocessing operations are needed before applying cluster-based methods.

The authors in [22] propose a novel cluster-based algorithm for detecting outliers. The proposed algorithm utilizes the *k*-means clustering method. Firstly, it computes the data center of each cluster. Secondly, it builds a dataset model by identifying the distance threshold to the center of each cluster. Thirdly, it optimizes the neighbor distribution of objects. The researcher in [23] introduces a fuzzy clustering approach to detect outlier documents. The main idea is that each document has a degree to be a candidate outlier. This assumption is considered for each document in the input collection. During the fuzzy clustering process, each document is assigned to each cluster with some degree of membership. Documents with very close degrees of membership to different clusters are candidates to be outliers. The proposed algorithm recomputes the objective function of each candidate document and the algorithm marks outliers those documents that increase the score of the objective function.

Other researchers employ other methods for text outlier detection. The authors in [3] utilize a low-rank approximation method to mark the outliers of the given text dataset. The researchers propose a matrix factorization algorithm that is based on Block Coordinate Descent (BCD) framework. The proposed technique processes iteratively the matrix representing the dataset to detect text outliers. The experimental results show important advantages of their algorithm over other algorithms in detecting text outliers. The researchers in [24] develop a character-level representation of text as numerical features for the outlier detection model. The proposed approach appears to be flexible enough for business purposes and allowable to use on a wide variety of unlabeled textual data sets without prior model training and tuning.

Some techniques have been proposed to cope with the curse of dimensionality. The goal is to choose a set of features that are more representative of the domain under mining and eliminate all other features. Mutual Information (MI) is one of the attractive methods that is used to solve this problem [25, 26]. But MI is a computationally expensive method, so some researchers estimate MI by using approximation methods. MI. The authors in [26] introduce a novel method, minimal Redundancy Maximal Relevance (mRMR) and they prove that mRMR is equivalent to Max-Dependency while other researchers propose a Normalized Mutual Information Feature Selection (NMIFS) method to improve the results [27].

Recently, several researchers utilized Deep Reinforcement Learning (DRL) algorithms for the task of outlier detection. Most of the recent DRL-based methods are supervised and partially supervised. Researchers in [28] propose an approach based on A3C RL with an adaptable deep neural network. For example, CNNs are fit for image tasks and RNNs are for sequential tasks traditionally, so the attention mechanism is considered as the actor. PPO-based anomaly detector is proposed in [29] named Meta Active Anomaly Detection (Meta-AAD) to balance short-term and long-term performance, which benefits the anomaly detector in the long term. The proposed method extracts transferable meta-features to make the extracted policy transferable between different datasets. Researchers in [30] introduce a DQN-based anomaly detector, Deep Q-learning with Partially Labeled Anomalies (DPLAN). They consider the problem of anomaly detection with a small set of partially labeled anomaly examples and a large-scale unlabeled dataset. Thus, the proposed method aims to introduce a model resulting in joint optimization of the detection of known and unknown anomalies. A generic policy-based RL framework is proposed in [31] to address the time series anomaly detection problem. The policy-based time series anomaly detector (PTAD) is progressively learned from the interactions with time-series data in the absence of constraints. The study adopts A3C to benefit from multi-workers and achieve the best performance. Experimental results show that it outperforms the value-based (DQN) temporal anomaly detector.

3. The Proposed Method

In this section, we present the proposed approach for detecting outliers in text documents, aiming to identify the most outlying document within the given corpus. Firstly, we define the problem of detecting outliers in text documents in a formal manner. Subsequently, we describe the key elements of the proposed system, which encompasses the underlying mathematical model

3.1. Formalization

The input corpus consists of interrelated text documents grouped into different categories depending on their contents. Some text documents are founded in a certain category, but their contents are irrelevant to the parent category. These text documents are called text document outliers. It is useful to have effective ways of detecting outlying text documents contained in some given category of interest to a user. Therefore, given a collection of text documents $D = (D_1, D_2, \dots, D_{|D|})$ belonging to a particular category where each document consists of a set of terms $D_i = \{T_1, T_2, \dots, T_{|D_i|}\}$, we need to detect the text document outliers $O = (O_1, O_2, \dots, O_{|O|})$ in the given category, where $O \subset$ D and $|O| << |D|$. The detected text documents as outliers will be the most outlying documents in the input collection.

The proposed method is divided into five main phases: Document Extraction, Features Extraction, Features Selection, Term-Weighting and Proximity Matrix Computation, and Outlier Detection as in Fig. 1. We summarize all notations used in this research in Table 1 for the readers' convenience.

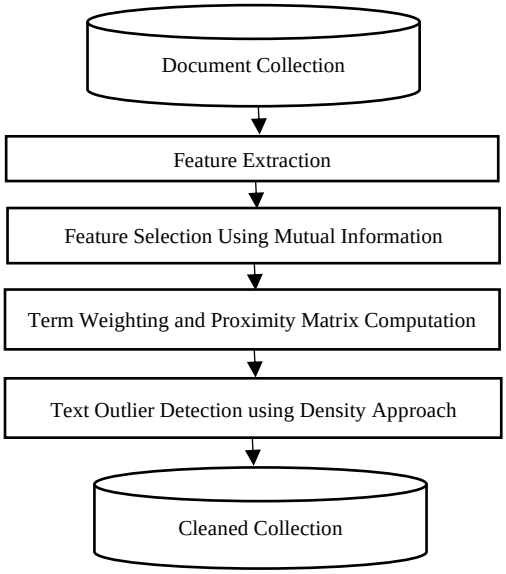

Fig.1. The proposed method

3.2. Document Extraction

The goal is to provide a collection of text documents sharing the same topic (e.g., health, computer, sport). We can use repositories of organizations such as universities or corporations which contain a large documents collection. One can use any of the existing tools (e.g., search engines or Web agents) to retrieve the required corpus. The extracted documents are refined by selecting textual sections (i.e., ignore any other data types such as images or audio), eliminating stop words, tokenizing each document into a set of tokens, converting tokens to the same case, stemming each token, and finally removing stop words because they don't have any useful information in text mining [32].

