DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

Hybridized Dimensionality Reduction Method for Machine Learning based Web Pages Classification

Thabit Sulaiman Sabbah
Faculty of Technology and Applied Sciences, Al-Quds Open University, Rammallah, Palestine
tazazmeh@qou.edu

Abstract— *Feature space high dimensionality is a well-known problem in text classification and web mining domains, it is caused mainly by the large number of vocabularies contained within web documents. Several methods were applied to select the most useful and important features over the years; however, the performance of such methods is still improvable from different aspects such as the computational cost and accuracy. This research presents an enhanced cosine similarity-based hybridization of two efficient feature selection methods for higher classification performance. The reduced feature sets are generated using the Random Projection (RP) and the Principal Component Analysis (PCA) methods, individually, then hybridized based on the cosine similarity values between features' vectors. The performance of the proposed method in terms of accuracy and F-measure was tested on a dataset of web pages based on several term weighting schemes. As compared to relevant methods, results of the proposed method show significantly higher accuracy and f-measure performance based on less feature set size.*

Index Terms— *Cosine similarity, Dimensionality Reduction, Feature selection, PCA, Random Projection.*

I. INTRODUCTION

The rapid increase in the volume of text data on the Internet is one of the major reasons that lead to the high dimensionality in text mining and web classification. Hence, the requirement of effective Dimensionality Reduction (DR) techniques increases. High dimensionality is a known provocation in various domains including Text Classification/Categorization (TC) and web mining [1]. However, the huge number of features increases computational complexity, reduces clustering methods performance, and shows the necessity of efficient DR techniques.

The primary aim of DR techniques is to select or maintain the informative and discriminative sub-feature space [1], without a significant drop effect in the method's performance. Generally, DR methods are of three main categories; Feature Extraction (FE), Feature Selection (FS), and Hybrid methods [1, 2] where each category works toward the primary aim but in different ways. The third category in which different methods from FS and FE categories are hybridized in various approaches is gaining increasing interest [1].

FS and FE methods are statistical-based methods that work on top of the statistical representation of the data. However, in text-classification for terrorism activities detection approaches, various types of features have been frequently employed such as syntactic, lexical, stylistic, n-grams, Bag of Words (BoW), and Parts of Speech (POS), or a combination of these features [3]. However, the significant features of the text are usually determined using the Term frequency-inverse document frequency (TFIDF) weighting method [1]. This work considers distinctive words contained in web pages as the features of

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*

DOI: https://doi.org/10.33103/uot.ijcccce.22.3.9

feature space, where these features are weighted using different weighting schemes as described in section III.

II. RELATED WORKS

This section summarizes the works related to dimensionality reduction methods.

A. Feature Selection (FS) and Extraction (FE) Methods

FS (or variable selection) methods are the methods that aim to choose a subset of features among the original feature space in which the selected features are best enough to obtain acceptable performance. These techniques are based on the assumption that many of the features (terms) included in texts are uninformative and less important, and their removal will not affect the quality of the classification but will increase the accuracy as well as decrease the complexity and processing time of classification [4]. However, on the other hand, their computational complexity is higher, and the selected features are biased toward the employed learning method. FS methods depend on the term weighting schemes [5], in which the terms (features) are weighted. Various weighting schemes were proposed in literature including Term Frequency (TF), Term Frequency – Inverse Document Frequency (TFIDF), in addition to many other modified weighting methods as presented in Sabbah, et al. [6]. The number of occurrences of a term is a major factor that most weighting schemes depend on in calculating the term's weight.

Similarly, the aim of FE methods is to select the most informative features, nevertheless by eliminating as few informative and redundant features as possible [7]. In the TC domain, the huge number of distinctive (unique) words (terms) contained in the corpus is the major reason for the high dimensionality problem, however, much of these features are redundant. In literature, many methods are proposed to reduce the complexity of data in the TC field; for example, Principal Component Analysis (PCA) [8] and Random Projections (RP) [9].

B. Hybridized Reduction Methods

Table summarizes many of the hybridized DR methods related to TC domain based on their integration method and testing domain. However, In general, each of the DR techniques mentioned above considers only one aspect of the features while working toward feature space reduction [1]. Therefore, much attention is given recently to hybrid feature reduction techniques. Hybrid methods work toward considering different aspects of features by integrating different reduction methods. However, on the other hand, in some works, the optimization methods such as Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) are applied for further tuning of the feature sub-lists.

As seen in Table , many of the hybrid approaches presented in the literature are in the consecutive form, in which different methods of DR are performed sequentially. Thus, the performance of such methods is generally limited or driven by the method which is applied first. On the other hand, some other hybridized DR methods presented above are based on set operations such as union, modified union, intersection, multi-intersection. This paper presents a hybridized DR method that utilized a controlled union operation through the cosine similarity between the reduced feature spaces produced by different FE methods.

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

TABLE I. HYBRIDIZED DIMENSIONALITY REDUCTION METHODS SUMMARY

Method of integration		Testing Domain	Reference	Integrated Methods*
Consecutive		TC	Sabbah, et al. [10], Uysal [11] Lin, et al. [12] Bharti and Singh [13] Uğuz [14] Sam, et al. [15]	TFIDF-SVM, IG-IGFSS, DFS-IGFSS TFIDF-ICSO MM-AC-PCA, MAD-AC-PCA IG-GA, IG-PCA PCA-ICA
Set operations	Union	TC	Sabbah, et al. [10] Sabbah, et al. [16] Sabbah, et al. [17] Selamat and Omatu [18]	(IG-Chi^2)-SVM TFIDF-SVM TF-DF-TFIDF-Glasgow-Entropy Entropy-PCA
		Dark Web Classification	Sabbah and Selamat [19]	TFIDF-PCA, TFIDF-RP
	Union,intersection, multi_Int*	TC	Bharti and Singh [1]	TV-DF, TV-DF-PCA
Optimization		TC, computer vision, data mining	Boutemedjet, et al. [20]	GD-EM

