ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/282586569

Hybridized Term-Weighting Method for Dark Web Classification

Article *in* Neurocomputing · November 2015

DOI: 10.1016/j.neucom.2015.09.063

CITATIONS
50

READS
748

5 authors, including:

Thabit Sabbah
Al-Quds Open University
34 PUBLICATIONS **353** CITATIONS
SEE PROFILE

Ali Selamat
University of Technology Malaysia
551 PUBLICATIONS **8,792** CITATIONS
SEE PROFILE

Md Hafiz Selamat
University of Technology Malaysia
32 PUBLICATIONS **315** CITATIONS
SEE PROFILE

Roliana Ibrahim
University of Technology Malaysia
173 PUBLICATIONS **2,533** CITATIONS
SEE PROFILE

All content following this page was uploaded by Thabit Sabbah on 12 February 2019.

The user has requested enhancement of the downloaded file.

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Hybridized Term-Weighting Method for Dark Web Classification

Thabit Sabbah[a], Ali Selamat[a,*], Md Hafiz Selamat[a], Roliana Ibrahim[a] and Hamido Fujita[b]

[a]*UTM-IRDA Digital Media Center of Excellence & Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*

[b]*Software and Information Science, Iwate Prefectural University, Takizawa, Japan*

Authors' copy

Abstract

The role of intelligence and security informatics based on statistical computations is becoming more significant in detecting terrorism activities proactively as the extremist groups are misusing many of the obtainable facilities on the Internet to incite violence and hatred. However, the performance of statistical methods is limited due to the inadequate accuracy produced by the inability of these methods to comprehend the texts created by humans. In this paper, we propose a hybridized feature selection method based on the basic term-weighting techniques for accurate terrorism activities detection in textual contexts. The proposed method combines the feature sets selected based on different individual feature selection methods into one feature space for effective web pages classification. Union and Symmetric Difference combination functions are proposed for dimensionality reduction of the combined feature space. The method is tested on a selected dataset from the Dark Web Forum Portal and benchmarked using various famous text classifiers. Experimental results show that the hybridized method efficiently identifies the terrorist activities content and outperforms the individual methods. Furthermore, the results revealed that the classification performance achieved by hybridizing few feature sets is relatively competitive in the number of features used for classification with higher hybridization levels. Moreover, the experiments of hybridizing functions show that the dimensionality of the feature sets is significantly reduced by applying the Symmetric Difference function for feature sets combination.

Keywords: Data mining; Term-Weighting; Feature Sets Combination; Text Classification; Terrorism Detection; Symmetric Difference; Support Vector Machines.

1 Introduction

Internet infrastructure furnishes extremist groups with a quick and easy access environment (Ki-moon, 2012). Terrorists are exploiting the facilities such as anonymity of communication, inexpensive development and maintenance, and the huge potential audience to spread their propaganda, support, instructions, and encourage

* Corresponding author.
Email : aselamat@utm.my (Ali Selamat), *Tel: +607-5531008 Fax: +607-5530160*

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

others to join extremism (Abbasi and Chen, 2007). Since September 11, 2001 (9/11), information technology researchers are studying and tracing these groups to prevent and reduce the potential terrorism in the world by analyzing different types of content of online resources related to such groups. The hidden and covert parts of the web used by terrorist or extremist groups online is referred as the Dark Web (DW) (Zhou *et al.*, 2006). Generally, DW content is categorized into indexable and multimedia types. Multimedia content consist of images, audio, and video files while the indexable category consists of the static and dynamic text-based files. Text content is the large category in web data (Choi *et al.*, 2014) and in DW collection (Fu *et al.*, 2010).

The effectiveness of terrorism detection by means of web analysis has been proven as it has been discovered that the group known as "Hamburg Cell" that was mainly responsible for the preparation of the September 11 attacks against the United States used Internet intensively (Corbin, 2003). Various statistical text classification based techniques were proposed for terrorism activities detection from web contents. The performance of text classification and DW detection methods is highly controlled by the feature selection technique rather than the classifier's kernel function (Leopold and Kindermann, 2002). Generally, feature selection methods are the methods concern with selecting the most informative features (terms) from the entire feature space. However, selecting such features requires the features to be weighted and ranked by the weighting scheme (Choi *et al.*, 2014). In the literature, many feature selection approaches based on different term weighting schemes (Liu *et al.*, 2009) were proposed. The assumption behind these techniques is that during text classification and DW detection as well, texts contain some less important and uninformative terms. Therefore, removing these terms will not affect the quality of the classification (Efron *et al.*, 2003). However, it will reduce the complexity and the required time as well as increasing the accuracy of the classification process (Wibowo and Williams, 2002).

Many of feature selection methods, are based on a single term weighting scheme (Bharti and Singh, 2015) such as Term Frequency (TF), and Term Frequency – Inverse Document Frequency (TF-IDF). Term weighting schemes are used frequently to rank the features (terms) of the text and determine the most relevant ones before classification (Choi *et al.*, 2014). However, various ideas about the significant terms of the text are exist in the literature and lead to different term weighting techniques. For example, the TF weighting scheme supports the idea that the most significant term in the document is the most frequently mentioned term (Salton and Buckley, 1988). However, Document Frequency (DF) scheme is based on the idea that the term appears in more documents has a more significant (Chen *et al.*, 2009). Moreover, other different term weighting schemes such as TF-IDF, Entropy, and Glasgow support different assumptions of the significant terms as shown in section 2.2.

However, the hybridized feature selection methods were proposed frequently in text classification domain, where various feature selection methods were combined in different hybridization forms (Bharti and Singh, 2015).

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Table 1 Summary of hybridized feature selection methods

Reference	Integrated Methods*	Method of integration	Testing Domain
Bharti and Singh (2015)	TV-DF TV-DF-PCA	Union, intersection, modified union	Text categorization
Bharti and Singh (2014)	MM-AC-PCA MAD-AC-PCA	PCA after (AC after MM) PCA after (AC after MAD)	Text categorization
Sahu and Mishra (2012)	k-means-PSO	PSO after k-means	Gene selection
Uğuz (2011)	IG-GA IG-PCA	GA after IG PCA after IG	Text categorization
Meng *et al.* (2011)	FCD-LSI	LSI after FCD	Spam detection
El Akadi *et al.* (2011)	MRMR-GA	GA after MRMR	Gene selection
Unler *et al.* (2011)	MRMR- PSO	PSO after MRMR	Different real-world datasets
Tsai and Hsiao (2010)	PCA-GA PCA-GA-DT	Union, intersection, muti-intersection	stock prediction
Boutemedjet *et al.* (2009)	GD-EM	Optimization of GD using EM	Text categorization computer vision data mining
Song and Park (2009)	LSI-GA	GA after LSI	Text categorization
Zhang *et al.* (2008)	MRMR- ReliefF	MRMR after ReliefF	Gene selection
Sam *et al.* (2006)	PCA-ICA	ICA after PCA	Text categorization
Zhang *et al.* (2005)	7 Different feature selection methods	Union	Handwritten numerals recognition
Selamat and Omatu (2004)	Entropy-PCA	Union	Text categorization

TV: Term Variance
DF: Document Frequency
MM: Mean-median
MAD : Mean Absolute Difference
AC: Absolute Cosine
PCA: Principal Component Analysis
PSO: Particle Swarm Optimization
IG: Information Gain
GA: Genetic Algorithm
GD: Generalized Dirichlet
FCD: Feature Contribution Degree
LSI: Latent Semantic Indexing
MRMR: Maximum Relevance Minimum Redundancy
DT: Decision Tree
ICA: Independent Component Analysis
EM: Expectation-Maximization

However, many remarks are taken on these methods, such as: First, most of these methods are based on the hybridization of only two or three term weighting schemes. Second: the Symmetric Difference (SD) combination function is not applied by any of these methods. Third: some of these methods are utilizing optimization techniques such as PSO, and PCA, which works on top of the original feature sets. Fourth: none of these methods were applied in DW detection domain. Therefore, in this work, it is intended to combine the features sets selected based on many various term weighting schemes in different semantics using the SD function. The combined feature set will be utilized in DW content classification domain. However, the performances of the proposed hybridized feature sets are compared to the performance of individual feature sets on one hand, while different semantics of combinations are cross compared to each other on the other hand.

The motivation behind this research came from the thought that individual term weighting techniques have been well researched especially the TF-IDF technique. However, improving the classification performance, which is one of the current issues in machine learning, requires computational algorithms that capable to achieve higher performance. The main goal of the research is to combine the feature sets generated based on different ideas (assumptions) of significant terms into one hybridized feature set, so as to take advantage of the strength of each of the methods while complimenting the weaknesses of other methods, in order to achieve higher classification performance.

In this research, it is expected that the combined feature set will increase the classification performance of the classifier as it consists of the most important features based on numerous different weighting schemes. The

Authors copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

verity of concepts in selecting the important features from the text on which different weighting schemes are based, allows the combined feature set to include different significant features, which in turn increases the separability of classifier. Although the combination of many feature sets may lead to a high dimensionality of the data, the Symmetric Difference (SD) combination function is proposed along with the Union function. Moreover, our experimental results show that the classification accuracy achieved based on combining few small feature sets is not only higher than the accuracy based on any individual techniques in many cases, but is also higher than the accuracy based on combining many large feature sets.

Although there are many existing works in the domain of DW based on statistical methods, the main contributions of this research encompass proposing a hybridized method that utilizes the use of a wider range of term weighting techniques simultaneously in order to achieve higher classification performance. Whereas, the Symmetric Difference (SD) hybridization method, which is capable to reduce significantly the dimensionality of the features in the combined feature set, compared with the UNION (UN) function is presented in this work. In addition, this work is applied in the domain of DW detection where none of such hybridized methods have been utilized.

The rest of this paper is organized as follows: Section 2 reviews the related work in DW detection, and discusses the term-weighting schemes considered in this study and briefly introduces the classifier used for benchmarking. However, Section 3 presents the proposed method while Section 4 describes the experimental setup. Section 5 presents and discusses the results, and finally, conclusions are presented in Section 6.

2 Related works

This section aims to summarize the works on DW detection techniques where the statistical text analysis is frequently utilized. In addition, this section presents discussions and explanations on the major term weighting schemes and classification techniques applied in DW detection and text classification are discussed and explained.

2.1 DW analysis and classification

Web classification techniques are at the core of DW analysis. These techniques utilize the *structural data* and the *content data* in the analysis. Structural data are the data related to the links between web data units (i.e. the web pages) in addition to the web page structural data, such as the tree-like structure that describes the Xml or Html files. Existing studies such as (Chen *et al.*, 2008; Qin *et al.*, 2011; Qin *et al.*, 2008; Zhou *et al.*, 2006) make use of the structure of the web data by employing Link-Based Bootstrapping (LBB) techniques to classify and detect dark content in different types of online data sources such as social networks, forums, and weblogs.

In classification techniques based on web *structural data*, the crawler is initiated by a list of predefined URLs, and then by using the favorite link search algorithms and backlink technique, the initial list of URLs is expanded based on the assumption that the extremist online resources are linked because of their community structure. Links amongst online resources are defined through the interaction activities (features) such as hyperlinks, friendship or other relationship, reply to e-mail or a message, comment or (Facebook like), following

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

tweets, etc. Experts filter the expanded list of URLs to reduce the collection and analyze irrelevant web pages. Finally, the extracted features are used to classify and analyze the web data.

Similarly, in *content-based* techniques, the content (especially textual content) of web pages is extracted and processed for classification or analysis. However, text content is the large category in web data and DW collection as mentioned earlier. Over the recent years, statistical text classification methods have been used intensively in detecting potential terrorist activities on the web, in which the text is represented as weighted features. The most common representation is the TF-IDF term weighting technique, which is widely used to determine the significant words (features) in the text (Choi *et al.*, 2014). However, many other statistical text representations are frequently used with text classification approaches to detect dark content such as lexical, syntactic, stylistic(Abbasi and Chen, 2005, 2008; Zheng *et al.*, 2006), domain-specific Bag of Words (BoW), Parts of Speech (PoS) (Greevy and Smeaton, 2004), and n-grams (Chen, 2008a; Choi *et al.*, 2011; Huang *et al.*, 2010; Tianjun *et al.*, 2009).

A combination of such features is used in few studies. For example (Zheng *et al.*, 2006) uses a feature set consisting of 270 features of four categories: lexical (character and word-based), structural, syntactic, and content-specific features to identify the authorship of an online message in English and Chinese language. By classifying the data using three different classification techniques namely, Decision trees, neural networks, and support vector machine. In the lexical character-based features, the total number of characters, uppercase characters, digits, spaces, and frequency of letters special characters are all considered as features. However, the total number of words, short words, average word length, average sentence length in terms of character and words, and total number of different words are considered as word-based features. In the category of structural features, the total number of lines, sentences, paragraphs, sentences per paragraph, words per paragraph and many other statistics are employed as features, while punctuations and 150 function words frequency are used as syntactic features and the frequency of eleven content-specific keywords as content specific features. Zheng *et al.* (2006) reported the accuracy of the Support Vector Machine (SVM) classifier as the highest achieved accuracy with value of 97.69% and 83.33% based on the 270 feature for the English and Chinese datasets respectively.