Table 1. Notations used in the paper

Notation	Description	Notation	Description
α	Frequency weight	I	Mutual information
S	Section	$P(x, y)$	Joint probability function of X and Y
T	Feature	$P_X(X)$	Marginal probability of X
w	Weight of a feature in a section	$\vec{D}$	Document vector
C	Count of a feature	Dis	Dissimilarity
D	Document	O	Document outlier
tf	Term frequency	k	Number of nearest neighbors
N	Number of documents in corpus	$k\text{-}dis$	k-dissimilarity
n	Number of documents contain T	ldd	Local document dissimilarity
W	Overall weight of a feature	$DLOF$	Document local outlier factor

3.3. Feature Extraction

Each document is conducted and all features in the document are extracted. The system counts the frequency of each feature and weights the frequency according to the section it belongs to as follows:

$$w(T_{kit}) = \begin{cases} \alpha & if\ S_t \in improrant\ sections, \alpha > 1 \\ 1 & otherwise \end{cases} \quad (1)$$

where, $w(T_{kit})$ is the weight of feature T_k in document D_i in section S_t. If the feature T_k recurs in different sections, then the frequency is given as:

$$tf_{ik} = \sum_{part} C_{ik} * \ w(T_{kit}) \quad (2)$$

where, tf_{ik} is the frequency of T_k which has a count C_{ik}in each section S_t within document D_i. This can be effective when some sections have importance more than other sections of the document (e. g. subject tag in HTML pages).

3.4. Feature Selection

In this phase, we aim to select the more representative subset of features, so the extracted features are refined according to some criteria for eliminating unrepresentative features [25]. The proposed method utilizes a Vector Space Model (VSM) as a representation scheme for the outlier detection algorithm. VSM contains all extracted features from all documents. The generated vector suffers from high dimensionality so we need to cope with this problem. The proposed system adopts mutual information as a filter method to selectively choose the most representative features of the given category. Let X and Y be two random variables with discrete values, the mutual information $I(X;\ Y)$ between X and Y is defined as:

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} P_{X,Y}(x,y) log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \tag{3}$$

$$P_X(x) = \sum_{y\in Y} P_{X,Y}(x,y) \tag{4}$$

$$P_Y(y) = \sum_{x\in X} P_{X,Y}(x,y) \tag{5}$$

where $P_{X,Y}(x,y)$ is the joint probability function of X and Y, $P_X(x)$ and $P_Y(y)$ are the marginal probability functions of X and Y respectively.

It is clear that this measure is non-negative $I(X;\ Y) \geq 0$ and symmetric $I(X,Y) = I(Y,X)$. In this research, we use a Normalized Mutual Information Feature Selection (NMIFS) method which was introduced in [27]. So, the selection criterion utilizes the average normalized MI between each feature in the vector and all other given features. Finally, the general vector is generated which represents the domain under the mining process and each document is represented as a vector with the same length of the general vector containing the frequency (F_i) of each feature (T_i) in that document (D_i) with respect to the general vector of VSM.

3.5. Term-weighting and Proximity Matrix Computation

The general vector of VSM consists of n features, $\{T_1, T_2, T_3, \dots, T_n\}$. Each feature T_k in each document D_i is weighted by tf_idf formula obtaining W_{ik}, hence, each document $D_i$becomes a vector of weighted features:

$$D_i = (W_{i1}, W_{i2}, W_{i3}, \dots, W_{in}) \tag{6}$$

where $W_{ik} > 0$ if T_k occurs in D_i, and $W_{ik} = 0$ otherwise.

We use tf_idf formula in obtaining W_{ik} as follows:

$$W_{ik} = \frac{tf_{ik} * \log(N/n_k)}{\sqrt{\sum_i (tf_{ik})^2 * \left(log\left(N/n_k\right)\right)^2}} \tag{7}$$

where tf_{ik} represents the frequency of T_k in D_i, N is the number of documents in the input collection, and n_k is the number of documents where T_k occurs. The formula is normalized to the interval [0, 1] by division on the denominator in (7).

The proximity matrix (PM) computation module is the last step before the outlier detection process. It aims to measure the dissimilarity (distance) between documents by using certain dissimilarity metrics. There are several metrics in the literature of information retrieval that can be used to compute distances between document vectors such as the Cosine metric, Euclidean metric, Manhattan metric, or Canberra metric. We use the Cosine formula to compute the dissimilarity between every pair of documents in the input collection. Cosine formula measures the angle between two document vectors $\overrightarrow{D_i}$ and $\overrightarrow{D_j}$ as follows:

$$Dis\left(\overrightarrow{D_i}, \overrightarrow{D_j}\right) = 1 - \left[\frac{\sum_k W_{ik} W_{jk}}{\sqrt{\sum_k {W_{ik}}^2}\sqrt{\sum_k {W_{jk}}^2}}\right] \tag{8}$$

where W_{ik} represents the weight of T_k in $\overrightarrow{D_i}$ and W_{jk} represents the weight of T_k in $\overrightarrow{D_j}$. The formula expresses that the closer to one the high dissimilar and the farther to one the lower dissimilar. The result of these computations is the proximity matrix.