AC: Absolute Cosine
Cor: Correlation
DF: Document Frequency
DFS: Distinguishing Feature Selector
EM: Expectation-Maximization
GA: Genetic Algorithm
GD: Generalized Dirichlet
ICA: Independent Component Analysis
ICSO: Improved Cat Swarm Optimization
IG: Information Gain
MAD: Mean Absolute Difference
MM: Mean-median
PCA: Principal Component Analysis
SVM: Support Vector Machines
TF: Term Frequency
TFIDF: Term Frequency-Inverse Document Frequency
TV: Term Variance
multi_Int*: Multi Intersection

III. PROPOSED HYBRIDIZED FEATURE SELECTION METHOD

The review of the hybridized DR methods shown in section II, reveals that many hybrid methods have been applied in the TC domain. However, a few methods have been utilized for Dark Web (DW) pages classification. Moreover, the review shows that the PCA feature extraction method has been applied as a part of the hybridization; nevertheless, a few of the hybridized methods have applied the random projection (RP) method as well as PCA and RP together. One method that utilized PCA and RP in the domain of DW detection is Sabbah and Selamat [19], in which the ordinary union integration of PCA and RP based on the TFIDF feature selection method was applied. The results of that work were bounded and findings could not be generalized. Thus, this research presents an enhanced DR hybridization method based on PCA and RP as feature selection methods to achieve higher DW classification performance.

This work proposes an enhanced hybridization technique of PCA and RP feature extraction methods, in which the feature space is firstly reduced to equal-size lower dimensional feature spaces using PCA and RP individually, and then the reduced feature spaces are merged, while the similar features are eliminated, where the cosine similarity is utilized in finding the similar feature.

Fig. 1 shows the processed flow of the proposed method, where the processes in the shaded region are the main contribution of this work.

In general, the main steps of the proposed method can be summarized as follows:

i. Apply required pre-processing algorithms on the corpus.
ii. Represent the corpus into Vector Space Model (VSM).
iii. Apply the Principal Component Analysis (PCA) feature extraction method, to produce the reduced feature space f_{PCA}.
iv. Apply the Random Projection (RP) feature extraction method, to produce the reduced feature space f_{RP}, which is of the same dimension as f_{PCA} generated in step iii.
v. Calculate the cosine similarity matrix between f_{PCA} and f_{RP}.
vi. Determine the threshold value.

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

vii. Remove the features where the similarity is greater than the threshold value t, while merging the feature spaces f_{PCA} and f_{RP} into one feature set (f_{final}).
viii. Perform Machine learning based Classification.
ix. Evaluate performance.

In details, the proposed method contains the following processes:

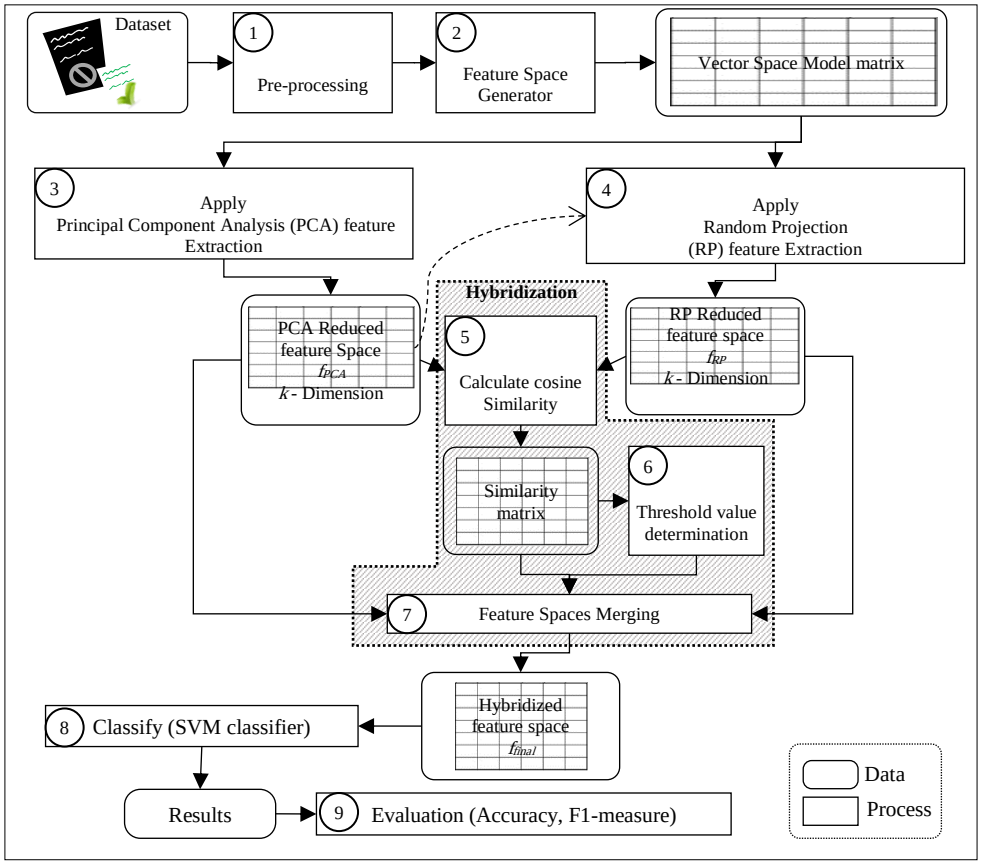


FIG. 1. PROPOSED HYBRIDIZED DIMENSIONALITY REDUCTION METHOD PROCESSES FLOW.