However, to address Arabic language specific issues such as infection, word elongation, and diacritics, an Arabic language parser is used in (Abbasi and Chen, 2005). A feature set of 418 features of the same categories as in (Zheng *et al.*, 2006) were extracted from the bilingual (English and Arabic) documents. However, additional technical structural features such as font size, embedded images and hyperlinks were also included. In their experiments, feature sets were added incrementally to two classifiers (C4.5 and SVM). Abbasi and Chen (2005) reported that SVM classifier with combination of all feature sets performs the best in terms of accuracy (97% and 94.83%) on both English and Arabic datasets, respectively.

In (Chen, 2008b), n-gram based features were used on the level of character, word, root, and collections, a set consisting of 7,556 features was extracted as indicators of documents from two Arabic forums Al Firdaws and Al Montada. A Recursive Feature Elimination (RFE) technique in conjunction with Information Gain (IG) heuristic were applied to reduce feature dimensions and identify the most appropriate and relevant features. However, the study aimed to measure the sentiment polarities expressed in selected radical international Jihadist

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Dark Web forums. Chen (2008b) reported that 22% of the features were included in the model and all sentiment classifiers demonstrated good results, with higher than 88% accuracy in sentiment polarities.

BoW and PoS in addition to bi-gram features were also used individually in (Greevy and Smeaton, 2004) to detect racism in text, where SVM was used as the classifier. Greevy and Smeaton (2004) conducted their experiments on four different datasets and reported that the highest classification accuracy was achieved based on the BoW using the polynomial function as the classifier kernel.

However, the performance of such statistical methods and traditional frequency-based term weighting schemes such as TF-IDF in the domain of terrorist activities detection is reported to be insufficient (Choi *et al.*, 2014; Greevy and Smeaton, 2004; Ran and Xianjiu, 2010). In general, the low performance of statistical methods in text classification comes from the inability of these methods to understand the semantic meanings of a text created by humans (Choi *et al.*, 2014). Therefore, some other techniques are proposed to overcome the deficiency of individual statistical methods. Techniques proposed by (Choi and Kim, 2012; Choi *et al.*, 2014; Hwang *et al.*, 2011) utilize knowledge-based tools that provide the conceptual hierarchy interconnections such as WordNet and Wikipedia to measure the semantic relations between concepts and determine the important features. However, the improvement in classification performance is not reported to be significant relative to the traditional statistical methods.

2.2 Term weighting techniques

In information retrieval, text classification, and web-classification domains, term-weighting techniques (also known as schemes of formulas) such as Term Frequency (TF), Document Frequency (DF), and Inverse Document Frequency (IDF) are widely used (Ran and Xianjiu, 2010) in addition to TF-IDF, Entropy (Selamat and Omatu, 2004), and the Glasgow techniques. However, there are many other term-weighting schemes were proposed and used in these domains (Crestani *et al.*, 1998; Yang and Pedersen, 1997) such as such as Term Variance (TV) (Luying et al., 2005), Term Strength (TS) (Yang, 1995), Chi-square (CHI) (Yanjun et al., 2008), Inverse Document Frequency (IDF) (Robertson, 2004), Information Gain (IG) (Luying et al., 2005; Quinlan, 1986), Odds Ratio (OR) (Mengle and Goharian, 2009), Gini Index (GI) (Shang et al., 2007), Improved Gini Index (GINI) (Mengle and Goharian, 2009), Mutual Information (MI) (Peng et al., 2005), and Balanced Term Weighting Scheme (BTWS) (Jung et al., 2001).

The Vector Space Model (VSM) which proposed by (Salton *et al.*, 1975) is still a common and effective way for statistical representation of a corpus (collection) of documents. In VSM model, each document is considered as a vector of terms such that $d = (t_1, t_2, \ldots, t_n)$, and a corresponding weights vector $w = (w_1, w_2, \ldots, w_n)$, where $w_1, w_2, \ldots, w_n$ are the weights of $t_1, t_2, \ldots, t_n$ respectively, based on the used term weighting scheme. VSM model can be visualized as a two-dimensional matrix, in which the rows represent documents and the columns represent terms (features) in the collection, as shown in Figure 1. In this section, we first introduce a simple example, then the mathematical formulations of the term weighting schemes in the scope of this study are listed, and finally, a brief discussion on each term weighting scheme is presented in addition to some calculations based on the given example.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

2.2.1 Working example

Consider the following collection of documents, each document contains many words separated by white spaces. This research considers each word as a single feature, so the feature space of this collection of documents will consist of the set of all words without replication, In Figure 1 the bolded words in the top row represent the feature space of this collection.

Document	Content
Doc 1	This document is short.
Doc 2	This content is dark.
Doc 3	This content in this document is dark.
Doc 4	We can detect this dark content, this terrorist content.

Figure 1 shows the representation of the collection in VSM where the raw frequency (term occurrence) is used as the weight *w*, hence, the empty cells indicate that the term is missing from the document. The collection frequency row in Figure 1 represents the summation of each term occurrences' in all documents. However, the Document Frequency row is the number of documents in which the term occurs, while the Document Length column represents the number of terms in the document without considering the repetition of the term, which is known also as the number of distinctive terms in the document. However, the bolded words at the top row in the figure represent the feature space.

Document \ Term	**can**	**content**	**dark**	**detect**	**document**	**in**	**is**	**short**	**terrorist**	**this**	**we**	Document Length*
Doc1					1		1	1		1		4
Doc2		1	1				1			1		4
Doc3		1	1		1	1	1			2		6
Doc4	1	2	1	1					1	2	1	7
Collection Frequency**	1	4	3	1	2	1	3	1	1	6	1	
Document Frequency***	1	3	3	1	2	1	3	1	1	4	1	

* The number of distinctive terms in the document.
** Summation of term occurrences in entire collection.
*** The number of document in which the term appears.

Figure 1 Collection of documents VSM representation

For more illustration on term weighting schemes, we will consider the above example as the case to be discussed in the following subsections, where brief discussions on the term weighting techniques used in this study are presented.

2.2.2 Mathematical formulas

As mentioned earlier, different term weighting schemes are based on different ideas (assumptions) about the important terms. However, this research considers five weighting schemes to be applied and combined as a hybridized feature selection method. Table 2 shows the mathematical formulation of term weighting schemes in the scope of this study, and summarized the based assumptions behind these schemes.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Table 2 Mathematical formulas of term weighting schemes in the scope of this study

Term weighting scheme	Formula	Important terms concept
Term frequency (TF)	$TF_{t,d} = \frac{fr_{t,d}}{\sqrt{\sum_{t=1}^{n} fr_{t,d}{}^2}}$	The more frequent term in the document indicates a more important term.
Document frequency (DF)	$DF_t = \sum_{d=1}^{N} \begin{cases} 1 & t \in d \\ 0 & t \notin d \end{cases}$	The term occurs in more different documents is the higher importance term.
Term frequency - Inverse document frequency (TF-IDF)	$TF - IDF_{t,d} = TF_{t,d} \cdot IDF_t$ where $IDF_t = \log\ (N/DF_t) + 1$	A less frequent term in the collection is a more significant term in the document, and in contrast, a more recurrent term in the collection, the less representative term of the document.
Glasgow	$w_{td} = \frac{\log(fr_{td}+1)}{\log(length_d)} \times \left(\log\left(\frac{N}{DF_t}\right)+1\right)$	A less frequent term in a short document is a more significant term than a term in a longer document.
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 $L_{td} = \begin{cases} 1+\log fr_{td}, & fr_{td} > 0 \\ 0, & fr_{td} = 0 \end{cases}$	A more frequent term is a more important term if occurs in fewer documents, taking the distribution of the term over the collection into account

Where in Table 2, the symbols have the following meanings:

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,

$length_d$ is the length of the vector that represents the distinctive terms of document *d*, and

F_t is the frequency of term *t* at the collection level.

However, these schemes were preferred and selected to be applied in this research because of their clear concepts about important terms and their popularity of usage in text classification and DW domains.

2.2.3 Term weighting discussions

i) Term frequency (TF)

In general, TF is concerned with the weight of a certain term in the document. In this research, the normalized TF (Salton and Buckley, 1988) is considered. The weight of the term *t* in document *d* (noted as TF_{td}) is calculated by finding the quotient of the raw frequency (occurrences) of the term *t* in the document *d*, and the Euclidean norm of the document *d* as shown in Table 2. Where the Euclidean norm of the document is defined as the square root of the summation of the square of frequencies of all terms in the document. TF weighting scheme

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

supports the assumption that more occurrences of a term in a shorter document indicate more significance of the term (Chen *et al.*, 2009; Salton and Buckley, 1988).

Based on the above definition, the collection of documents in the considered example is represented as in Figure 2, where the bolded cells indicate the highest TF value, and the underlined cells is the second highest value in the matrix that will be used for more illustration of section 3.2.

Term / Document	can	content	dark	detect	document	in	is	short	terrorist	this	we
Doc1					0.5000		0.5000	0.5000		0.5000	
Doc2		0.5000	0.5000				0.5000			0.5000	
Doc3		0.3333	0.3333		0.3333	0.3333	0.3333			**0.6667**	
Doc4	0.2774	0.5547	0.2774	0.2774					0.2774	0.5547	0.2774

Figure 2 VSM representation based on TF scheme

As seen in Figure 2, the highest TF value (0.6667) belongs to the term "this" in Doc3. The term "this" occurs two times in Doc3 and in Doc4. However, since TF is normalized to the document's Euclidean norm and Doc3 is shorter than Doc4, therefore, the TF value of term "this" is higher in Doc3, which reflects the assumption that more occurrences of a term in a shorter document indicates more significance of the term in the document.

TF weighting scheme and its different variants have been utilized in earlier researches in the domain of DW analysis such as (Al-Zaidy *et al.*, 2011; Alghamdi and Selamat, 2012; Chaurasia *et al.*, 2012; O'Callaghan *et al.*, 2013; Sun *et al.*, 2011; Wadhwa and Bhatia, 2013; Yang *et al.*, 2011; Zimbra and Chen, 2012), sentiment analysis (Abbasi *et al.*, 2008; Gohary *et al.*, 2013), and in text classification and categorization (Agrawal and Phatak, 2013; Choi and Kim, 2012; Gayathri and Marimuthu, 2013; Huang *et al.*, 2010; Iezzi, 2012).

ii) Document frequency (DF)

DF weight of a term *t* (denoted by DF_t) represents the number of documents within the collection in which the term *t* is found (Alghamdi and Selamat, 2012). DF is a global term weighting technique and based on the assumption that the terms that often occur frequently within the collection are more important (Chen *et al.*, 2009). Global term weighting techniques such as DF and IDF measure the term at the collection level, causing many terms to have the same weight, and the weight of certain terms in different documents to be the same too.

Based on the considered example, it can be seen from Figure 1, that the term "this" is the term with the highest DF value followed by the terms "content", "dark", and "is".

DF was utilized for detecting topics in Arabic DW content (Alghamdi and Selamat, 2012). However, in the domain of text mining and text classification, DF has been used in some studies such as (Joho and Sanderson, 2007) and (Rafrafi *et al.*, 2012).

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

iii) Term frequency-inverse document frequency (TF-IDF)

TF-IDF is an ultimate ranking measure that represents the terms within the collection of documents and reflects the assumption that a less frequent term in the collection is a more significant term in the document and vice versa. TF-IDF is used in many information-retrieval applications (Chianga *et al.*, 2008).

TF-IDF of a term *t* in document *d*, denoted by ($TF\text{-}IDF_{td}$), is the dot product of the term frequency (*TF*) and the inverse document frequency (*IDF*) of the term. Figure 3 shows the VSM matrix of the collection in the considered example based on the TF-IDF weight scheme.

IDF is a very important global statistical measure that supports the assumption that a more frequent term in the collection is considered to be less important. Additionally, IDF can be safely used to make the factor less insensitive, since it depends on the log function, which is a monotonically increasing function. Furthermore, IDF is always positive because the denominator (DF_t) is always less than or equal to *N* which represents the total number of documents in the collection.

However, as mentioned above, IDF is a global term weighting technique that measures the weight of a term on the level of the corpus, causing mostly thousands of terms to have the same IDF weight. This case makes it irrational to select for example the terms having the top 50, or 100 weights as the feature set for classification, so the IDF measure is not suitable to be used as a standalone feature selection technique as compared with other techniques in this study. However, it is mentioned here because of its importance in calculating the TF-IDF weight.