Algorithm *nmiDLOF*

Input: Document Collection (D_n) No. of Outliers K
Output: K outlier documents

1. for each document $D_i \in D_n$
   extract each feature T_k
   count T_k w.r.t. eq.1 and eq. 2
2. generate the general feature vector $F = \{T_{ik}: \forall\, T_i \in D_k\}$
3. for each feature $T_k \in F$
   compute NMI $\forall\, T_k \in F$ using Eq. 3
5. generate F' with m features with high NMI
6. for each document $D_i \in D_n$
   compute $W_{ik}\ \forall\, T_i \in d_k$ w.r.t. F' using eq. 7
7. compute PM of D_n
8. for each document $D_i \in D_n$
   compute $DLOF$ using eq. 9 and eq. 10
9. output K documents with the highest $DLOF$

3.6. Text Outlier Detection

It is the process of applying the outlier detection technique on PM for identifying text outlier documents. LOF algorithm is introduced by Breunig et al. [17] where the density of a certain object (document) has an important role and there is no explicit classification of whether each object is an outlier or not. The algorithm computes a Local Outlier Factor (LOF) for each object which indicates how strongly it can be considered an outlier. The LOF is a density-based algorithm in the literature of structured data mining. It computes the local density of each object concerning *k*-nearest neighbors through multiple steps. In the case of free-text documents, more preprocessing operations are needed to prepare the input into an applicable form that is appropriate for using data mining approaches. Here the proximity matrix is the input to the density algorithm. In our case, we use the LOF algorithm on the computed proximity matrix for text document outlier detection. For the sake of formalization LOF algorithm, we introduce the following definitions where the target dataset is a collection of documents:

Definition 1: *Outlier Documents*
Given a collection of documents $D = (D_1, D_2, D_3, \dots, D_n)$ belonging to a particular category, we need to detect the document outliers $O = (O_1, O_2, \dots, O_k)$ in the given category, where $O \subset D$ and $k << n$. The detected Web documents as outliers will be the most outlying documents in the input collection concerning the textual data found in documents.

Definition 2: *k_dissimilarity of a document D_i*
The *k*-dissimilarity of a document D_i, designated by *k*-dis(D_i), is defined as the dissimilarity $dis(D_i, D_j)$ between D_i and a document D_j such that:

- for at least *k* documents $D'_j \in D - \{D_i\}$ it holds that $dis(D_i, D'_j) \leq dis(D_i, D_j)$, and
- for at most *k-1* documents $D'_j \in D - \{D_i\}$ it holds that $dis(D_i, D'_j) < dis(D_i, D_j)$.

Definition 3: *local document density*
The local document density of a document D_i, designated by $ldd_k(D_i)$, is defined as the inversion of the average reachability dissimilarity of the *k*-NNs of D_i and given by:

$$ldd_k(D_i) = \frac{1}{\frac{(\sum_{y \in N_k(x)} max\{k\_dis(d_j), dis(d_i, d_j)\}}{|N_k(x)|}} \tag{9}$$

where $N_k(D_i)$is the *k*-NNs of a document D_i.

Definition 4: *Document Local Outlier Factor*
The document local outlier factor of D_i, designated by $DLOF_k(D_i)$, is defined as:

$$DLOF_k(D_i) = \frac{\sum_{y \in N_k(D_i)} \frac{ldd_k(D_j)}{ldd_k(D_i)}}{|N_k(D_i)|} \tag{10}$$

The computed $DLOF_k(D_i)$determines the amount of anomaly of a document D_i. It is the average ratio of the local document density with respect to k-NN documents. Equation (10) expresses that the lower local document density of D_i's is, and the higher the local document densities of d_i's k-NN are, the higher is the $DLOF$ value of D_i. When DLOF is close to 1 means that the document locates inside a cluster deeply and it is not an outlier document.

We summarize all the aforementioned steps in the above algorithm which is named $nmiDLOF$; normalized mutual information Document Local Outlier Factor.

4. Experimental Results and Discussions

This section illustrates our experiments on text document outlier analysis using the proposed method; $nmiDLOF$. We compare the proposed method with a conventional method of density-based LOF algorithm which relies on term-frequency inverse document-frequency (tf_idf) in a feature selection process which is denoted with $tfLOF$ in the evaluation process. Firstly, the description of the datasets is introduced. Secondly, we define the metrics that are used in the evaluation process. Thirdly, we present and discuss the results of the experiments. The proposed algorithm was implemented in Python 10 on a system with an Intel CORE i7, 8GB RAM, and Windows 10 operating system.

4.1. Datasets

We use two high-dimensional collections of datasets to measure the performance of the proposed method, $nmiDLOF$, and compare with conventional method, $tfLOF$. The first collection is considered synthetic and small collection while the second one is considered real collection and contains larger number of documents. It should be pointed out that these collections are not originally designed for outlier analysis, consequently, we prepare them as in [3].

- **Newsgroup collection:** this collection contains ten newsgroup datasets and they have been prepared for research in text mining. They are appropriate for several ML algorithms such as classification and clustering of text documents. These datasets are available on [33]. We prepared six different datasets to be used in our experiments including Food, Sports, Business, Graphics, Space and Historical which are described in Table 2. In every experiment of Newsgroup collection, we remove features with a frequency less than 2 and set the number of neighbors (k) equal to the best result achieved by trying values from 20 to 50. We set the weight of section in (1) equal to 1 because the documents do not contain distinguishable sections (i.e., $\alpha = 1$).
- **Reuters-21578 collection:** this collection contains a huge number of text documents that are indexed and classified into a set of categories and uploaded on UCI machine learning repository [34]. Several researches utilized this collection in classification and clustering techniques of text documents [3, 35]. We use six datasets of Reuters-21578 collection including Atheism, Christian, Electronics, Hockey, Graphics and Politics which are illustrated in Table 2. In every experiment of Reuters-21578 collection, we remove features with a frequency less than 5 and set the number of neighbors (k) equal to the best result achieved by trying values from 200 to 500 and we set $\alpha = 1$.