A. Pre-processing

This process consists of various activities such as filtering, tokenizing, and stemming. Filtering involves stop-words removal based on the common Arabic stop word list[1], in addition to removing numbers, symbols, special characters such as punctuations, and non-Arabic characters and diacritics [21]. To achieve the stimming process, this research considered Larkey's Light Stemmer algorithm [22].

B. Feature Space Generator

In this research, the Rapidminer software[2] was used to handle the VSM representation of the corpora, while the schemes shown in Table were used as separately as term weighting schemes.

[1] http://arabicstopwords.sourceforge.net/
[2] http://rapidminer.com/

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

C. Principal Component Analysis (PCA) application

As a linear technique of dimensionality reduction, The PCA is extensively used in many fields such as image processing and text classification [23] and [24], respectively. Mathematically, in PCA the data presented in a feature space of certain dimensionality is transformed into a new coordinate system in which the features are orthogonal, the mathematical details of PCA are presented in [25]. In this research, a Matlab implementation of PCA was used for feature space reduction.

TABLE II. HYBRIDIZED DIMENSIONALITY REDUCTION METHODS SUMMARY

Scheme	Formula	Notation description
TF	$TF_{t,d} = \frac{fr_{t,d}}{\sqrt{\sum_{t=1}^{n} fr_{t,d}{}^2}}$	TF: Term Frequency DF: Document Frequency TFIDF: Term Frequency - Inverse Document Frequency fr_{td} is the raw frequency of term *t* in document *d*, *n* is the number of distinctive terms in document *d*, *N* is the number of documents in the collection, $length_d$ the length of the vector that represents the distinctive terms of document *d*, F_t is the frequency of term *t* on the collection level
DF	$DF_t = \sum_{d=1}^{N} \begin{cases} 1 & t \in d \\ 0 & t \notin d \end{cases}$	
TFIDF	$TF - IDF_{t,d} = TF_{t,d} \cdot IDF_t$ where $IDF_t = \log\ (N/DF_t) + 1$	
Entropy	$w_{td} = L_{td} \times G_t$ Where $G_t = \frac{1+\sum_{j=1}^{N} \frac{fr_{td}}{F_t} \log\left(\frac{fr_{td}}{F_t}+1\right)}{\log N}$ and $L_{td} = \begin{cases} 1 + \log fr_{td}, & fr_{td} > 0 \\ 0, & fr_{td} = 0 \end{cases}$	

D. Random Projection (RP) application

RP is a common effective and modest linear DR technique [26]. The RP method utilizes the random projection matrices to generate a lower-dimensional space of the given data [27]. RP mathematical formulation and proofs can be found in [9]. However, in this study, the matrix R is constructed as a sparse random matrix, in which the values are of values: 0, -1, and +1 with specified probabilities of 2/3, 1/6, and 1/6 respectively based on Achlioptas [28] work, as in (1).

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1k} \\ r_{21} & \cdots & r_{2k} \\ \vdots & r_{ij} & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ r_{n1} & \cdots & r_{nk} \end{pmatrix}_{n \times k} , where\ r_{ij} = \sqrt{3} \begin{cases} +1\ with\ probability\ \frac{1}{6} \\ 0\ \ with\ probability\ \frac{2}{3} \\ -1\ with\ probability\ \frac{1}{6} \end{cases} \quad (1)$$

E. Hybridization

The hybridization process in the proposed method includes three steps; the calculation of similarity matrix between the reduced feature spaces f_{PCA} and f_{RP}, threshold value determination, and the removal of similar features where the cosine similarity is greater than the threshold value while merging the feature sets. This subsection presents these three steps.

Cosine Similarity Calculation

This study calculates the similarity between the feature vectors using the cosine similarity method as it is one of the widely used similarity methods in the domain of statistical text mining. Moreover, the cosine similarity method is more suitable than other similarity measures to hybridize the reduced feature sets, as the features in each of these feature sets are described as orthogonal numerical vectors. In the hybridization process, the features that the absolute value of their similarity is greater than a certain threshold value

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

are removed, as these vectors are similar to each other's, and the remaining features that are not similar are used for the classification.

Threshold Value Determination

The cosine similarity values are ranged between -1 and +1, where the minus sign indicated the opposite direction of the feature vector. The higher absolute value of the similarity indicates the more similar features in the same or the opposite direction. The proposed method is based on a unique idea, and no such existing studies are using the same method of hybridization, therefore, the threshold value could not be determined based on literature or from previous studies. Thus, the threshold value in this research is determined based on the experimental results. *Fig. 2* shows the average performance of all weighting schemes for different threshold values.

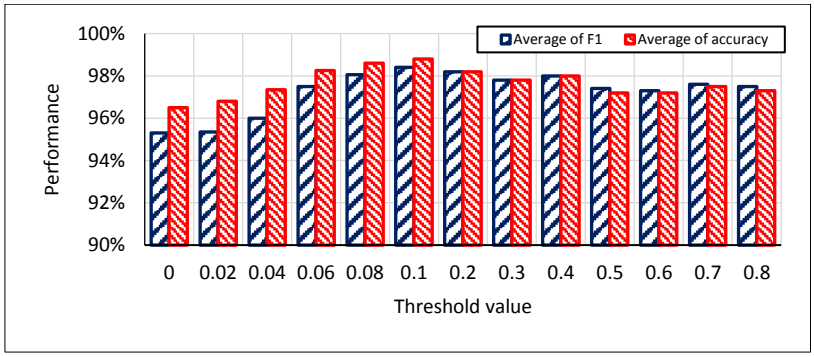


FIG. 2. AVERAGE PERFORMANCE OF TERM WEIGHTING SCHEMES ON DIFFERENT THRESHOLD VALUES.