Term / Document	can	content	dark	detect	document	in	is	short	terrorist	this	we
Doc1					0.8466		0.6438	**1.1931**		0.5000	
Doc2		0.6438	0.6438				0.6438			0.5000	
Doc3		0.4292	0.4292		0.5644	0.7954	0.4292			0.6667	
Doc4	0.6618	0.7143	0.3571	0.6618					0.6618	0.5547	0.6618

Figure 3 VSM weights matrix based on TF-IDF scheme

In Figure 3, it is seen that the highest TF-IDF value in the matrix belongs to the term "short", which appears once in the collection. Even though, there are many other terms that appear once in the collection such as "detect" and "terrorist". However, the term "short" appears in a shorter document, therefore it achieved higher weight as the assumption behind the TF-IDF scheme demands.

TF-IDF weighting technique was utilized by (Ting *et al.*, 2013) for health professionals web information retrieval, and by (Paik, 2013) for effective ranking of terms, and for text clustering by (Iezzi, 2012), and by (Alghamdi and Selamat, 2012; Elovici *et al.*, 2005; L'Huillier *et al.*, 2010; Yang *et al.*, 2009) for DW analysis.

iv) Glasgow

Glasgow weighting scheme was proposed by (Sanderson and Ruthven, 1996). The main aim of this weighting scheme is to prevent terms in longer documents to be more favored because of the existence of many instances of insignificant terms in such documents. Figure 4 shows the Glasgow based weights matrix of the example in consideration.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Document \ Term	can	content	dark	detect	document	in	is	short	terrorist	this	we
Doc1					0.4952		0.3766	**0.6979**		0.2925	
Doc2		0.3766	0.3766				0.3766			0.2925	
Doc3		0.2067	0.2067		0.2718	0.3831	0.2067			0.2851	
Doc4	0.3002	0.2920	0.1620	0.3002					0.3002	0.2268	0.3002

Figure 4 Glasgow based VSM weights matrix

As seen in Figure 4, the term "short" is weighted the highest even as its frequency is low, however, other low-frequency terms such as "detect" and "terrorist" are weighted lower since they appear in the longer document.

v) Entropy

Entropy is the most sophisticated weighting scheme which is based on the probabilistic analysis and information-theoretic ideas. Entropy technique considers that a more frequent term is the more important term if occurs in less documents, taking the distribution of the term over the collection into account (Dumais, 1991; Selamat and Omatu, 2004; Wu *et al.*, 2008). Figure 5 shows the Entropy values of terms in corresponding documents of the example under consideration.

Document \ Term	can	content	dark	detect	document	in	is	short	terrorist	this	we
Doc1					1.0138		0.9289	1.2213		0.8968	
Doc2		0.9481	0.9289				0.9289			0.8968	
Doc3		0.9481	0.9289		1.0138	1.2213	0.9289			1.5183	
Doc4	1.2213	**1.6052**	0.9289	1.2213					1.2213	1.5183	1.2213

Figure 5 Entropy based VSM weights matrix

In Figure 5, the term "content" is weighted the highest based on Entropy weighting scheme.

2.3 Classification Techniques

2.3.1 Support vector machine (SVM)

Support Vector Machine (SVM) (Boser et al., 1992) is based on the procedure of learning a linear hyperplane from a training set that separates positive examples from negative examples, which represents the Dark and non-Dark documents in our case. The hyperplane is located at the point in the hyperspace that maximizes the distance between the closest positive and negative examples that known as support vectors. The linear classifier is based on two elements: the weight vector $\ddot{W}$ perpendicular to the hyperplane (which accounts for the training whose components represent feature), and a bias $\ddot{b}$ which determines the offset of the hyperplane from the origin. An unlabeled example $\ddot{x}$ is classified positive if $f(\ddot{x}) = \ddot{W}\ddot{X} + b \geq 0$, otherwise, it is classified as negative. Hence, SVM is a binary classifier. Many existing works have applied SVM as text classifier such as (Chen *et al.*, 2001; Joachims, 1998; Tong and Koller, 2002; Yang and Liu, 1999).

There are several advantages of SVM as text classifier. First, SVM can handle exponential or even infinitely many features because it does not have to represent examples in its transformed space, the only thing that needs to be computed efficiently is the similarity of two examples. Redundant features (that can be predicted

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

from other features), and high dimensional features are well-handled, thus SVM does not need an aggressive feature selection (Joachims, 1998).

The superiority of SVM classifier is proven in comparison with other classifiers such as C4.5 in the domain of DW analysis (Chen, 2007; Zheng *et al.*, 2006). Moreover, SVM has been shown to be one of the best performing and accurate classification approaches in many other different domains (Lee *et al.*, 2012).

2.3.2 K nearest neighbor (KNN)

The KNN is an effective, simple classification algorithm (Man et al., 2009), and used widely in the domain of text classification (Gayathri and Marimuthu, 2013; Harish *et al.*, 2010). KNN classifier defines the class of the test example based on the labels of the K closest neighbors of the training samples. The distance between the test example and its neighbors has been calculated effectively by the Euclidean distance(Selamat *et al.*, 2009), which is defined as the square root of the summation of the squared differences between corresponding features in feature vectors. However, the efficiency of KNN algorithm in terms of classification time, inversely proportional with the data dimensionality. This weakness comes from the fact that KNN does not have a real training phase, which causes a high computational cost at the classification (Man *et al.*, 2009);

2.3.3 Decision Trees (DT)

DT classifier is a tree in which internal nodes are labelled by terms (Mitchell et al., 1990); the branches departing from them are labelled by tests on the weight that the term has in the test document, and the leaves are labelled by categories. Such a classifier categorize a test document d_j by recursively testing for weight W_i by the term labeling the internal nodes having a vector, until a leaf node is reached; the label of this node is then assigned to the group. Such classifiers most frequently use binary document presentations, and thus take the form of binary trees.

DT uses a "divide and conquer" strategy to recursively partition and develops tree classifiers. The process of partitioning the leaves and branches recursively is repeated until the leaf of the tree generated contains predetermined terms based on a given training example.

DT has been applied as a text classifier by various researches (Apte *et al.*, 1998; Johnson *et al.*, 2002; Quinlan, 1986; Schapire and Singer, 2000; Vens *et al.*, 2006; Weiss *et al.*, 1999). The main advantage of DT is that it is simple to interpret and easy to be developed. In addition, DT is able to "grow" adaptively without retraining the sample data. Even though, DT involves training examples, when apply as a text classifier it functions, more like a keyword based filter.

2.3.4 Naive Bayes (NB)

The NB classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naïve) independence assumption. A more descriptive term for the underlying probability model would be independent feature model (Rennie et al., 2003). NB classifier can be illustrated by considering the problem of

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

classifying documents according to their content, such as dark and non-dark. The probability that a document d_j belongs to the class c_i can be written as:

$$\mathrm{P}(c_i|d_j) = \frac{p(c_i)\prod_{k=1}^{|T|} p(w_{kj}|c_i)}{p(d_j)}$$

Further derivation of Naïve Bayes can be referred to in the work of (Rennie et al., 2003) and (McCallum and Nigam, 1998).

Depending on the precise nature of the probability model, NB classifier can be trained very efficiently in supervised learning setting. In many practical applications, NB models are suitable and adapts well for text categorization as stated by (Koller and Sahami, 1997; Larkey and Croft, 1996; Lewis and Gale, 1994). In addition to its implementation simplicity, the training computational efficiency of NB algorithm is linear in both the number of instances and attributes. However, NB has some drawback such as it is only suitable for linear models, and it is based on statistical independence.

2.3.5 Extreme learning machine (ELM)

ELM is one of the Artificial Neural Networks (ANN) variants that is proposed to overcome some of ANN drawbacks (Liu et al., 2005) such as the over-fitting problem and the high computational time (Olatunji et al., 2010). ELM model is a feed-forward neural network that consist of one hidden layer in which the number of hidden nodes are chosen randomly, and the output weights are determined analytically. Given a Training set $\mathrm{X}= \{(x_i,t_i) \mid x_i \in R^n, t_i \in R^m, i=1,\ldots,\mathrm{N}\}$, hidden neuron number $\tilde{\mathrm{N}}$ and an activation function $g(z)$ then the ELM learning algorithm is performed by doing the following (Guang-Bin et al., 2004):

1. Initialize random input weights $\boldsymbol{W}$ and bias $\boldsymbol{B}$ to all hidden neurons.

2. Calculate output matrix H of the hidden layer.

3. Calculate the output weight $\beta = \mathrm{H}^{+}\mathrm{T}$.

Such that H^{+} is the Moore–Penrose generalized inverse of the hidden layer output matrix H, where H is the matrix in which the elements are the values of applying the activation function $g(z)$ on the input weights $\boldsymbol{W}$, bias $\boldsymbol{B}$ and the training set $\boldsymbol{X}$, such that

$$\mathrm{H}(W_{w=1}^{\tilde{N}}, B_{b=1}^{\tilde{N}}, X_{x=1}^{N}) = \begin{bmatrix} g(w_1.x_1+b_1) & \cdots & g(w_{\tilde{N}}.x_{\tilde{N}}+b_{\tilde{N}}) \\ \vdots & \ddots & \vdots \\ g(w_1.x_N+b_1) & \cdots & g(w_{\tilde{N}}.x_N+b_{\tilde{N}}) \end{bmatrix}_{N\times\tilde{N}}, \beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\tilde{N}}^T \end{bmatrix} \text{ and } \mathrm{T} = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}$$

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

However, further details on the workings of ELM algorithm can be found in (Guang-Bin et al., 2004; Huang et al., 2006).

3 Proposed Hybridized Term-Weighting method

This section presents the proposed hybridized term-weighting (HTW) method for accurate terrorist activities detection in dark web content. HTW hybridizes the feature sets selected based on TF, DF, TF-IDF, Glasgow, and Entropy term-weighting techniques. For the convenience of notation, the weighting techniques in this research are noted as A, B, C, D, and E which stands for TF, DF, TF-IDF, Glasgow, and Entropy respectively. Whereas, the UNION (UN) and Symmetric Difference (SD) functions were considered in hybridizing the Top K selected features sets. However, the hybridized feature set includes the important features based of different weighting scheme so as to take advantage of the strength of each of the these schemes while complimenting the weaknesses of other schemes. Figure 6 shows the framework of the proposed Hybridized term-weighting method.

3.1 The HTW framework

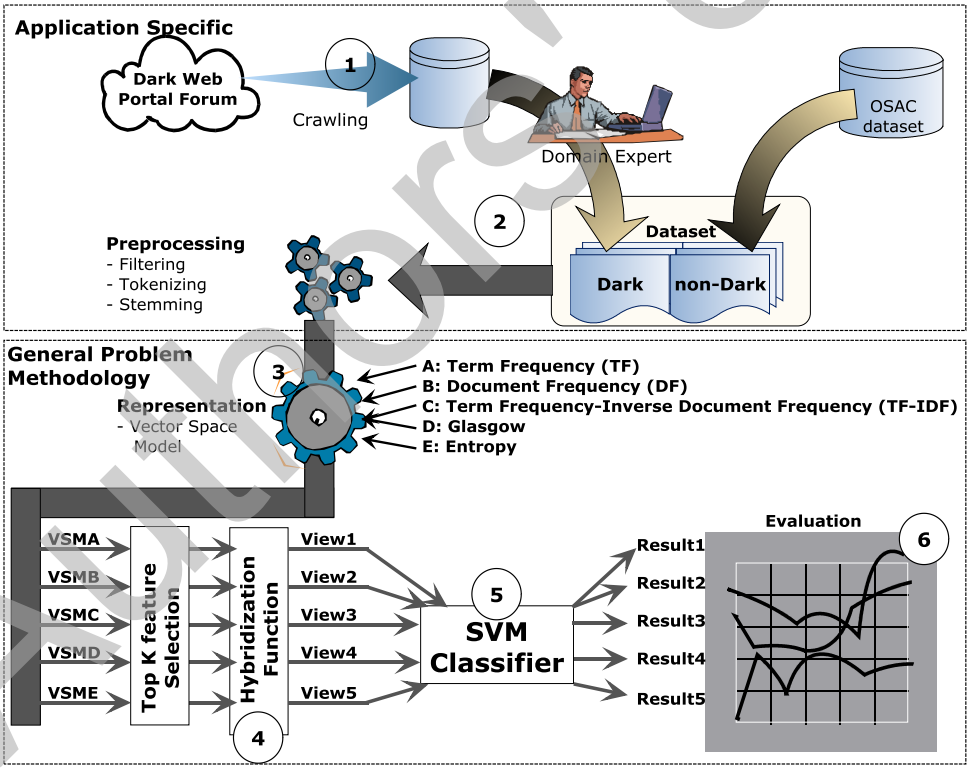


Figure 6 Hybridized term-weighting (HTW) framework

Figure 6 shows the framework of the proposed HTW method. Generally, the framework is divided into two parts; the upper part is Application Specific while the lower part is the General Problem Methodology. The Application Specific part contains the processes that are related to this research specifically, which are the crawling and the preprocessing of the dataset considered by this study. However, the General Problem

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Methodology contains the processes that showed followed in case of applying HTW method for solving the general text classification problem.