Table 2. Description of the newsgroup and reuters-21578 datasets

Collection	Dataset	Number of Documents	Number of Outliers	No. of features
Newsgroup	Food	111	11	664
	Sports	118	18	804
	Business	118	18	918
	Graphics	119	19	931
	Space	110	10	1416
	Historical	114	14	1965
Reuters-21578	Atheism	1000	135	4361
	Christian	996	128	4532
	Electronics	1000	130	4808
	Hockey	998	114	5438
	Graphics	1003	168	5836
	Politics	1006	182	6914

The proposed system prepares each document involving lower case conversion, stemming, and stop-words removal. Once the preprocessing phase is completed the general feature vector is generated. The last column in the above tables explains the number of features extracted from each dataset. The system uses these features to generate the representative feature vector by the Normalized Mutual Information technique (NMI).

4.2. Evaluation Metrics

The effectiveness of text document outlier detection techniques is evaluated in terms of the ROC (Receiver Operating Characteristics) curve. We use the AUC (Area Under the Curve) metric to evaluate the proposed method, $nmiDLOF$ against the conventional method, $tfLOF$ and to analyze the results. AUC-ROC is one of the most important metrics in evaluating the performance of the classification models. It measures the capability of the model in distinguishing between classes (in our case two classes; outlier and normal). The higher the AUC, the better technique is at classifying the outlier text document and the normal text documents. It plots the true positive rate (TPR) against the false positive rate (FPR). The TPR is computed as the recall in IR while the FPR is the fraction of the falsely detected positive text documents. Based on the confusion matrix in machine learning the definitions of TPR and FPR are as follows:

$$TPR = \frac{TP}{TP+FN} \tag{11}$$

$$FPR = 1 - \frac{TP}{TP+FP} \tag{12}$$

where TP is True Positive, FP is False Positive, TN is True Negative and FN is False Negative.

We also use the running time to evaluate the performance of text outlier detection techniques and measure the effects of mutual information method on the detection process.

4.3. Results and Discussions

Table 3 explains the results achieved over the twelve datasets with respect to the evaluation metrics: the AUC and the execution time. The ROC curves and the AUC results on Newsgroup dataset are depicted in Figures 2 to 7 while the ROC curves and the AUC results on Reuters-21578 datasets are depicted in Figures 8 to 13. The $nmiDLOF$ method scores the best AUC values over all datasets used. Firstly, let's consider the AUC values on the Newsgroup datasets: Food, Sports, Business, Graphics, Space, and Historical. The proposed method, $nmiDLOF$, achieved the values of 0.9430, 0.9972, 0.9817, 0.8684, 0.8040, and 0.9946 respectively while the conventional method, $tfLOF$, achieved the values of 0.7603, 0.9966, 0.8301, 0.8610, 0.8637, and 0.9089 respectively. Secondly, on the Reuters-21578 datasets: Atheism, Christian, Electronics, Hockey, Graphics, and Politics, the same scenario is repeated where $nmiDLOF$ produced the best values of 0.7603, 0.9966, 0.8301, 0.8610, 0.8637 and 0.9089 respectively, while $tfLOF$ produced the values of 0.6558, 0.9476, 0.7832, 0.8246, 0.7866 and 0.8327 respectively. Finally, as a result of these values, the proposed method, $nmiDLOF$, outperforms the conventional method, $tfLOF$, on average of 5.73% of AUC scores which is considered a remarkable improvement and asserts that superiority of $nmiDLOF$ over $tfLOF$.

Table 3. Results of $tfLOF$ and $nmiDLOF$ on all datasets

Collection	Dataset	Running Time (ms)		AUC	
		$tfLOF$	$nmiDLOF$	$tfLOF$	$nmiDLOF$
Newsgroup	Food	**458**	792	0.9060	**0.9430**
	Sports	**474**	833	0.9268	**0.9972**
	Business	**570**	984	0.9167	**0.9817**
	Graphics	**563**	1106	0.8184	**0.8684**
	Space	**904**	1631	0.7460	**0.8040**
	Historical	**1575**	2304	0.9769	**0.9946**
Reuters-21578	Atheism	**16684**	33509	0.6558	**0.7603**
	Christian	**16766**	35899	0.9476	**0.9966**
	Electronics	**17884**	37043	0.7832	**0.8301**
	Hockey	**19087**	38634	0.8246	**0.8610**
	Graphics	**21016**	44893	0.7866	**0.8637**
	Politics	**25390**	49789	0.8327	**0.9089**

The obtained results consistently demonstrate the superior performance of the proposed method, $nmiDLOF$, compared to the conventional method, $tfLOF$. This superiority arises from the implementation of Normalized Mutual Information (NMI) in the selection of features, which facilitates the generation of a more representative general feature vector that effectively characterizes the mining process within the domain. Additionally, the proposed method leverages the LOF algorithm and mutual information to its advantage. By considering the frequency of occurrences for each pair of extracted features, the method is able to identify meaningful outliers within text documents. This validation of the proposed technique's superiority can be attributed to the distinctive qualities of the mutual information measure. Unlike other dependency metrics, mutual information offers the ability to quantify the relationship between any two features. Furthermore, the mutual information measure remains invariant when the general feature vector undergoes transformations.