The results in *Fig. 2*, was generated by conducting the classification using the proposed method using different weighting schemes, based on the shown threshold values on x-axis, then the F1 and accuracy performances were averaged. It is noticed from *Fig. 2* that the average F1 and accuracy performances of all schemes are the highest based on the hybridization at the threshold value t = 0.1, therefore, this study considered the value 0.1 as the threshold value for the hybridization.

Feature Spaces Merging

As seen *Fig. 1*, in the similarity matrix is used as an input to the "hybridize" process in addition to the feature sets f_{PCA} and f_{RP}, and the determined threshold value. In the hybridization process, the feature sets f_{PCA} and f_{RP} are merged together into one feature space named f_{final}, however, the features of a cosine similarity greater than the threshold value (t) are removed from the feature sets while merging.

F. Classification

The hybridization method proposed in this study is able to generate a reduced feature set in order to effectively increase the performance of DW classification based on the hybridization of RP and PCA methods. This research considered the Support Vector Machine (SVM) classifier to carry out the classification processes. SVM classifier is found to be superior in comparison to the other classifiers in the domain of DW analysis [29]. Moreover, the experiments conducted in this research were based on the K-fold cross-validation where K=10. The following consecutive sections discuss the results of these experiments.

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

IV. EXPERIMENTAL SETUP

This section describes the dataset considered in this study, followed by an explanation of the evaluation metrics.

A. Dataset

This study, similar to many of existing works in DW classification and detection, depends on the Dark Web Portal Forum (DWPF) as the source of experimental datasets [30-37]. For the requirements of this research, thousands of documents were crawled from DWFP, among them 1000 documents were labeled as dark by a native Arabic expert. The Open Source Arabic Corpora (OSAC) [38] were utilized to obtain the non-dark documents. The dataset considered in this research is an unbalanced dataset that consists of 1,000 documents, where 25% of the documents are of the dark label and the remaining 75% of documents are of the non-dark label. This dataset was considered to examine the proposed methods intensively under the condition of reality simulation, which is described by Correa and Sureka [39] where the amount of dark content is much smaller than the non-dark content. Table shows the statistical information of the dataset.

TABLE III. DATASET DESCRIPTION

Class Label	Property	Unbalanced Dataset
Dark	Number of documents	250
	Percentage	25%
	maximum document length (in words)	4279
	minimum document length (in words)	469
	average document length (in words)	1628.9
Non-Dark	Number of documents	750
	Percentage	75%
	maximum document length (in words)	5654
	minimum document length (in words)	447
	average document length (in words)	1421.7
Total Number of Documents		**1000**
Total Number of features		**109263**

B. Evaluation

To evaluate the work presented in this work, F1 and Accuracy measures were employed. These evaluation measures in addition to Precision and Recall are the common evaluation measures in the domain of TC and dark web detection [40].

The performance measurements Precision, Recall, F1, and Accuracy are calculated as in (3)-(6):

$$Precision = |TP| / (|TP|+|FP|) \quad (3)$$

$$Recall = |TP| / (|TP| + |FN|) \quad (4)$$

$$F1 = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (5)$$

$$Accuracy = (|TP|+|TN|) / (|TP|+|FP|+|FN|+|TN|) \quad (6)$$

The results in this study are presented only using F1 and accuracy to shorten and avoid unnecessary redundancy of data.

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

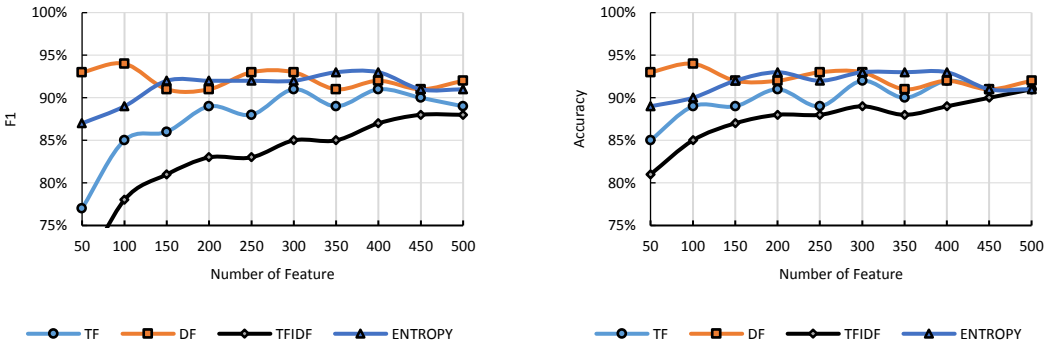
DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

V. EXPERIMENTS AND RESULTS DISCUSSION

A series of experiments were conducted on the considered dataset. the data is represented in VSM based on the different term weighting schemes, and then the Principal Component Analysis (PCA) and the Random Projection (RP) methods were applied on the VSM matrix, individually. The classification performance based on the feature sets of the first 50 to 500 features (in intervals of 50 features) is evaluated. These feature sets were hybridized using the proposed method, and finally, the hybridized feature sets' performance was evaluated and compared to the individual PCA and RP methods' performance. Section VI discusses these results and presents the significant test results.

A. Individual Methods Performance

This section presents the performance results using F1 and accuracy measures of the methods Random Projection (RP) and Principal Component Analysis (PCA) on the considered dataset individually.