In the Application Specific processes, firstly, web pages were collected from the Internet (part 1 of the framework). For this purpose, a special crawler was developed to download and store the web pages from the Dark Web Forum Portal (DWFP). The implemented crawler consisted of many modules as shown in Figure 7.

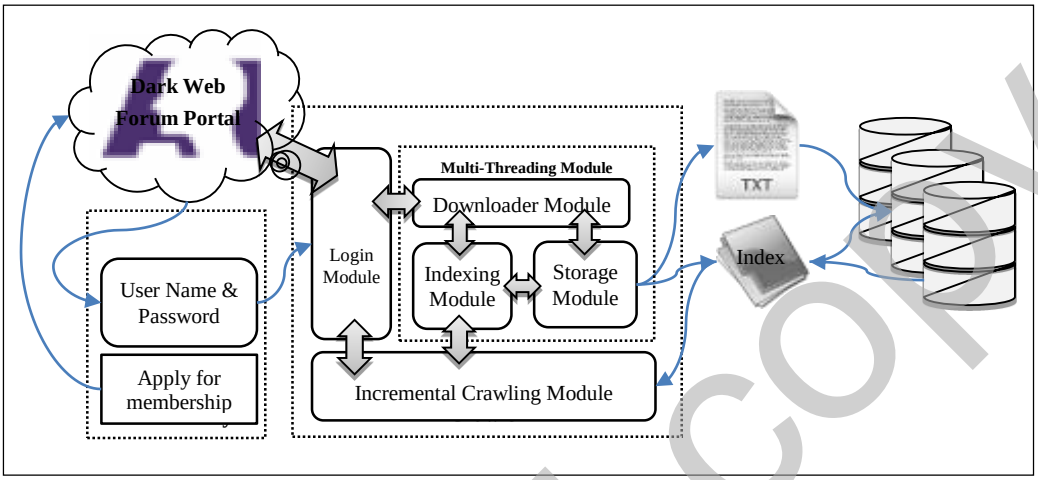


Figure 7 DWFP focused crawler

In the crawler, the (Login Module) is responsible for passing the obtained login information to DWFP and establishing the connection. Then, the (Download Module) in collaboration with Indexing and Storage modules works in a multithreaded environment to download, index, and store the documents from DWFP to the local storage media. The Incremental Crawling Module is designed based on the incremental updating approach (Cho and Garcia-Molina, 2000; Fu *et al.*, 2010) to resume downloading the documents in case of lost connection for any reason. The considered dataset has been constructed based on the downloaded documents and the Open Source Arabic Corpora (OSAC) (Saad and Ashour, 2010), as explained in section 4.1.

Then, pre-processing algorithms were applied..Pre-processing included filtering, tokenizing, and stemming (part 2 in Figure 6). Filtering involves removing the meaningless words (also known as stop words), removing non-Arabic characters, numbers, symbols, and special characters such as punctuations, Arabic diacritics, and other characters (Chisholm and Kolda, 1999; Gohary *et al.*, 2013; Last *et al.*, 2006). However, tokenizing means converting the document's text from one block (string) into separated strings in which each string consists of one word, while stemming is the process of removing suffixes, prefixes, and infixes from the words (Ceri *et al.*, 2013). After the stemming process, the words of the document's text are compounded again into one string, and then saved into a text file for further processing.

However, in the General Problem Methodology, the documents of the dataset are presented in VSM form where different term-weighting techniques are used in calculating terms' weights; for every term-weighting technique, the representation module generates a different VSM matrix (referred to as VSMA, VSMB, ..., VSME in the framework). Then, in part 4, the hybridization function is applied (hybridization functions are explained in

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Section 3.3), and then the classification is performed in part 5 using an SVM classifier (however, for benchmarking purposes, other classifiers are performed in this part), and finally results are evaluated.

3.2 Top K feature selection

As mentioned in section 2.2, VSM model can be visualized as a two-dimensional matrix, as shown in Figure 8. However, high data dimensionality is a well-known problem in text classification domain (Selamat and Omatu, 2004); therefore, Feature Selection (FS) and Feature Extraction (FE) methods are usually applied to reduce the data dimensionality caused by the huge number of distinctive terms (words) in the corpus (Bingham and Mannila, 2001; Lee *et al.*, 2008). FE methods such as Principal Component Analysis (PCA), Factor Analysis (FA), and Random Projection (RP) aim to reduce data dimensionality by eliminating the irrelevant and redundant features as much as possible. The use of these methods is not in the scope of this research. However, FS methods aim to identify and select the most significant features of the feature space. In this research, we apply the Top *K* weighted features as a feature selection method, in which the features that have the highest *K* weights are selected from the feature space to form the base of the classification.

Top K features selection can be explained as follows:

Let V be the feature space weighted matrix as shown in Figure 8, where *n* is the number of features in the space, and *m* is the number of documents in the corpus and w_{ij} is the weight of feature *i* in document *j*.

$$V = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & w_{ij} & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix}$$

Figure 8 Vector space model (VSM) matrix

Then, the output of Top K FS method will be the weighted matrix *O*, in which *K* is the number of selected features (represented as columns in the matrix) where *K* is much less than *n*. In the proposed method, Algorithm 1 is applied to each VSM matrix to construct the top *K* features of the VSM matrixes.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Algorithm 1: Top K feature selection from VSM matrix

- INPUT

$$V = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & w_{ij} & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix},$$

/ V is a matrix of size m × n, where m represents the number of documents in the collection, and n represents the number of collection's distinctive features. */*

/ w_{ij} represents the calculated weight of i^{th} term in j^{th} document. */*

/ K, number of features to be selected as Top where K << n*/*

- VARAIBLES

$R = [w_{11} \quad \cdots \quad w_{1n}]$ */* vector variable to store row of V matrix temporarily */*

$$RowsMaxs = \begin{bmatrix} w_{11} & w_{12} \\ \vdots & \vdots \\ w_{1m} & w_{2m} \end{bmatrix},$$

/ RowsMaxs is a matrix of two columns to store the maximum value of each term cross all documents, first columns is index column and second contains values. */*

mx */* temporary real variable */*

r,c,x */* temporary integer variables */*

- OUTPUT

$$O = \begin{bmatrix} w_{11} & \cdots & w_{1K} \\ \vdots & w_{ij} & \vdots \\ w_{m1} & \cdots & w_{mK} \end{bmatrix},$$

/ O is a matrix of size m × K, where m represents the number of documents in the collection, and K represent the number of selected features. */*

/ w, i, j represents the same as in input.*/*

- BEGIN

r ← 0 *// initial value*

Repeat

r ← r + 1 *//increment r variable*

R ← V_r *// store value of r^{th} row of matrix V into vector R*

mx ← max(R) // get the maximum value of Vector R

While *i < length (R)*

If (*R[i]=mx***)** RowsMaxs[ ] ← [*i,mx*] *// append column index and value in RowsMaxs matrix*

Loop

Until r >= m

SORT (RowsMaxs) *// sort RowsMaxs matrix based on mx values in descending order*

c ← 0 // reset to initial value

O←∅ *// initialize the output matrix*

Repeat

c ← c + 1 //increment r variable

x ← RowMaxs[r,1] *// store the index column from RowMaxs into x variable*

O_c ← V_x *// store x^{th} column of V matrix into c^{th} column of matrix O*

Until *c >= K*

END

In Algorithm 1, we start with the VSM matrix *V*. First, the maximum value of each row and its index are saved into the (*RowMaxs*) array, however if the maximum value is replicated in the row, then all corresponding indexes will be saved in the (*RowMaxs*) array. Then, the (*RowMaxs*) array is sorted in descending order so that the top K values can be taken from the top of the array. Lastly, the columns corresponding to the top K value indexes are taken from the matrix *V* and appended as columns to the output matrix. Based on the example shown in subsection 2.2.1, Figure 9 shows the top K feature set based on different term weighting schemes where K=2.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

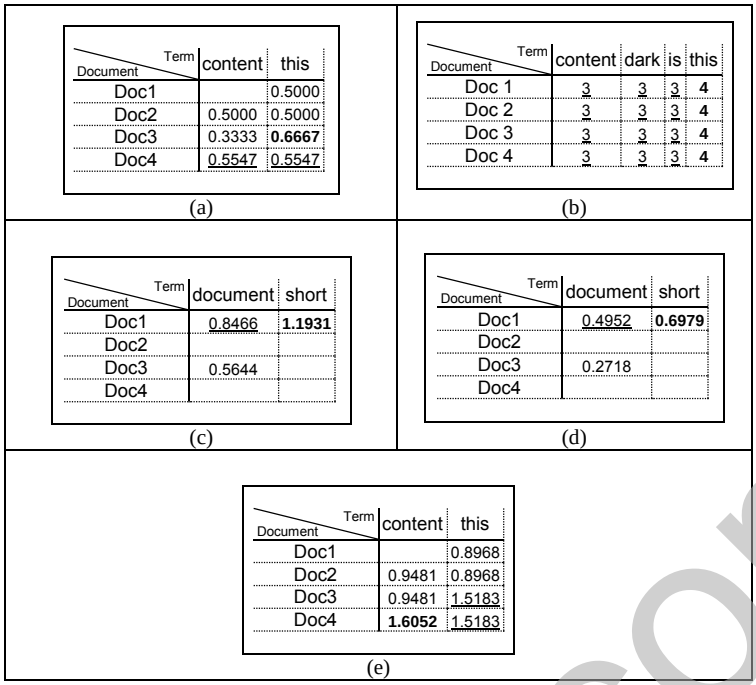


Figure 9 Top K feature set based on different term weighting schemes where (*K*=2)

In Figure 9, the output of applying Algorithm 1 (with $K = 2$) on the VSM matrixes based on different term weighting schemes is shown, where (a) is the output based on TF, (b) based on DF, (c) based on TF-IDF, (d) based on Glasgow, and (e) based on Entropy weighting scheme. It can be seen that in some cases, the selected features are the same such as in (c) and (d), however, in some other cases the features are different, since the assumptions about the significant terms are different.

3.3 Hybridization

In this research, two different hybridization functions are considered; UNION and Symmetric Difference functions to combine the sets of top K features selected from the VSM matrices in five different views. The aim of testing these two functions is to reduce the dimensionality of features in the combined feature set. The view number represents the level of hybridization, i.e. the number of sets to be combined, as illustrated in Figure 10.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

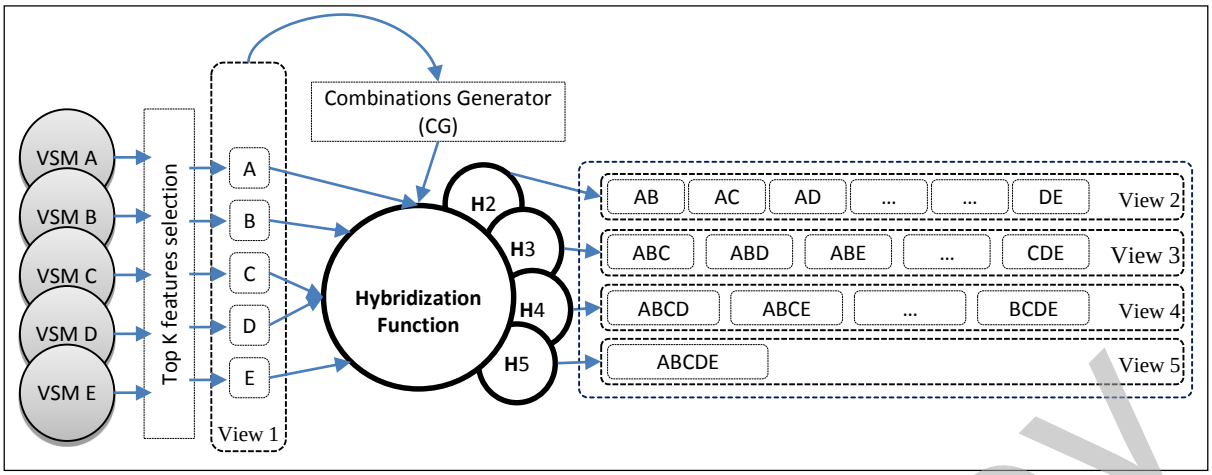


Figure 10 Top K feature sets hybridization illustration

During the hybridization process, the Combinations Generator (CG) generates the sequence of sets to be hybridized in each view, for example in View 2, feature sets would be hybridized in pairs. Therefore, the output of CG will be all possible combinations from two sets among the five sets (A, B, ..., and E) such as AB, AC, and so on. However, the order of sets in the combination is not important, i.e. there is no difference between for example AB and BA, since hybridizing functions are based on sets where the order of elements is not important. Therefore the number of combined feature sets in Views 2 to 5 can be calculated by the "Binomial Coefficient" as in equation (1), where n is the total number of VSM matrixes, which is 5, and r is the number of feature sets to be combined.

$$\text{Number of sets in View}(r) = C(n,r) = {}^{n}C_{r} = {}_{n}C_{r} = \binom{n}{r} = \frac{n!}{r!(n-r)!} \quad (1)$$

The following is the description of these views and the mathematical formulation of the hybridization function.