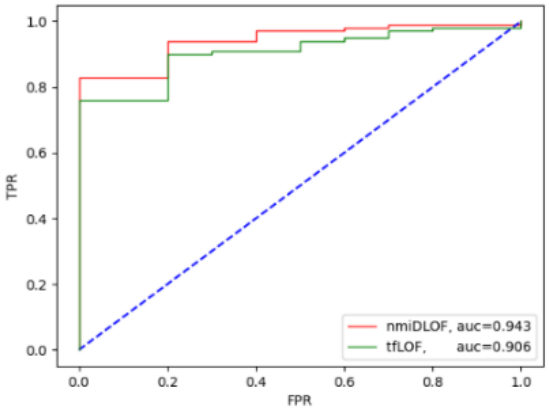

Fig.2. Newsgroup, food

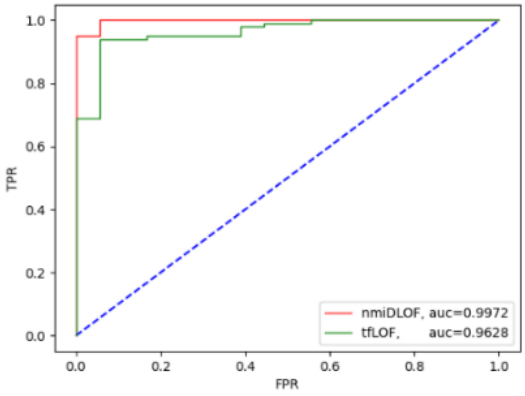

Fig.3. Newsgroup, sports

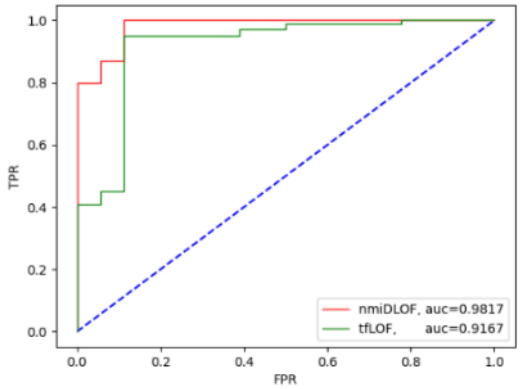

Fig.4. Newsgroup, business

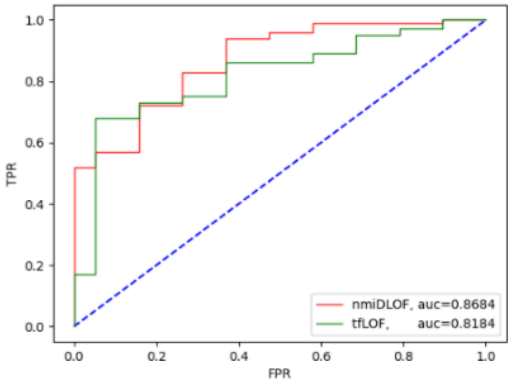

Fig.5. Newsgroup, graphics

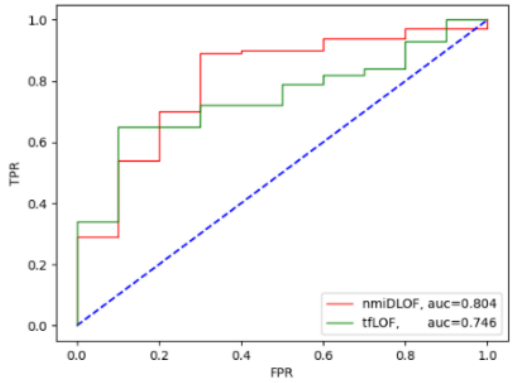

Fig.6. Newsgroup, space

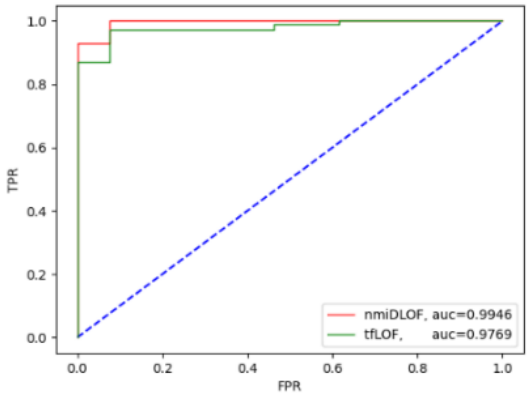

Fig.7. Newsgroup, historical

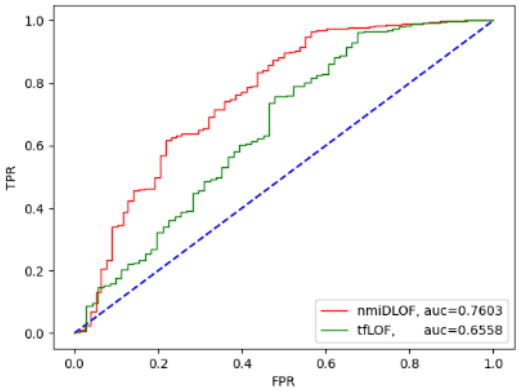

Fig.8. Reuters-21578, atheism

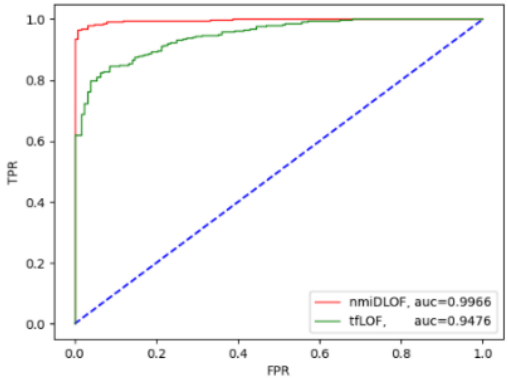

Fig.9. Reurters-21578, christian

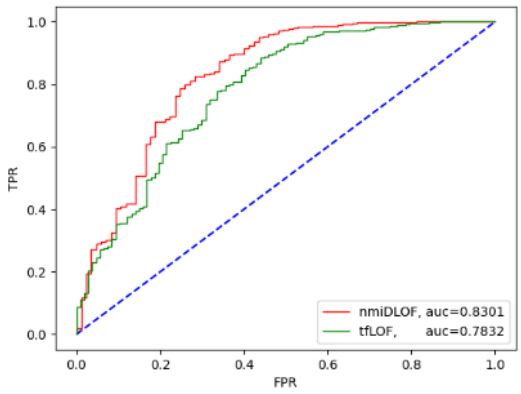

Fig.10. Reuters-21578, electronics

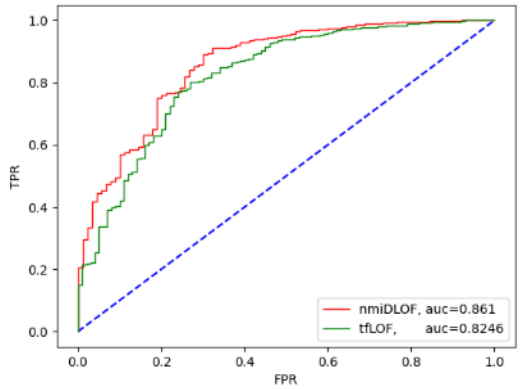

Fig.11. Reuters-21578, hockey

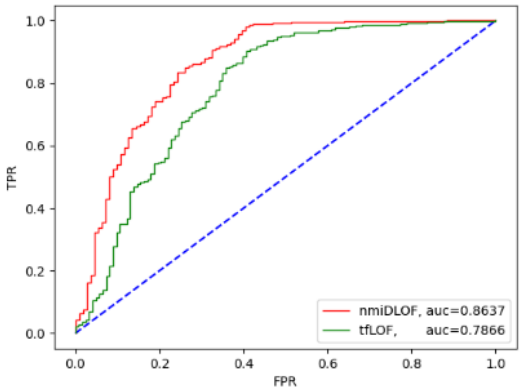

Fig.12. Reuters-21578, graphics

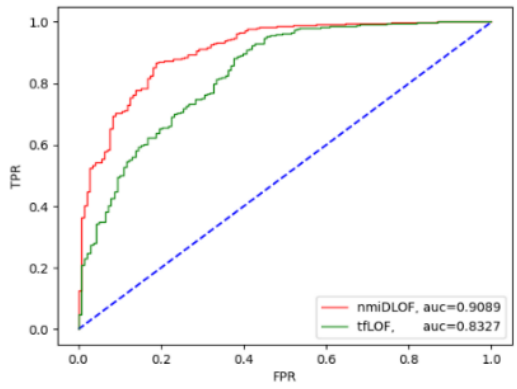

Fig.13. Reuters-21578, politics

Consequently, the use of mutual information allows for the calculation of shared information between pairs of features within the feature space. When two features exhibit a high degree of correlation and share information, it becomes possible to substitute the two features with a single representative feature. Conversely, if two features are entirely independent, the computed mutual information will be zero, indicating the absence of shared information. This characteristic of mutual information extraction implicitly incorporates contextual information, leading to the creation of a more representative general feature vector that accurately characterizes the mining process within the domain. By employing mutual information as a feature selection method, the proposed technique enables the LOF algorithm to effectively differentiate between normal and outlier text documents. Consequently, the utilization of mutual information enhances the accuracy of the computed outlier factors for input text documents, resulting in the superior performance of the proposed technique compared to the conventional method.

The evaluation process incorporates the consideration of running time as a second criterion. To analyze this aspect, Table 3, Fig. 14 and Fig.15 present the running time comparison between the two methods across all datasets. The utilized datasets are characterized by their high dimensionality, containing a substantial number of generated features (hundreds for the Newsgroup corpus and thousands for the Reuters-21578 corpus). The results clearly demonstrate that the proposed method, $nmiDLOF$, requires a longer execution time compared to the $tfLOF$ method. This discrepancy arises from the computational requirements associated with calculating mutual information across the generated features. During the general vector phase, the proposed method computes the occurrences of each feature pair across all documents within the given corpus, which leads to increased execution time. It is worth noting that the running time is further influenced by the number of features generated. As the quantity of features increases, the execution time of the method proportionally rises. However, it is important to emphasize that in the mining process, effectiveness holds greater significance than running time. Thus, the increased execution time of the proposed method can be justified by its superior performance in terms of accuracy and representation.