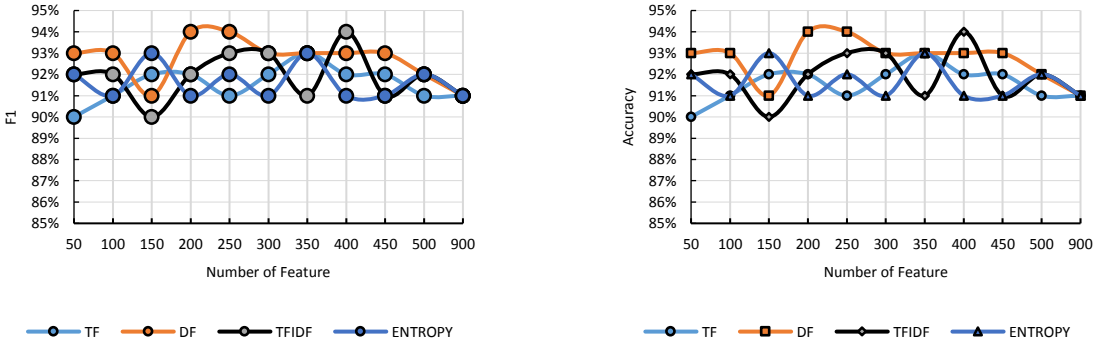


(a) F1 performance (b) Accuracy performance

FIG. 3. RANDOM PROJECTION PERFORMANCE USING DIFFERENT TERM WEIGHTING SCHEMES.

As seen in (*Fig. 3* a), the F1 performance is upper bounded by 94.0%. Although, the F1 performance is less than 85.0% for some schemes such as TFIDF based on small feature sets, however, the performance increases with the increase in the number of features.

The high F1 performance of the RP method indicates the ability of this method to detect the true positive samples (i.e., documents that contain dark contents) in an imbalanced dataset which simulates the real environment on the Internet where the dark content is much less than the non-dark content.



(a) F1 performance (b) Accuracy performance

FIG. 4. PRINCIPAL COMPONENT ANALYSIS PERFORMANCE USING DIFFERENT TERM WEIGHTING SCHEMES.

Received 6/October/2021; Accepted 19/November/2021

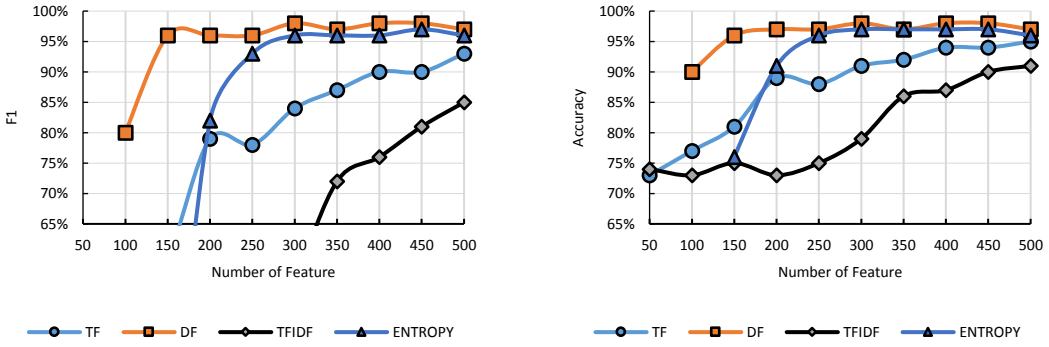
ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

Fig. 4 shows the F1 and accuracy performance measurements of PCA feature extraction method, where it is seen that the F1 and accuracy performance of different term weighting schemes are closer to each other than the corresponding results of RP method.

B. Hybridized Feature Selection Method Performance

Fig. 5 shows the performance measures of the proposed Hybridized feature selection method using different term weighting schemes.



(a) F1 performance (b) Accuracy performance

FIG. 5. HYBRIDIZED FEATURE SELECTION METHOD PERFORMANCE.

It is seen from (*Fig. 5* a) that the DF weighting scheme has achieved the highest F1 performance 98.0% based on the hybridization of small feature sets (300 features). Similarly, the accuracy performance depicted in (*Fig. 5* b) shows that the DF scheme has achieved the highest accuracy performance based on the same hybridized feature set.

Conversely, Table shows the numbers of features in the hybridized sets, where the numbers in bold represent the sizes of feature sets where the corresponding scheme has achieved its highest F1 performance.

Table IV shows that hybridization of RP and PCA based on different weighting schemes accordingly generates different number of features in the hybridized feature sets. For example, the hybridization of a f_{PCA} and f_{RP} of sizes 200 features per each, generates different sizes of hybridized feature sets based on the weighting scheme such as 54 features based on TF, 142 features based on DF, and so on. The high F1 and accuracy performance while a small number of features in the hybridized feature set of some schemes such as DF indicate that the proposed method has the ability to generate reduced feature sets that are capable to achieve higher detection performance under the condition of a smaller number of positive samples.

TABLE IV. NUMBER OF FEATURES IN HYBRIDIZED FEATURE SETS BASED ON DIFFERENT WEIGHTING SCHEMES

Number of feature per f_{PCA} and f_{RP}	Number of features in the final hybridized feature set (f_{final}) of different weighting schemes			
	TF	DF	TFIDF	ENTROPY
50	2	0	2	0
100	10	8	10	0
150	20	60	18	2
200	54	142	32	30
250	100	242	52	102
300	172	**342**	88	196
350	260	442	132	296
400	350	542	182	396
450	444	642	248	**494**

Received 6/October/2021; Accepted 19/November/2021

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

500	**544**	742	**326**	594
Average	196	351	109	264

VI. RESULTS DISCUSSION

The results of the proposed hybridized feature selection (HFS) method are discussed in this subsection by comparing the results of the relevant feature extraction methods PCA and RP. Additionally, the significance tests are used to testify the significance of the obtained results of the HFS method, compared to the results obtained by the relevant feature extraction methods.

A. Discussion of Hybridized Feature Selection Method Results

This subsection discusses the results of HFS method and presents the significance of these results.