View 1 is the weighted matrix of the top K features selected to represent the document for each VSM model A, B, C, D, and E. The selected feature sets will be referred to as the set S= {*A, B, C, D, E*}, where

A is the set of top K features selected based on TF,

B is the set of top K features selected based on DF,

C is the set of top K features selected based on TF-IDF,

D is the set of top K features selected based on Glasgow,

E is the set of top K features selected based on Entropy,

and

View 2: Hybridization of feature sets A, B, C, D, and E based on **two** sets,

View 3: Hybridization of feature sets A, B, C, D, and E based on **three** sets,

View 4: Hybridization of feature sets A, B, C, D, and E based on **four** sets,

View 5: Hybridization of feature sets A, B, C, D, and E based on **five** sets.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

As mentioned before, two set operations are considered in this study as the hybridization functions. The following subsequent subsections show these functions and their mathematical formulation.

3.3.1 Union combination function

In mathematics, the union of two sets is the set of elements that are members of either of the two sets. The union operation is commonly denoted by∪. The union between two sets A and B is defined as the set of elements in A or in B or in both A and B. However, the union operation of multiple sets is defined as follows: Let S be the set of individual sets S_1,S_2,S_3,…S_n, then the combined set *M* is equal to $M = \bigcup_{i \in S} A_i, \{A_i \mid i \in I\}$, where the feature *x* belongs to *M* such that $x \in M \Leftrightarrow \exists A \in M, x \in A$.

3.3.2 Symmetric difference (SD) combination function

The SD function (commonly denoted by ∆ in mathematics) of two sets is defined as the set of elements that exist in one of the two sets but not in both. In other words, the symmetric difference of sets A and B is the elements that are in either of A or B and not in their intersection. However, the symmetric difference of n sets S_1, S_2, S_3,…, S_n, is expressed by $\Delta M = \left\{a \in \bigcup M : \left|\{A \in M : a \in A\}\right| \text{is odd}\right\}$, which means that *M* contains the elements that are in an odd number of the sets. The proof of the symmetric difference combination formula can be proven as follows: let ∆ denote the symmetric difference of two sets A and B, then

$$A \Delta B = (\mathrm{A} \cup \mathrm{B}) - (\mathrm{A} \cap \mathrm{B}) = (\mathrm{A} - \mathrm{B}) \cap (\mathrm{B} - \mathrm{A})$$

And the following theorem can be made:

Theorem. If A_1, …, A_n are sets, then

$$x \in (A_1 \Delta \cdots \Delta A_{n-1}) \Delta A_n \Leftrightarrow x \textit{ is in exactly an odd number of } A_i$$

$$\Leftrightarrow \left|\left|\{i \mid 1 \le i \le n, x \in A_i\}\right|\right| \textit{is odd}$$

Proof.

- We proceed by induction on *n*. The result is true if *n*=1 or 2.
- Assume the result holds for k. Then $x \in (A_1 \Delta \cdots \Delta A_n) \Delta A_{n+1}$ if and only if x is in exactly one of $A_1 \Delta \cdots \Delta A_n$ and ΔA_{n+1}.
- If x exists in an **even** number of sets from among $A_1, \cdots, A_{n+1}$, then it either exists in an even number of sets from among $A_1, \cdots, A_n$ and not in A_{n+1}; in which case it exists in neither $A_1 \Delta \cdots \Delta A_n$(by the induction hypothesis) nor in A_{n+1}., hence not in the symmetric difference; or else it exists in an odd number of sets from among $A_1, \cdots, A_n$ (and hence exists in

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

$A_1 \Delta \cdots \Delta A_n$ by the induction hypothesis) and in A_{n+1}, and so it does not exist in the symmetric difference (since it exists in both operands).

- If x exists in an **odd** number of sets from among $A_1, \cdots, A_{n+1}$, then it either exists in an even number of sets from among $A_1, \cdots, A_n$ (and hence not in $A_1 \Delta \cdots \Delta A_n$), and in A_{n+1}; or it exists in an odd number of sets from among $A_1, \cdots, A_n$ and not in A_{n+1}. Either way, it exists in exactly one of $A_1 \Delta \cdots \Delta A_n$ and A_{n+1}, hence exists in $(A_1 \Delta \cdots \Delta A_n) \Delta A_{n+1}$.
- Thus, $x \in (A_1 \Delta \cdots \Delta A_n) \Delta A_{n+1}$ if and only if it exists in exactly an odd number of sets from among $A_1 \Delta \cdots \Delta A_{n+1}$. □

Based on the example under consideration, the hybridized feature sets of the top 2 features selected by TF, DF, and TF-IDF weighting schemes are generated, Figure 11 shows the hybridized feature sets AB and AD.

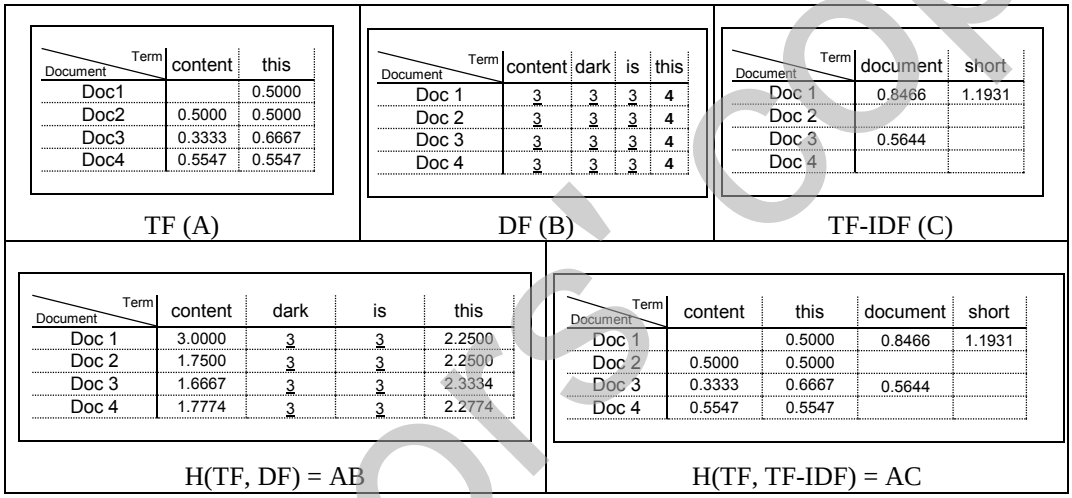


Figure 11 Example of hybridized feature sets based on TF, DF and TF-IDF weighting scheme

In Figure 11, the hybridization of the top 2 feature sets based on TF and TF-IDF is done by appending the columns of matrix C to the matrix A since there are no common features between A and C matrixes. However, in hybridizing TF with DF, the features "content" and "this" are common between the two matrixes; therefore, the average value of the weights is considered in the combined matrix.

3.4 Validation

In the classification part, and for benchmarking purposes, many of famous classifiers that are widely used in text classification were applied. Discussions on the classifiers used in this study are shown in section 2.3. However, as in text classification, many steps are performed for results validation; first, the weighting matrix is divided into two parts: one part for training and the other for testing. Then, the classifier uses the training data to build up the classification model (CM) during the learning phase. Then, the generated CM is used to label the testing data during the testing phase. Stratified *K*-fold cross validation method is commonly used with classification, in which samples are divided into *K* mutually exclusive (equal or approximately equal) subsets (known as folds) (Selamat and Ng, 2011), and then the classification process is run *K* rounds. However, as

Author's copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

stratified folding, each of the resulted folds will contain samples from both classes in proportions that are equal to the classes' proportions in the full dataset. In each round, the classifier uses *K*-1 folds as training data and the remaining fold as testing data. However, a specific fold could not be used for testing more than once. The results of each round are saved, and then merged together to get the final classification results. This method of validation has two main advantages; first is that all samples in the dataset are used for training, and every sample is used once for testing the model. Therefore, this research applied the stratified 10-fold cross-validation method with different benchmarked classifiers.

3.5 Evaluation

The famous information-retrieval measurements (accuracy, precision, recall, and F-measure) are used widely in terrorism-detection approaches (Choi *et al.*, 2014; Fu *et al.*, 2010; Greevy and Smeaton, 2004; Xianshan and Guangzhu, 2012; Zimbra and Chen, 2012). However, evaluating by precision and recall in isolation from each other does not make sense, as it is known that higher levels of precision may be obtained at the price of low values of recall (Man *et al.*, 2009). Therefore, the F-measure is normally used as the combination measure of precision and recall. In this research, the F-measure and accuracy measurements will be used to evaluate the proposed method; however, in the context of dark content detection evaluation, the following terms are used to indicate the corresponding definitions:

True Positive (TP): Number of dark documents correctly identified by the classifier.

False Positive (FP): Number of dark documents incorrectly identified by the classifier.

True Negative (TN): Number of non-dark documents correctly identified by the classifier.

False Negative (FN): Number of non-dark documents incorrectly identified by the classifier.

Moreover, the evaluation measures are calculated as follows:

- $Accuracy = \frac{|TP|+|TN|}{|TP|+|FP|+|FN|+|TN|}$
- $Precision(Dark) = \frac{|TP|}{|TP|+|FP|}$
- $Recall(Dark) = \frac{|TP|}{|TP|+|FN|}$
- $F-measure = 2 \times \frac{Precision \times Recall}{Precision+Recall}$

Moreover, the paired sample T-test and ANOVA statistical tests are employed to examine the significance of the differences between the number of features and the classification performance measurements within and between different views.

4 Experimental setup

This section shows the results of the conducted experiments and presents a detailed discussion of these results from different perspectives. As mentioned in section (3.3), The top K features were selected from the VSM of each term weighting scheme (A, B, …, E) as View 1, and then the other views (View 2, View 3, View 4, View 5) were created by combining these feature sets in pairs, trebles, quads, and quintuple. Table 4 shows the distribution of sets of different views, however, the order of the sets is not important, since the hybridization functions are set

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

functions, in which the result of combining the sets A with set B, for example, is the same result of combining set B with set A.

4.1 Dataset

To test and evaluate the HTW method, an experiment was conducted. First, thousands of Arabic web pages were downloaded from the DWFP. DWFP is the largest collection of crawled terrorist-related documents (Fu *et al.*, 2010). The data on DWFP is collected from 17 Arabic forums, and many other forums in other languages such as English, German, French, and Russian languages (Fu *et al.*, 2010). However, this research focuses on Arabic web pages. Therefore, a native Arabic domain expert labeled the web pages manually. The expert considered a document as dark, if the document contained material related to terrorist activities such as weapons or explosives manufacturing, attack planes, bombing, and other such activities, as shown in Figure 12**Error! Reference source not found.**. Then non-dark files were added from the Open Source Arabic Corpora (OSAC) (Saad and Ashour, 2010). OSAC dataset includes 22,429 Arabic text documents collected from many websites, where the documents are distributed into 11 categories (History, Entertainments, Economics, Education & Family, Religious discussions, Heath, Sports, Astronomy, Stories, Low, and Cooking Recipes).

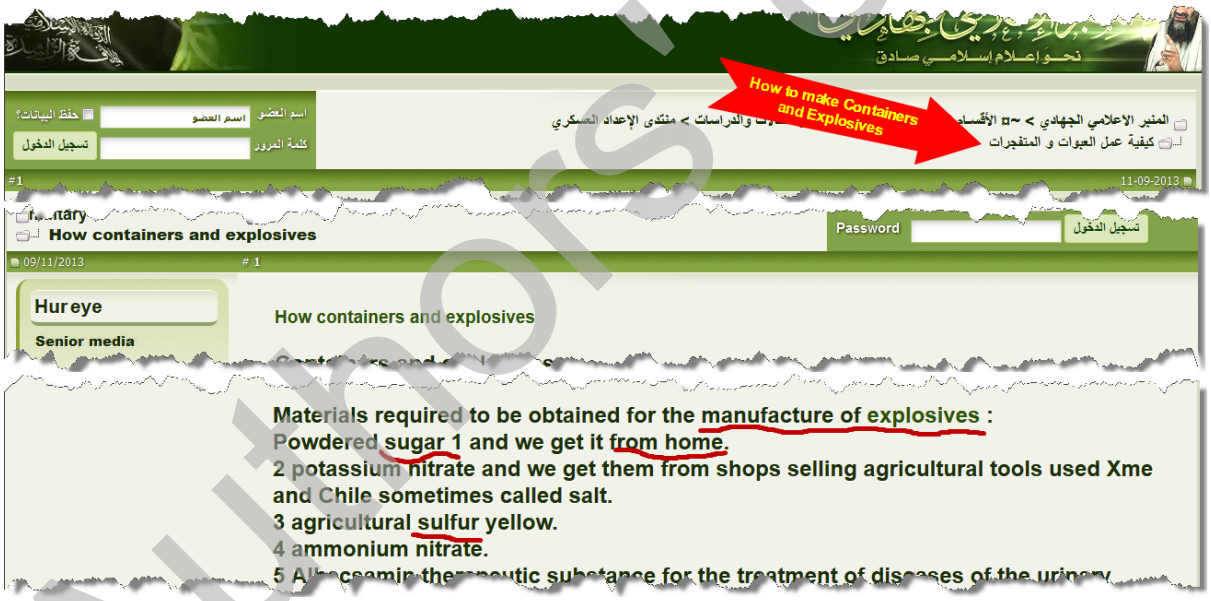


Figure 12 Dark content example

Figure 12**Error! Reference source not found.** shows a snapshot of an Arabic dark forum page besides its Google based translation, in which the poster explains in details about explosives manufacturing. Although there is no regular structure or format of dark content on the web, dark content is widely presented in text format (Fu *et al.*, 2010).