In future research, efforts will be directed towards minimizing the running time of $nmiDLOF$. This could involve exploring alternative approaches such as the utilization of estimation methods for mutual information. By employing these estimation techniques, it may be possible to reduce the computational burden and alleviate the running time constraints associated with $nmiDLOF$, without compromising its effectiveness in outlier detection within text mining processes.

5. Conclusions

In this paper, we introduce a hybrid unsupervised method for detecting text document outliers which addresses the main challenge in domains with sparse high-dimensional nature such as text mining. The proposed framework consists of several phases: document extraction, feature extraction, feature selection, term-weighting and proximity matrix representation, and finally text outlier detection. The distinguishing characteristic of the proposed method is that it merges two robust techniques. First, it utilizes a normalized unsupervised mutual information method to select subset features that are more representative of the domain under the mining process and mitigate the impact of high dimensionality and sparsity. The selected subset of features is used to represent document collection for the outlier detection algorithm. Secondly, the proposed method adopts a famous density-based technique, LOF, for detecting outlier documents according to the computed document local outlier factor which computes the amount of anomaly of each document with respect to *k*-NN documents. Moreover, the proposed method has the capability to be used in other fields such as Web content outliers where most data contained is text. The experimental results on twelve synthetic and real datasets with respect to the ROC AUC metric showed that the proposed technique, $nmiDLOF$, overcomes the $tfLOF$ technique with a remarkable improvement in terms of accuracy. The reason is that the proposed technique can construct a vector of features that characterizes the domain under mining more accurately due to the use of Normalized Mutual Information in selecting features and generating a more representative general feature vector that represents the domain under mining. Moreover, The proposed method takes advantage of the LOF algorithm and mutual information. The $nmiDLOF$ technique considers the number of occurrences of each pair of the extracted features which produces meaningful text document outliers. On the other hand, the $tfLOF$ method takes into account the whole dimensional space in searching for outliers which causes some document outliers may be detected as normal documents. Although the proposed technique outperformed the conventional method in terms of effectiveness, the comparative analysis showed that $nmiDLOF$ technique consumes more execution time than $tfLOF$ technique due to the computation of mutual information across the generated features from all documents in the input corpus. Thus, in the future we will consider how to eliminate the running time of $nmiDLOF$ such as using one of the estimation methods of MI.