Comparison of Hybridized Feature Selection (HFS) Method Results

The F1 and accuracy performance of different weighting schemes based on the proposed HFS method using feature sets of sizes from 50 to 500 in intervals of 50 features have been averaged in order to simplify the comparison between the F1 and accuracy performance. Table V shows the average F1 and accuracy performance of different weighting schemes considered by this study, where the higher performance (F1 and accuracy) values are shown in bold. However, *Fig. 6* shows the comparison of the average performance.

TABLE V. AVERAGE F1 AND ACCURACY PERFORMANCE AND AVERAGE NUMBER OF FEATURES OF DIFFERENT SCHEMES

Weighting scheme	F1 (%)	Accuracy (%)	Number of features
TF	70.5	87.4	196
DF	**95.1**	**96.4**	351
TFIDF	46.1	80.3	109
ENTROPY	84.8	93.4	264

From Table V, it is seen that DF scheme have achieved higher average performance for both measures F1 and accuracy.

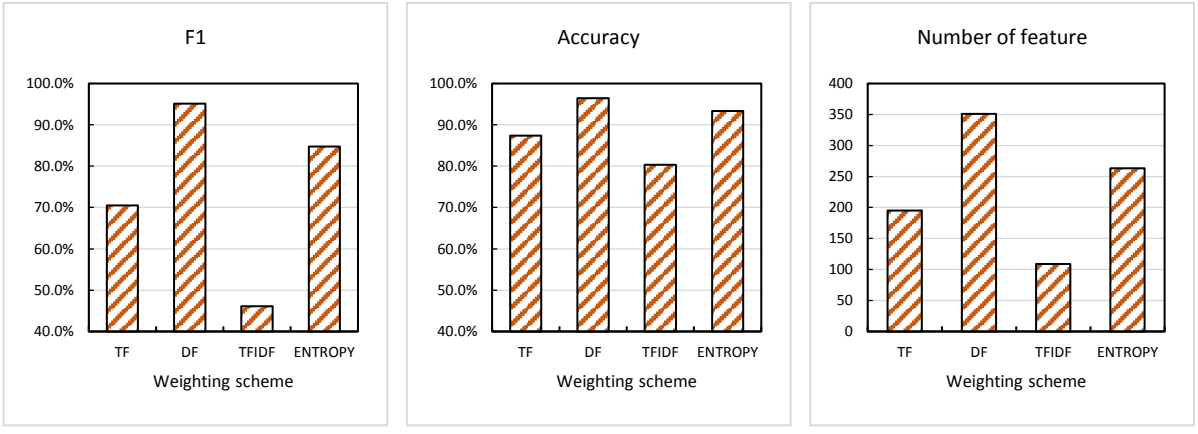


FIG. 6. COMPARISON OF AVERAGED PERFORMANCE OF DIFFERENT WEIGHTING SCHEMES USING THE ENHANCED HYBRIDIZED FEATURE SELECTION METHOD.

The results in Table and the comparison shown in *Fig. 6* show that the HFS method has not only a higher ability to detect the positive samples (i.e. the DW content) but is also able to lower the error rate of both types, in the case of less number of DW documents.

Received 6/October/2021; Accepted 19/November/2021

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijcccce.22.3.9

Moreover, it is noticeable that the hybridized feature selection method has achieved higher performance based on a relatively small feature set size. However, these results show that the HFS has a good performance in the environment that simulates the real web where the amount of Dark Web content is the minority.

Significance Tests

Fig. 7 and ***Error! Reference source not found.*** show the p-values of the F1 and accuracy paired sample T-Test for all pairs of weighting schemes. However, the null hypothesizes to be tested for each pair of term weighting schemes S_a and S_b is as follows:

H_f: the difference between the mean F1 performance of the HFS method based on S_a and S_b is statistically not significant.
H_a: the difference between the mean accuracy performance of the HFS method based on S_a and S_b is statistically not significant.

In *Fig. 7* the top row and the leftmost column are the different weighting schemes considered in this study. However, the value in the cell where a column and a row intersect represents the p-value of the paired sample T-Test of the pair formed of the schemes in the top row and the leftmost column. For example, the *p*-value of the paired sample T-Test of the pair of schemes TF-DF is 0.003.

	F1				Accuracy			
	TF	**DF**	**TFIDF**	**ENTROPY**	**TF**	**DF**	**TFIDF**	**ENTROPY**
TF		0.003*	0.002*	0.709		0.001*	0.002*	0.076**
DF			0.001*	0.214			0.000*	0.150
TFIDF				0.015**				0.003*
ENTROPY								

*Significant at α = 0.01 **significant at α = 0.05

FIG. 7. SUMMARY OF RESULTS FROM PAIRED SAMPLE T-TEST ON HFS METHOD F1 AND ACCURACY PERFORMANCE BASED ON DIFFERENT SCHEMES.

From *Fig. 7* the differences between HFS performances based on different weighting schemes are seen to be significant at 0.01 and 0.05 levels for most pairs. For each pair where the P-value is less than alpha, the null hypothesis is rejected and the alternative hypothesis is to be accepted. For example, the p-value of the paired sample T-Test of the pair TFIDF-DF is 0.001. However, this p-value is less than α = 0.01, which means that the difference in HFS F1 performance based on the scheme TFIDF on one hand and the scheme DF, on the other hand, is not due to the chance, and the scheme with higher F1 performance is preferable.

Comparison of Hybridized Feature Selection (HFS) Method Results with Results of PCA and RP Feature Extraction Methods

In this comparison, the average F1 and accuracy performance, as well as the number of features, of the HFS is compared to the corresponding average performance of the methods PCA and RP that presented in the Section V. This comparison is conducted to c the differences and the improvement of performance measurements achieved by applying the HFS method using some weighting schemes.

Fig. 8 shows the Average F1 and Accuracy comparison between HFS and the PCA and RP feature extraction methods using different weighting schemes.