The dataset used for the experiments in this research is a balanced dataset, which consists of 1000 documents (50% of the documents are labeled as dark and 50% are non-dark). In the literature, many existing DW detection studies such as (Aknine *et al.*, 2005; Greevy and Smeaton, 2004) use a balanced dataset in which the

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

number of samples in both classes (Dark/non-Dark) are equal. However, in reality, the amount of dark content on the web is much less than the non-dark content (Correa and Sureka, 2013), therefore, the performance of dark content detection using imbalanced datasets will be the focus of our future work.

As in many of existing works, in this research, the DW detection problem is treated as a specific binary text classification problem, in which the dark content is to be distinguished from other topics. However, the dataset considered in this research reflects the general text classification problem as the dataset contains documents that belong to many various topics. Hence, even though application area of this research is the DW detection, the proposed method is still applicable and valid for the general text classification problem, as the general text classification problem is frequently treated as multi binary classification problems. Furthermore, the proposed method is a statistical text classification approach, therefore the focus on Arabic text in this research will not limit the generalization of the method, as text is represented and analyzed statistically, which is the same paradigm followed by general statistical text classification approaches regardless of text underlying language.

4.2 Pre-processing

Documents in the dataset were pre-processed. Pre-processing included stop-words removal based on the common Arabic stop word list[1], filtering against non-Arabic letters, numbers, symbols, and Latin text, and then stemming where the Larkey's Light Stemmer algorithm (Larkey *et al.*, 2007) is applied. The next step in the experiment was the representation of the documents in VSM model, based on different term-weighting techniques A, B, C, D, and E, which stands for TF, DF, TF-IDF, Glasgow, and Entropy weighting techniques, respectively. A special Java application was developed to perform the representation and weights calculation based on the Lucene 4.3 package[2]. As shown in Figure 6, the generated VSM matrices are named VSMA, VSMB, VSMC, VSMD, and VSME, and saved to the local storage to be used in the next step.

4.3 HTW application

To apply the HTW method, a series of experiments were conducted; the top 50 to 500 features (in intervals of 50 features) were selected from VSM matrixes as View 1, and then the remaining views were constructed by hybridizing the feature sets of View 1 based on two, three, four, and five sets. The total number of feature sets in different views is 31 sets of each top K features. The Rapid Miner software[3] (v5.3) was used to handle the VSM matrices, however, the classifiers were implemented in Matlab to perform the classification based on the parameters specified in **Error! Reference source not found.**. The results of this experiment are shown and discussed in the next section.

Table 3 Classifiers parameters

Classifier	Parameters
SVM	Kernel: Linear | default values for other parameters
KNN	K = 25 | Distance function : Euclidean

[1] http://arabicstopwords.sourceforge.net/
[2] http://lucene.apache.org/
[3] http://rapidminer.com/

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

DT	Default
NB	Default
ELM	Default

Table 4 Hybridized Feature sets in views

View	View 1	View 2	View 3	View 4	View 5
Feature Sets	A	AB	ABC	ABCD	ABCDE
	B	AC	ABD	ABCE	
	C	AD	ABE	ABDE	
	D	AE	ACD	ACDE	
	E	BC	ACE	BCDE	
		BD	ADE		
		BE	BCD		
		CD	BCE		
		CE	BDE		
		DE	CDE		
Number of sets in View	5	10	10	5	1

In this section, the classification performance results shown in subsections 0 to 5.5 are based on the SVM classifier, however, the benchmarking with KNN and ELM classifiers are shown in subsection 5.6.

5 Results and discussion

This section presents and discusses results of the conducted experiments from different point views.

5.1 Individual scheme classification performance

Figure 13 shows the classification performance of the term weighting schemes, where it can be seen that the schemes E and D which stand for Entropy and Glasgow schemes respectively are generally outperforming other schemes in F-measure, and Accuracy, however, TF-IDF scheme performance is the least in the same perspectives.

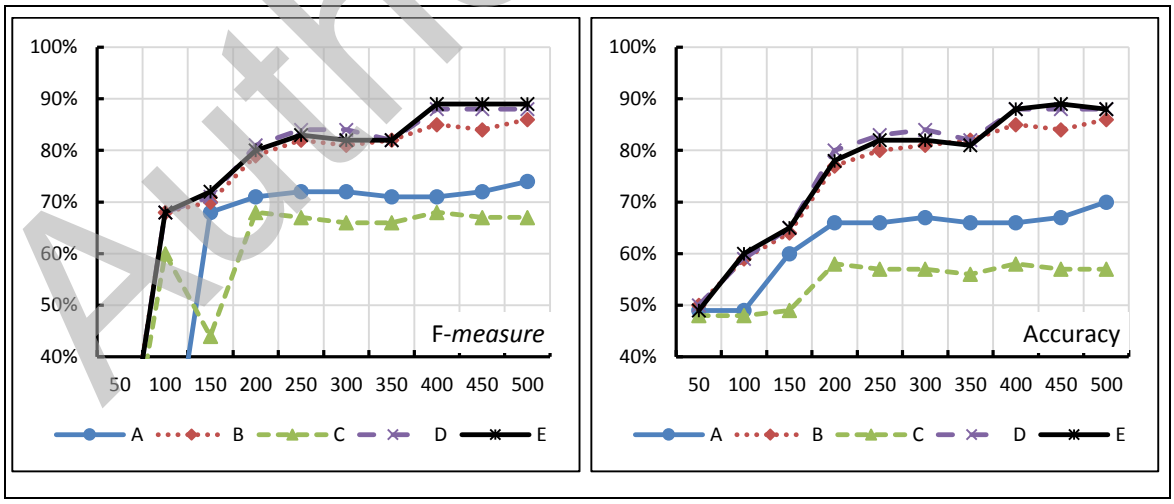


Figure 13 Individual term weighting techniques classification performance

In addition, it can be seen that the performance of all schemes was poor and unstable for less than 200 features, however, as the number of features increases, the classification performance in terms of f-measure and

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

accuracy gradually improved and stabilized. The highest classification accuracy of 89% is achieved by scheme E based on 450 features in the feature set.

5.2 Comparison of number of features based on hybridization function

In this research two hybridization functions were applied; the UNION (UN) and the Symmetric Difference (SD). Figure 14 shows the difference between the numbers of features in the hybridized feature sets.

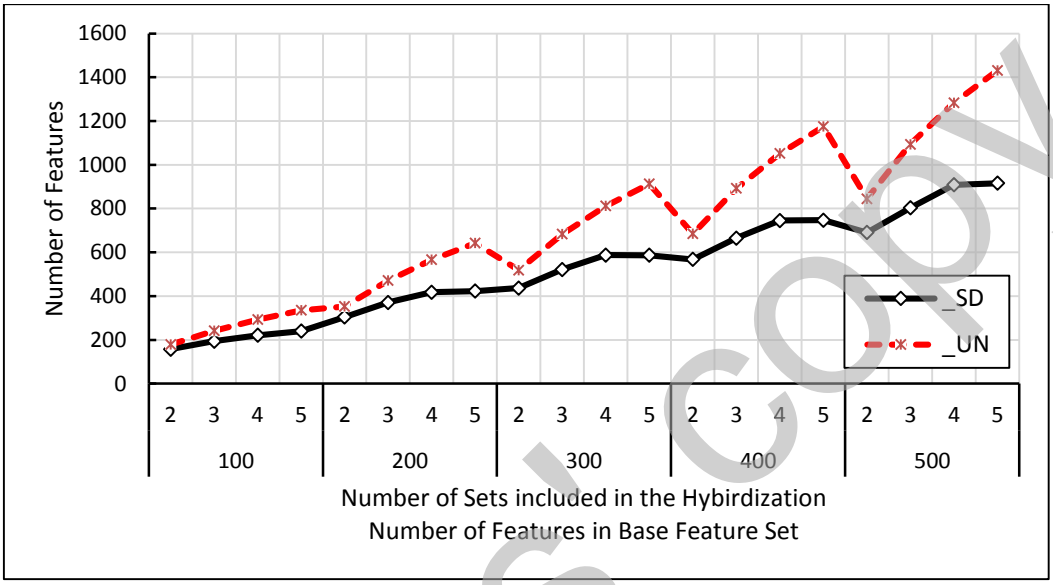


Figure 14 Number of features in the hybridized feature sets based on different combination functions

It can be seen from Figure 14 that, in general, the number of features in the hybridized feature set generated by the SD combination function is less than the number based on the UN function. Moreover, the difference in features increases as the number of sets included in the hybridization increases, and as the number of features in the base feature set increases. The reason behind this is that the UN function does not remove the common features between the feature sets included in the combination. However, the SD function removes the features that occur in an even number of sets based on the definition of the SD function as discussed in subsection 3.3.2.

5.3 View based classification performance comparison based on hybridization function

Figure 15 shows the comparison of view based classification performance in terms of F-measure based on different hybridization functions, in which for each sub-graph the x-axis represents the number of features in the combined feature sets, and each curve represents a different combination of term weighting techniques.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

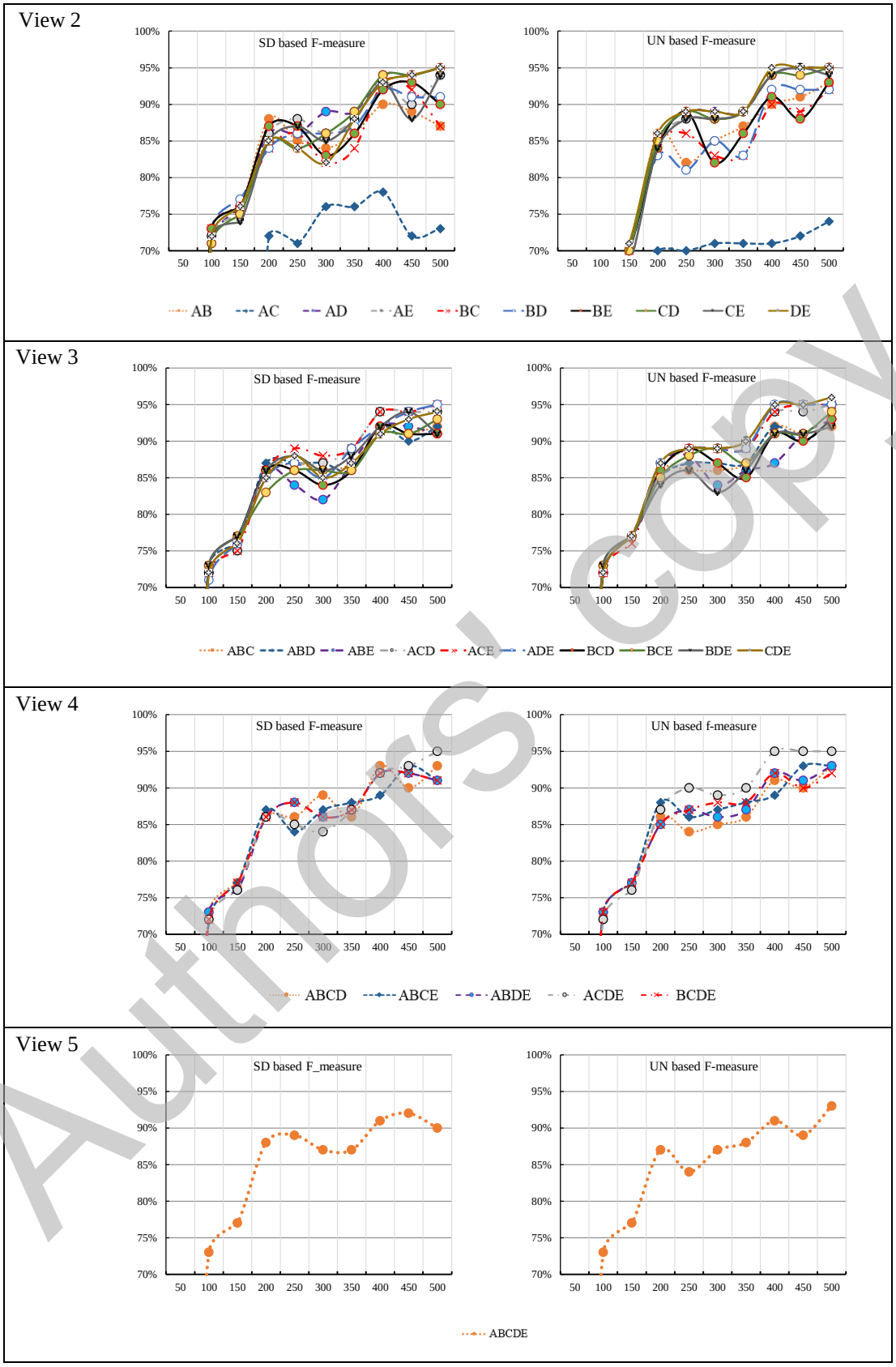


Figure 15 View based classification F-measure performance based on different hybridization function