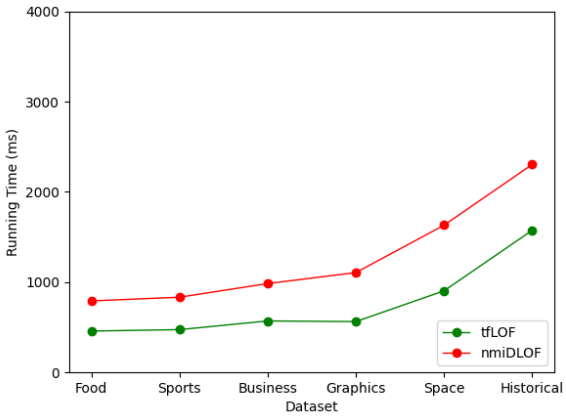

Fig.14. Running time on newsgroup collection

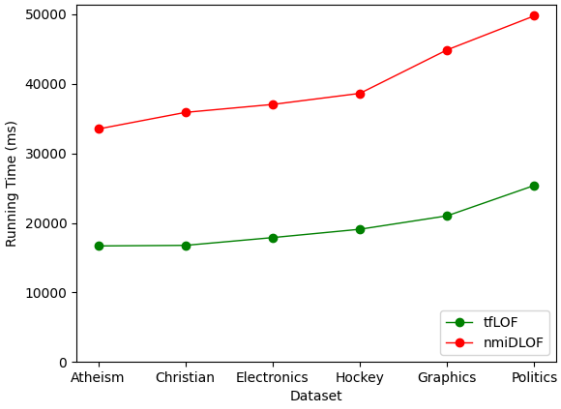

Fig.15. Running time on reuters-21578 collection

References

[1] M. Agyemang, K. Barker, and R. Alhajj, "Web outlier mining: Discovering outliers from web datasets," in Intelligent Data Analysis, 2005. doi: 10.3233/ida-2005-9505.

[2] "IBM Blog." https://www.ibm.com/blog/ (accessed Jul. 08, 2023).

[3] R. Kannan, H. Woo, C. C. Aggarwal, and H. Park, "Outlier detection for text data," *Proc West Mark Ed Assoc Conf*, pp. 489–497, 2017, doi: 10.1137/1.9781611974973.55.

[4] W. A. Mohotti and R. Nayak, "Efficient Outlier Detection in Text Corpus Using Rare Frequency and Ranking," *ACM Trans Knowl Discov Data*, vol. 14, no. 6, 2020, doi: 10.1145/3399712.

[5] M. Agyemang, K. Barker, and R. S. Alhajj, "WCOND-mine: Algorithm for detecting web content outliers from web documents," in *Proceedings - IEEE Symposium on Computers and Communications*, 2005. doi: 10.1109/ISCC.2005.155.

[6] C. C. Aggarwal and C. X. Zhai, *Mining text data*, vol. 9781461432234. 2013. doi: 10.1007/978-1-4614-3223-4.

[7] J. Zhang, X. Long, and T. Suel, "Performance of compressed inverted list caching in search engines," in *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, 2008. doi: 10.1145/1367497.1367550.

[8] A. Hotho, A. Nürnberger, and G. Paaß, "A Brief Survey of Text Mining," *Journal for Language Technology and Computational Linguistics*, vol. 20, no. 1, 2005, doi: 10.21248/jlcl.20.2005.68.

[9] D. Chakraborty, V. Narayanan, and A. Ghosh, "Integration of deep feature extraction and ensemble learning for outlier detection," *Pattern Recognit*, vol. 89, 2019, doi: 10.1016/j.patcog.2019.01.002.

[10] Y. Liu *et al.*, "Generative Adversarial Active Learning for Unsupervised Outlier Detection," *IEEE Trans Knowl Data Eng*, vol. 32, no. 8, 2020, doi: 10.1109/TKDE.2019.2905606.

[11] K. E. and N. R., "Algorithms for Mining Distance-Based Outliers in Large Datasets," in *Proceedings of the VLDB Conference, New York, USA*, 1998.

[12] S. D. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003. doi: 10.1145/956750.956758.

[13] M. I. Petrovskiy, "Outlier detection algorithms in data mining systems," *Programming and Computer Software*, vol. 29, no. 4, pp. 228–237, 2003, doi: 10.1023/A:1024974810270/METRICS.

[14] E. Fouche, Y. Meng, F. Guo, H. Zhuang, K. Bohm, and J. Han, "Mining text outliers in document directories," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2020. doi: 10.1109/ICDM50108.2020.00024.

[15] F. Cao, X. Wu, L. Yu, and J. Liang, "An outlier detection algorithm for categorical matrix-object data," *Appl Soft Comput*, vol. 104, 2021, doi: 10.1016/j.asoc.2021.107182.