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

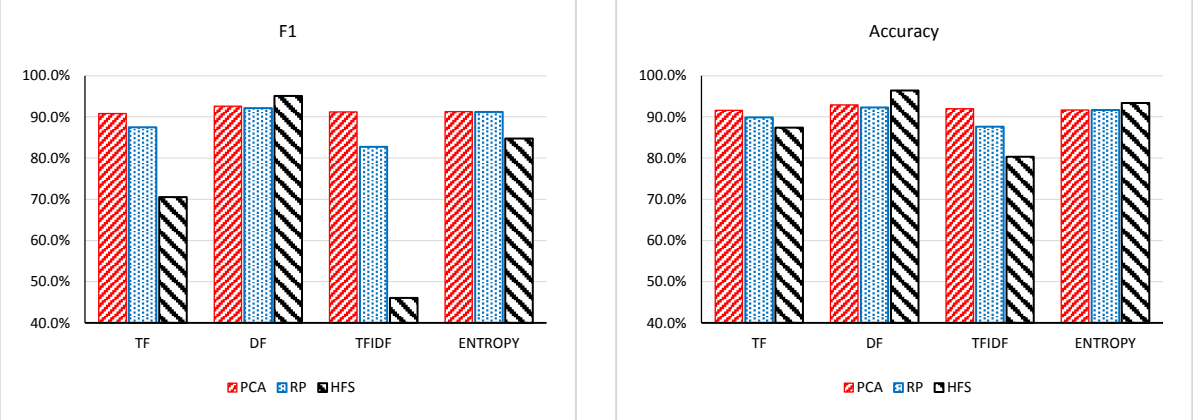


FIG. 8. COMPARISON OF AVERAGE F1 AND ACCURACY PERFORMANCE BETWEEN HFS AND RELEVANT METHODS.

The comparison in *Fig. 8* shows that the HFS method has achieved higher average F1 and Accuracy performance based on the DF. Similarly, the PCA method has achieved higher F1 and Accuracy performance based on the schemes TF, TFIDF, and Entropy. The highest average F1 performance has been achieved by the methods HFS based on the scheme DF with the value of 95.1%.

VII. CONCLUSIONS AND FUTURE DIRECTIONS

This research presented the proposed hybridized feature selection method, in which the hybridization is applied on PCA and RP feature reduction methods based on cosine similarity. The proposed method was tested and benchmarked in the domain of Dark Web (DW) pages classification using an imbalanced dataset and based on several term weighting schemes. The results of the proposed HFS method were showed, discussed, and compared to the results of the relevant methods. Based on many term-weighting schemes, in terms of the accuracy and F1 measures, the proposed HFS method outperformed individual feature reduction methods significantly, whereas the generated feature space of HFS was of lower dimensionality. In the future, the proposed method will be tested using public datasets in text classification and other domains where high dimensionality problem is the main concern.

REFERENCES

[1] K. K. Bharti and P. K. Singh, "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3105-3114, 2015.

[2] C. Liu *et al.*, "A new validity index of feature subset for evaluating the dimensionality reduction algorithms," *Knowledge-Based Systems*, vol. 121, pp. 83-98, 2017/04/01/ 2017.

[3] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing- style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378-393, 2006.

[4] W. Wibowo and H. E. Williams, "Simple and accurate feature selection for hierarchical categorisation," in *Proceedings of the 2002 ACM symposium on Document engineering*, McLean, Virginia, USA, pp. 111-118, 585079: ACM, 2002.

[5] Y. Liu, H. T. Loh, and A. X. Sun, "Imbalanced text classification: A term weighting approach," (in English), *Expert Systems with Applications*, vol. 36, no. 1, pp. 690-701, Jan 2009.

[6] T. Sabbah *et al.*, "Modified frequency-based term weighting schemes for text classification," *Applied Soft Computing*, vol. 58, pp. 193-206, 2017.

[7] C. O. S. Sorzano, J. Vargas, and A. Pascual-Montano, "A survey of dimensionality reduction techniques," *arXiv preprint arXiv:1403.2877*, 2014.

[8] W. Buntine, S. Perttu, and V. Tuulos, "Using Discrete PCA on Web Pages," in *Workshop on Statistical Approaches to Web Mining (SAWM'04)*, Pisa, Italy, pp. 1-14, 2004.

Received 6/October/2021; Accepted 19/November/2021

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.22.3.9

[9] S. Kaski, "Dimensionality reduction by random mapping: Fast similarity computation for clustering," in *Proceedings of The 1998 IEEE International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence*, Anchorage, AK, USA, vol.1: IEEE, pp. 413-418, 1998.

[10] T. Sabbah, M. Ayyash, and M. Ashraf, "Hybrid support vector machine based feature selection method for text classification," *The International Arab Journal of Information Technology,*, vol. 15, no. 3A, p. 599-609, 2018.

[11] A. K. Uysal, "An improved global feature selection scheme for text classification," *Expert Systems with Applications*, vol. 43, pp. 82-92, 2016.

[12] K.-C. Lin, K.-Y. Zhang, Y.-H. Huang, J. C. Hung, and N. Yen, "Feature selection based on an improved cat swarm optimization algorithm for big data classification," *The Journal of Supercomputing*, journal article pp. 1-12, 2016.

[13] K. K. Bharti and P. K. Singh, "A three-stage unsupervised dimension reduction method for text clustering," *Journal of Computational Science*, vol. 5, no. 2, pp. 156-169, 3// 2014.

[14] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024-1032, 10// 2011.

[15] L. Z. Sam, M. A. Maarof, and A. Selamat, "Automated Web Pages Classification with Integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as Feature Reduction," in *Proceedings of the International Conference on Man-Machine Systems (ICoMM06)*, Langkawi, Malaysia, 2006.