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

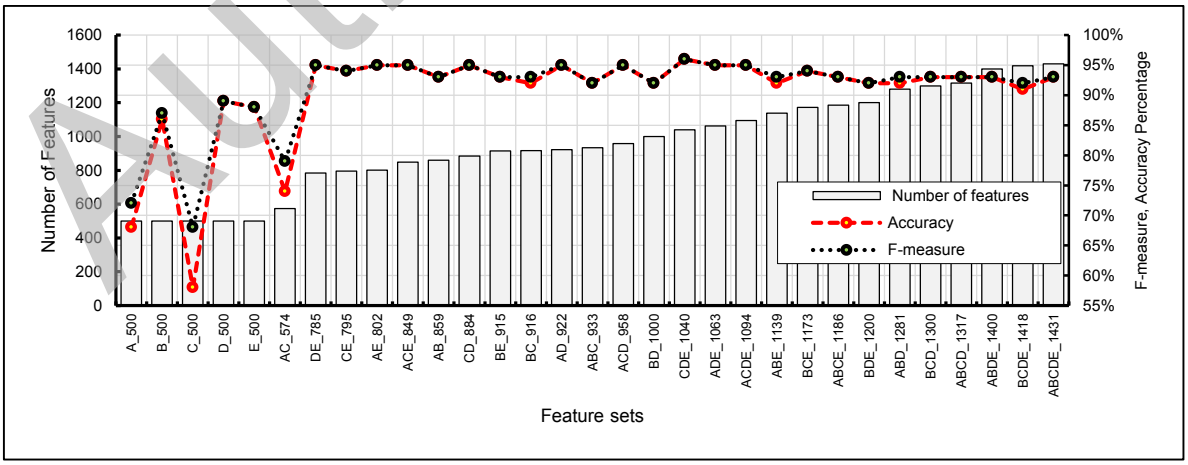
From Figure 15 it can be seen that the classification F-measure performance based on the combined term weighting techniques is higher than the F-measure based on individual schemes (Figure 13). Moreover, it is clear that, in general, the classification performance based on UN hybridizing function outperforms the performance based on SD function. The higher values of F-measure based on the UN function are due to the larger number of features included in the combined feature set as shown in Figure 14. However, the difference in F-measure between SD and UN based classification is either not found or not high as the difference between the numbers of features shown in Figure 14. For example, the F-measure value based on the DE feature set based on 500 features for both SD and UN hybridization functions is the same. However, from Figure 14 it can observed that the difference in the number of features based on combining two sets of 500 features in each by the SD and the UN is about 150 features on average.

In addition, it is seen in general that the classification performance is increasing, as the number of combined sets increases, due to the combined feature set which becomes more discriminant since the included features are coming based on a wide range of assumptions about the significant terms in the text, which increases the classifier distinguishing ability (Chen et al., 2009).

Furthermore, it is noticeable that the F-measure performance based on SD and UN functions for some combinations (e.g. CD, CE, and DE) in View 2 is higher than the F-measure performance in View 5 where five feature sets are combined. This case shows that even the feature set in View 5 is based on different assumptions about the significant terms but still it contains some noise features that affect the classifier performance.

5.4 Hybridized feature sets based classification performance comparison

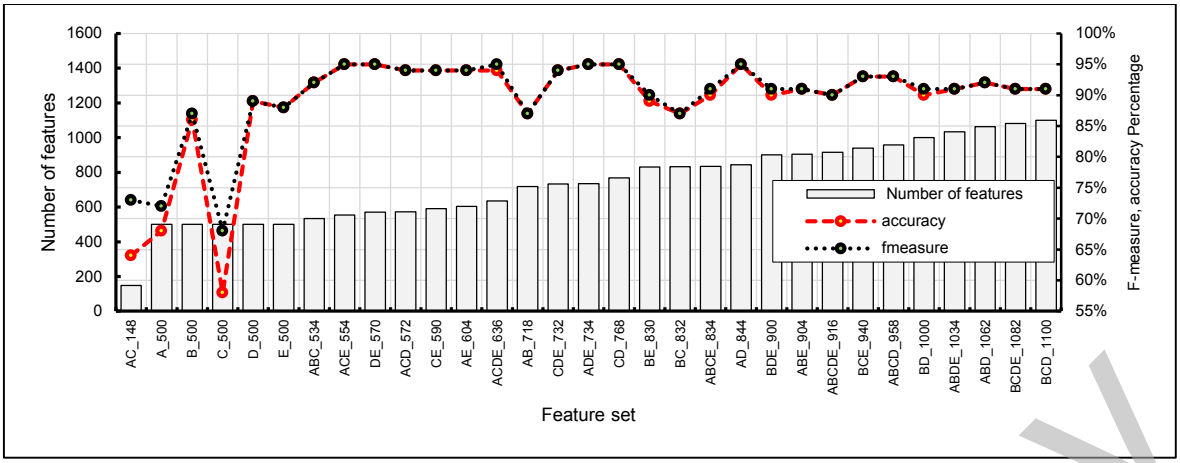
Figure 16 shows the comparison of classification performance of different feature sets at the level of 500 features, where the subfigure (a) shows the comparison based on UN hybridization function and (b) shows the comparison based on SD function. In both subfigures, the right vertical axis shows the percentage of classification performance measurements, while the left vertical axis shows the number of features in the feature set, and the horizontal axis shows the individual feature sets and all other hybridized feature sets generated in different views.



(a) UN function based hybridization

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X



(b) SD function based hybridization

Figure 16 Feature set classification performance comparison

It can be seen in Figure 16, that for UN based hybridization feature sets the maximum F-measure and accuracy value achieved is 96.00% on the levels of 1040 features by the hybridized feature set CDE. However, for SD based feature sets, the maximum classification performance measurement value achieved is 95.00% on the levels of 554, 570, 734, 768, and 844 features by the hybridized feature set ACE, DE, ADE,CD, and AD respectively. Furthermore, it is seen that the uppermost number of features in the UN based hybridized feature sets belongs to the ABCDE set (1431 features). However, for the SD based hybridized feature sets, the set ABCDE does not contain the highest number of features and does not achieve the maximum classification accuracy, because of the exclusion of some features from the hybridized feature set by the SD combination function. These results indicates ability of SD function to reduce the feature space dimensionality while achieving approximately the same classification performance based on UN function where the feature space is of high dimension. The aim of the proposed hybridized method is to achieve such high classification performance based on fewer features. Therefore, it can be concluded that the hybridization of the term weighting schemes TF, TFIDF, and Entropy (i.e. feature set ACE) has achieved the best performance among the SD based hybridized feature sets.

Moreover, it is noticeable that the performance measurements are decreasing while the number of features is increasing especially for the UN based hybridization. This case indicates that the included features in the large feature sets cause less discrimination ability to the classifier.

In addition to the discussions based on the values of evaluation measurements, and because the differences between these values sometimes are very small, the differences are statistically tested to determine the significance of these results, as explained in the next subsection.

5.5 Statistical significance tests

In this subsection, the statistical tests are used to examine the significance of the differences in the means of classification performance on the one hand, and the differences in the number of features in the combined feature sets based on different hybridization function, on the other hand, in different views.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

5.5.1 Statistical test within views

Table 5 shows the results of the paired sample T-test for classification performance measurements and the number of features within views.

Table 5 Paired-sample T-test for classification performance measurements and number of features within views

View	Pair		t	df	P value
View 2	SD f-measure	– UN f-measure	-1.26271	99	0.20966
	SD accuracy	– UN accuracy	-5.19673	99	0.00000*
	SD features	– UN features	-9.08045	99	0.00000*
View 3	SD f-measure	– UN f-measure	-3.81888	99	0.00023*
	SD accuracy	– UN accuracy	-3.47314	99	0.00076*
	SD features	– UN features	-15.10313	99	0.00000*
View 4	SD f-measure	– UN f-measure	-1.57796	49	0.12101
	SD accuracy	– UN accuracy	-1.32542	49	0.19118
	SD features	– UN features	-12.48351	49	0.00000*
View 5	SD f-measure	– UN f-measure	0.72761	9	0.48535
	SD accuracy	– UN accuracy	0.60984	9	0.55705
	SD features	– UN features	-5.72256	9	0.00029*

df: degree of freedom, * Significant at 0.01 level

The statistical test results in Table 5 shows that the difference in the number of features between SD and UN based feature sets is significant in all views at 0.01 alpha level. Furthermore, the difference between SD and UN based f-measure and accuracy is statistically not significant in View 4 and View 5, however, the differences between SD and UN based accuracies in View 2 and View 3 are statistically significant at 0.01 level with p-value of 0.00000 and 0.00076 respectively. Moreover, the differences between SD and UN based f-measure means are not significant in all views except for View 3 where the p-value is 0.00023.

5.5.2 Statistical test between views

The ANOVA statistical test is applied to test the significance of the differences in classification performance measurements (i.e. f-measure and accuracy) means among different views, where the SD based classification performance measurements are considered. The results of this test are shown in Table 6, where it can be seen that the differences between the means of f-measure achieved by different views are statistically significant at any number of features in the feature sets, except at the level of 400 features.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Table 6 ANOVA statistical significance test for classification performance between groups

	F-measure			Accuracy		
Number of Features	F	p-Value	F critical	F	p-Value	F critical
50	4.01224	0.01151**	4.13996	180.95161	0.00000*	4.13996
100	4.64836	0.00578*	4.13996	9.22356	0.00009*	4.13996
150	6.03965	0.00142*	4.13996	4.02508	0.01135**	4.13996
200	6.80489	0.00069*	4.13996	5.02882	0.00388*	4.13996
250	4.88678	0.00450*	4.13996	3.37680	0.02365**	4.13996
300	4.48402	0.00688*	4.13996	3.68321	0.01665**	4.13996
350	7.76648	0.00029*	4.13996	5.16277	0.00338*	4.13996
400	0.88555	0.48632	4.13996	4.55177	0.00640*	4.13996
450	3.23630	0.02785**	4.13996	3.38814	0.02334**	4.13996
500	6.35738	0.00105*	4.13996	3.23364	0.02793**	4.13996

* Significant at 0.01 level
** Significant at 0.05 level

Moreover, it is seen from Table 6, that the differences in means of accuracies achieved by different views are statistically significant at any number of features in the feature sets.

5.6 Benchmarking among classifiers

Table 7 shows the cross classifier benchmarking averaged performance measures, where the base feature sets consist of 500 features for each. The results of hybridization using different functions and levels are also shown. However, the hybridization level represents the number of combined sets.

Table 7 ANOVA statistical significance test for classification performance between groups

				Classifier				
Performance Measure	Hybridization Function	Hybridization level	Mean number of features*	SVM	KNN	DT	NB	ELM
F-measure	SD	1	500	80.80	85.65	37.82	7.43	83.42
		2	690	90.10	86.48	44.09	1.88	84.76
		3	803	92.80	85.68	49.14	6.49	85.44
		4	909	92.20	88.94	41.74	11.04	86.97
		5	916	90.00	87.89	54.55	15.52	86.47
	UN	1	500	80.80	85.65	37.82	7.43	83.42
		2	845	92.40	87.04	54.40	8.72	87.72
		3	1094	93.80	88.49	46.43	14.27	89.31
		4	1283	93.20	90.14	53.28	16.27	89.98
		5	1431	93.00	92.25	58.33	18.12	90.43
Accuracy	SD	1	500	77.80	83.88	50.00	50.48	81.82
		2	690	89.00	85.08	50.00	49.79	83.36
		3	803	92.70	83.71	50.00	50.53	84.66
		4	909	91.80	88.00	50.00	51.56	86.42
		5	916	90.00	87.10	50.00	53.20	86.60
	UN	1	500	77.80	83.88	50.00	50.48	81.82
		2	845	91.80	84.99	50.00	49.77	87.10
		3	1094	93.60	86.78	50.00	50.16	88.96
		4	1283	93.00	88.98	50.00	50.60	89.84
		5	1431	93.00	91.80	50.00	51.20	90.50

* Rounded number of features

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

It is seen in Table 7 that the SVM classifier, followed by the KNN and the ELM, achieved the highest performance measures. These results emphasize the claim of superiority of SVM over other classifiers in the domain of DW analysis (Chen, 2007; Zheng *et al.*, 2006). However, a more important conclusion that can be read from Table 7 is that the performance measures are proportionally increasing as the increase in the level of hybridization, in general. This result puts emphasize on our claim that the use of a feature set generated based on different ideas of significant terms will help in achieving higher classification performance. Moreover, it is noticeable that classification performance based on the UN hybridization function are higher than those based SD function, however the mean numbers of features are much less for the sets combined by the SD function than the sets combined by the UN function.