[16] M. Koppel and S. Seidman, "Detecting pseudepigraphic texts using novel similarity measures," *Digital Scholarship in the*

Humanities, vol. 33, no. 1, 2018, doi: 10.1093/llc/fqx011.
[17] M. M. Breuniq, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *SIGMOD Record (ACM Special Interest Group on Management of Data)*, vol. 29, no. 2, 2000, doi: 10.1145/335191.335388.
[18] A. H. Tanira, A. A. Rafea, and H. A. Hassan, "A density-based approach wiith vsm and weighted n-grams for detecting web document outliers," in *Proceedings of the ISCA 24th International Conference on Computer Applications in Industry and Engineering, CAINE 2011*, 2011.
[19] B. Tang and H. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, 2017, doi: 10.1016/j.neucom.2017.02.039.
[20] J. Ning, L. Chen, and J. Chen, "Relative density-based outlier detection algorithm," in *ACM International Conference Proceeding Series*, 2018. doi: 10.1145/3297156.3297236.
[21] L. Duan, L. Xu, Y. Liu, and J. Lee, "Cluster-based outlier detection," *Ann Oper Res*, vol. 168, no. 1, 2009, doi: 10.1007/s10479-008-0371-9.
[22] W. Baoyi, L. Xiangyu, and Z. Shaomin, "An improved outlier detection algorithm K-LOF based on density," *Computing, Performance and Communication Systems*, vol. 2, no. 1, pp. 1–7, Dec. 2017, doi: 10.23977/CPCS.2017.21001.
[23] F. Lazhar, "Fuzzy clustering-based semi-supervised approach for outlier detection in big text data," *Progress in Artificial Intelligence*, vol. 8, no. 1, 2019, doi: 10.1007/s13748-018-0165-5.
[24] M. Mohaghegh and A. Abdurakhmanov, "Anomaly Detection in Text Data Sets using Character-Level Representation," in *Journal of Physics: Conference Series*, 2021. doi: 10.1088/1742-6596/1880/1/012028.
[25] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans Neural Netw*, vol. 5, no. 4, 1994, doi: 10.1109/72.298224.
[26] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, 2005, doi: 10.1109/TPAMI.2005.159.
[27] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans Neural Netw*, vol. 20, no. 2, 2009, doi: 10.1109/TNN.2008.2005601.
[28] K. Zhou, W. Wang, T. Hu, and K. Deng, "Application of improved asynchronous advantage actor critic reinforcement learning model on anomaly detection," *Entropy*, vol. 23, no. 3, 2021, doi: 10.3390/e23030274.
[29] D. Zha, K. H. Lai, M. Wan, and X. Hu, "Meta-AAD: Active anomaly detection with deep reinforcement learning," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2020. doi: 10.1109/ICDM50108.2020.00086.
[30] G. Pang, A. Van Den Hengel, C. Shen, and L. Cao, "Toward Deep Supervised Anomaly Detection: Reinforcement Learning from Partially Labeled Anomaly Data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2021. doi: 10.1145/3447548.3467417.
[31] M. Yu and S. Sun, "Policy-based reinforcement learning for time series anomaly detection," *Eng Appl Artif Intell*, vol. 95, 2020, doi: 10.1016/j.engappai.2020.103919.
[32] S. M. Weiss, N. Indurkhya, T. Zhang, and F. J. Damerau, *Text mining: Predictive methods for analyzing unstructured information*. 2005. doi: 10.1007/978-0-387-34555-0.
[33] "(10)Dataset Text Document Classification | Kaggle." https://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification (accessed Jul. 08, 2023).
[34] "Reuters-21578 Text Categorization Collection - UCI Machine Learning Repository." https://archive.ics.uci.edu/dataset/137/reuters+21578+text+categorization+collection (accessed Jul. 08, 2023).
[35] C. Vidyadhari, N. Sandhya, and P. Premchand, "Automatic Incremental Clustering Using Bat-Grey Wolf Optimizer-Based MapReduce Framework for Effective Management of High-Dimensional Data," *International Journal of Ambient Computing and Intelligence*, vol. 11, no. 4, 2020, doi: 10.4018/IJACI.2020100105.

Authors' Profiles

Ayman H. Tanira is a lecturer of computer science at Palestine Technical College-Deir El-Ballah (PTCDB), Palestine. He was granted his bachelor's degree in computer science from Mu'tah University, Jordan and he was the top student. Mr. Tanira obtained his master's degree from Cairo University, Egypt, and also, he was the top student among his colleagues. Mr. Tanira is currently a Ph.D. student of computer engineering at the Islamic University of Gaza (IUGaza). His research focuses on text mining, deep learning, information security, and Blockchain. Mr. Tanira published a set of papers in international conferences and journals.

Wesam S. Ashour is a professor of computer engineering at the Islamic University of Gaza (IUG), Palestine. He has graduated in 2000 with B.Sc. in Electrical and Computer Engineering from Islamic University of Gaza then he got a studentship and traveling to UK for M.Sc. Prof. Ashour has finished his M.Sc. in Multimedia with Distinction in 2004 from the University of Birmingham, UK. In 2005, he has got a scholarship from the University of the West of Scotland (UWS), UK, for his PhD. During his PhD study, he has worked in UWS as a teaching assistant and lab demonstrator for some modules. Prof. Ashour has been the head of the Computer Engineering Department at IUG for periods 2009-2010 and 2013-2015. Also, Prof. Ashour has occupied the position of Assistant Vice President for Research and Graduate Affairs at IUG for the period August 2015 – August 2017, the position of Vice Dean of External Relations Affairs for the period August 2017 – August 2018, and the position Deputy Dean of the Faculty of Engineering 2018-2019, 2022-now.

How to cite this paper: Ayman H. Tanira, Wesam M. Ashour, "A Hybrid Unsupervised Density-based Approach with Mutual Information for Text Outlier Detection", International Journal of Intelligent Systems and Applications(IJISA), Vol.15, No.5, pp.41-56, 2023. DOI:10.5815/ijisa.2023.05.04