[16] T. Sabbah, M. Ayyash, and M. Ashraf, "Support Vector Machine based Feature Selection Method for Text Classification," in *The International Arab Conference on Information Technology (ACIT'2017)*, Yassmine Hammamet, Tunisia, 2017: http://acit2k.org.

[17] T. Sabbah, A. Selamat, M. H. Selamat, R. Ibrahim, and H. Fujita, "Hybridized term-weighting method for Dark Web classification," *Neurocomputing*, vol. 173, Part 3, pp. 1908-1926, 2016.

[18] A. Selamat and S. Omatu, "Web page feature selection and classification using neural networks," (in English), *Information Sciences*, vol. 158, no. 1, pp. 69-88, Jan 2004.

[19] T. Sabbah and A. Selamat, "Hybridized Feature Set for Accurate Arabic Dark Web Pages Classification," in *The 14th International Conference On Intelligent Software Methodologies, Tools And Techniques (SoMeT2015)*, Naples, Italy, 2015.

[20] S. Boutemedjet, N. Bouguila, and D. Ziou, "A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1429-1443, 2009.

[21] A. F. E. Gohary, T. I. Sultan, M. A. Hana, and M. M. E. Dosoky, "A Computational Approach for Analyzing and Detecting Emotions in Arabic Text," *International Journal of Engineering Research and Applications (IJERA)*, vol. 3, no. 3, pp. 100-107, 2013.

[22] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light Stemming for Arabic Information Retrieval," in *Arabic Computational Morphology*, vol. 38, A. Soudi, A. d. Bosch, and G. Neumann, Eds. (Text, Speech and Language Technology, Netherlands: Springer, pp. 221-243, 2007.

[23] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis," in *Proceedings of the 20th international joint conference on Artifical intelligence*, Hyderabad, India, 2007, pp. 1606-1611, 1625535: Morgan Kaufmann Publishers Inc.

[24] K. Kwang In, M. O. Franz, and B. Scholkopf, "Iterative kernel principal component analysis for image modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 9, pp. 1351-1366, 2005.

[25] I. Fodor, "A Survey of Dimension Reduction Techniques, US Department of Energy, available online at: www.llnl.gov/tid/lof/documents/pdf/240921.pdf," ed, 2002.

[26] A. Kabán and R. Durrant, "Dimension-Adaptive Bounds on Compressive FLD Classification," in *Algorithmic Learning Theory*, vol. 8139, S. Jain, R. Munos, F. Stephan, and T. Zeugmann, Eds. (Lecture Notes in Computer Science, Berlin, Germany: Springer Berlin Heidelberg, 2013, pp. 294-308.

[27] T. Kohonen *et al.*, "Self organization of a massive document collection," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 574-585, 2000.

[28] D. Achlioptas, "Database-friendly random projections," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, Santa Barbara, California, USA, pp. 274-281, 375608: ACM, 2001.

[29] H. Chen, "Exploring extremism and terrorism on the web: The Dark Web project," in *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics (PAISI 2007)*, Chengdu, China, vol. 4430 LNCS, pp. 1-20: Springer Verlag, 2007.

[30] T. Sabbah and A. Selamat, "Hybridized Feature Set for Accurate Arabic Dark Web Pages Classification," Cham, 2015, pp. 175-189: Springer International Publishing.

[31] T. Anwar and M. Abulaish, "Identifying cliques in dark web forums - An agglomerative clustering approach," in *Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI 2012)*, Washington, DC, USA, pp. 171-173: IEEE, 2012.

[32] S. A. Rios and R. Munoz, "Dark web portal overlapping community detection based on topic models," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD '12)*, Beijing, China, pp. 1-7: Association for Computing Machinery, 2012.

Received 6/October/2021; Accepted 19/November/2021

 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*

DOI: https://doi.org/10.33103/uot.ijcccce.22.3.9

[33] C. C. Yang, X. Tang, and X. Gong, "Identifying dark web clusters with temporal coherence analysis," in *Proceedings of the 2011 IEEE International Conference on Intelligence and Security Informatics (ISI 2011)*, Beijing, China, pp. 167-172: IEEE, 2011.

[34] G. L'Huillier, H. Alvarez, F. Aguilera, and S. A. Rios, "Topic-based Social Network Analysis for Virtual Communities of Interests in the Dark Web," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics(ISI-KDD 2010)*, Washington, DC, USA, pp. 66-73: Association for Computing Machinery, 2010.

[35] C. C. Yang, X. Tang, and B. M. Thuraisingham, "An analysis of user influence ranking algorithms on Dark Web Forums," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD 2010)*, Washington, DC, USA, pp. 10:1-10:7: Association for Computing Machinery, 2010.

[36] S. Kramer, "Anomaly detection in extremist web forums using a dynamical systems approach," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD 2010)*, Washington, DC, USA, pp. 8:1-8:10: Association for Computing Machinery, 2010,

[37] T. Sabbah, A. Selamat, and M. H. Selamat, "Revealing Terrorism Contents form Web Page Using Frequency Weighting Techniques," in *International Conference on Artificial Life and Robotics (ICAROB)*, Japan, 2014.

[38] M. K. Saad and W. Ashour, "OSAC: Open Source Arabic Corpora," in *Proceedings of the 6th International Conference on Electrical and Computer Systems*, Lefke, Cyprus, pp. 118-123, 2010.

[39] D. Correa and A. Sureka, "Solutions to Detect and Analyze Online Radicalization : A Survey," *CoRR*, vol. abs/1301.4916, 2013.

[40] D. Choi, B. Ko, H. Kim, and P. Kim, "Text analysis for detecting terrorism-related articles on the web," *Journal of Network and Computer Applications*, vol. 38, pp. 16-21, 2014.

Received 6/October/2021; Accepted 19/November/2021
 ISSN (Print) 1811-9212 *ISSN (Online) 2617-3352*