The classification performance measures achieved based on the NB and the DT classifiers are poor as seen in Table 7, these poor results are due to the high sparsity of feature sets. For more clarification of this point, it is known that DT classifier prefers the inputs where the attributes have many values (Quinlan, 1986). However, the NB classifier also suffers from some major back draws such as over-fitting and underflow that cause the low performance (Chandra and Gupta, 2011).

6 Conclusion and future work

This paper introduced the HTW method for accurate terrorism activities detection in textual content. Using the different term-weighting techniques, HTW combines small feature sets generated by the basic TF, DF, IDF, TF-IDF, Glasgow, and Entropy term-weighting techniques into one feature set in different views. Moreover, the UNION and Symmetric Difference combination functions were utilized as the hybridization functions. A selected balanced dataset downloaded from the DWFP is used in testing, evaluating and benchmarking the proposed method using many common text classifiers. The results show that it is possible to achieve higher classification performance by combining few small feature sets based on different assumptions about the most significant terms in the text. In addition, results show that the number of features in the combined feature set based on the Symmetric Difference function is significantly less than the number of features in the UNION function based feature set. Moreover, the benchmarking results support our assumption that a higher classification performance could be achieved by combining feature sets generated based on different assumptions about the most significant terms in the text. Our future work will focus on proposing a modified term-weighting technique for achieving higher DW content detection accuracy in balanced and unbalanced datasets.

Acknowledgments

The Universiti Teknologi Malaysia (UTM) under research grant 03H02 and Ministry of Science, Technology & Innovations Malaysia, under research grant 4S062 are hereby acknowledged for some of the facilities utilized during the course of this research work. Moreover, The AL-QUDS OPEN UNIVERSITY – PALESTINE is acknowledged for supporting and funding the first author during his PhD study. Additionally, Dr. Mahmood Ashraf, Department of Computer Science, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan, is acknowledged for his suggestions and contributions to this work.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

References

Abbasi, A., and Chen, H. (2005). Applying Authorship Analysis to Extremist-Group Web Forum Messages. *IEEE Intelligent Systems, 20*(5), 67-75.

Abbasi, A., and Chen, H. (2007). *Affect intensity analysis of dark web forums*. Proceedings of the 2007 IEEE international conference on Intelligence and Security Informatics (ISI 2007). May 23-24, 2007. New Brunswick, NJ, United states, 282-288.

Abbasi, A., and Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *Acm Transactions on Information Systems, 26*(2), 7.

Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *Acm Transactions on Information Systems, 26*(3), 12.

Agrawal, R., and Phatak, M. (2013). *A novel algorithm for automatic document clustering*. Proceedings of the 2013 IEEE 3rd International Advance Computing Conference (IACC). February 22-23, 2013. Ghaziabad, India, 877-882.

Aknine, S., Slodzian, A., and Quenum, G. (2005). Web personalisation for users protection: a multi-agent method. In B. Mobasher, and S. S. Anand (Eds.). *Intelligent Techniques for Web Personalization* (pp. 306-323): Springer.

Al-Zaidy, R., Fung, B. C. M., and Youssef, A. M. (2011). *Towards discovering criminal communities from textual data*. Proceedings of the 2011 ACM Symposium on Applied Computing. March 21-25, 2011. TaiChung, Taiwan, 172-177.

Alghamdi, H. M., and Selamat, A. (2012). *Topic detections in Arabic Dark websites using improved Vector Space Model*. Proceedings of the 4th Conference on Data Mining and Optimization (DMO 2012). September 2-4, 2012. Langkawi, Malaysia, 6-12.

Apte, C., Damerau, F., and Weiss, S. M. (1998). *Text Mining with Decision Trees and Decision Rules*. Proceedings of the Conference on Automated Learning and Discorery. June 11-13, 1998. Pittsburgh, PA, USA, 1-4.

Bharti, K. K., and Singh, P. K. (2014). A three-stage unsupervised dimension reduction method for text clustering. *Journal of Computational Science, 5*(2), 156-169.

Bharti, K. K., and Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications, 42*(6), 3105-3114.

Bingham, E., and Mannila, H. (2001). *Random projection in dimensionality reduction: applications to image and text data*. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. . August 26 - 29, 2001. San Francisco, CA, USA, 245-250.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*. Proceedings of the fifth annual workshop on Computational learning theory. July 27 - 29, 1992. Pittsburgh, PA, USA, 144-152.

Boutemedjet, S., Bouguila, N., and Ziou, D. (2009). A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(8), 1429-1443.

Ceri, S., Bozzon, A., Brambilla, M., Valle, E., Fraternali, P., and Quarteroni, S. (2013). An Introduction to Information Retrieval. In S. Ceri, A. Bozzon, M. Brambilla, E. Valle, P. Fraternali, and S. Quarteroni (Eds.). *Web Information Retrieval* (pp. 3-11). Berlin, Germany: Springer Berlin Heidelberg.

Chandra, B., and Gupta, M. (2011). Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data. *Expert Systems with Applications, 38*(3), 1293-1298.

Chaurasia, N., Dhakar, M., Chharia, A., Tiwari, A., and Gupta, R. (2012). *Exploring the Current Trends and Future Prospects in Terrorist Network Mining*. Proceedings of The Second International Conference on Computer Science, Engineering and Applications (CCSEA 2012). May 26-27, 2012. Delhi, India, 379–385.

Chen, C.-M., Lee, H.-M., and Chang, Y.-J. (2009). Two novel feature selection approaches for web page classification. *Expert Systems with Applications, 36*(1), 260-272.

Chen, D., Bourlard, H., and Thiran, J.-P. (2001). *Text identification in complex background using SVM*. Proceedings of the International Conference on Computer Vision and Pattern Recognition. December 8-14, 2001. Kauai, Hawaii, USA, 621-626.

Chen, H. (2007). *Exploring extremism and terrorism on the web: The Dark Web project*. Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics (PAISI 2007). April 11-12, 2007. Chengdu, China, 1-20.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Chen, H. (2008a). *IEDs in the Dark Web: Genre classification of improvised explosive device web pages*. Proceedings of the 2008 IEEE International Conference on Intelligence and Security Informatics (ISI 2008). June 17-20, 2008. Taipei, Taiwan, 94-97.

Chen, H. (2008b). *Sentiment and affect analysis of Dark Web forums: measuring radicalization on the internet*. Proceedings of the 2008 IEEE International Conference on Intelligence and Security Informatics (ISI 2008). June 17-20, 2008. Taipei, Taiwan, 104-109.

Chen, H., Chung, W., Qin, J., Reid, E., Sageman, M., and Weimann, G. (2008). Uncovering the Dark Web: A case study of Jjihad on the Web. *Journal of the American Society for Information Science and Technology, 59*(8), 1347-1359.

Chianga, D.-A., Keha, H.-C., Huanga, H.-H., and Chyrb, D. (2008). The Chinese text categorization system with association rule and category priority. *Expert Systems with Applications, 35*(1-2), 102-110.

Chisholm, E., and Kolda, T. G. (1999). New term weighting formulas for the vector space method in information retrieval. *Computer Science and Mathematics Division, Oak Ridge National Laboratory*.

Cho, J., and Garcia-Molina, H. (2000). *The Evolution of the Web and Implications for an Incremental Crawler*. Proceedings of the 26th International Conference on Very Large Data Bases. September 10-14, 2000. Cairo, Egypt, 200-209.

Choi, D., and Kim, P. (2012). Automatic Image Annotation Using Semantic Text Analysis. In G. Quirchmayr, J. Basl, I. You, L. Xu, and E. Weippl (Eds.). *Multidisciplinary Research and Practice for Information Systems* (Vol. 7465, pp. 479-487). Berlin, Germany: Springer Berlin Heidelberg.

Choi, D., Ko, B., Hwang, M., and Kim, P. (2011). Building Knowledge Domain N-Gram Model for Mobile Devices. *Information-an International Interdisciplinary Journal, 14*(11), 3583-3590.

Choi, D., Ko, B., Kim, H., and Kim, P. (2014). Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications, 38*, 16-21.

Corbin, J. (2003). *Al-Qaeda: In Search of the Terror Network That Threatens the World*. New York, USA: Thunder Mouth Press/Nation Books.

Correa, D., and Sureka, A. (2013). Solutions to Detect and Analyze Online Radicalization : A Survey. *CoRR, abs/1301.4916*.

Crestani, F., Sandersony, M., Theophylactou, M., and Lalmas, M. (1998). Short queries, natural language and spoken document retrieval: Experiments at Glasgow University. In: E.M. Voorhees and D.K. Harman (eds), The Sixth Text REtrieval Conference (TREC-6), 667–86. [NIST Special Publication 500–240] Available at: http://trec.nist.gov/pubs/trec6/papers/glasgow.ps.gz (accessed 5 December 2005).

Dumais, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers, 23*(2), 229-236.

Efron, M., Zhang, J., and Marchionini, G. (2003). *Comparing feature selection criteria for term clustering applications*. Proceedings of ACM SIGIR 2003. July 28 - August 1, 2003. Toronto, Canada, 28-31.

El Akadi, A., Amine, A., El Ouardighi, A., and Aboutajdine, D. (2011). A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowledge and Information Systems, 26*(3), 487-500.

Elovici, Y., Shapira, B., Last, M., Zaafrany, O., Friedman, M., Schneider, M., and Kandel, A. (2005). Content-Based Detection of Terrorists Browsing the Web Using an Advanced Terror Detection System (ATDS). In P. Kantor, G. Muresan, F. Roberts, D. Zeng, F.-Y. Wang, H. Chen, and R. Merkle (Eds.). *Intelligence and Security Informatics* (Vol. 3495, pp. 244-255). Berlin, Germany: Springer Berlin Heidelberg.

Fu, T., Abbasi, A., and Chen, H. (2010). A focused crawler for dark web forums. *Journal of the American Society for Information Science and Technology, 61*(6), 1213-1231.

Gayathri, K., and Marimuthu, A. (2013). *Text document pre-processing with the KNN for classification using the SVM*. Proceedings of the 7th International Conference on Intelligent Systems and Control (ISCO). January 4-5, 2013. Tamil Nadu, India, 453-457.

Gohary, A. F. E., Sultan, T. I., Hana, M. A., and Dosoky, M. M. E. (2013). A Computational Approach for Analyzing and Detecting Emotions in Arabic Text. *International Journal of Engineering Research and Applications (IJERA), 3*(3), 100-107.

Greevy, E., and Smeaton, A. F. (2004). *Classifying racist texts using a support vector machine*. Proceedings of the The 27th annual international ACM SIGIR conference on Research and development in information retrieval. July 25-29, 2004. Sheffield, United Kingdom, 468-469.

Guang-Bin, H., Qin-Yu, Z., and Chee-Kheong, S. (2004). *Extreme learning machine: a new learning scheme of feedforward neural networks*. Proceedings of the 2004 IEEE International Joint Conference on Neural Networks. July 25-29, 2004. Budapest, Hungary, 985-990 vol.982.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Harish, B. S., Guru, D. S., Manjunath, S., and Kiranagi, B. B. (2010). *A symbolic approach for text classification based on dissimilarity measure*. Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia. December 28 - 30, 2010. Allahabad, India, 104-108.

Huang, C., Fu, T., and Chen, H. (2010). Text-based video content classification for online video-sharing sites. *Journal of the American Society for Information Science and Technology, 61*(5), 891-906.

Huang, G. B., Zhu, Q. Y., and Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing, 70*(1-3), 489-501.

Hwang, M., Choi, C., and Kim, P. (2011). Automatic Enrichment of Semantic Relation Network and Its Application to Word Sense Disambiguation. *IEEE Transactions on Knowledge and Data Engineering, 23*(6), 845-858.

Iezzi, D. F. (2012). Centrality Measures for Text Clustering. *Communications in Statistics - Theory and Methods, 41*(16-17), 3179-3197.

Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. In C. Nédellec, and C. Rouveirol (Eds.). *Machine Learning: ECML-98* (Vol. 1398, pp. 137-142). Berlin, Germany: Springer Berlin Heidelberg.

Johnson, D. E., Oles, F. J., Zhang, T., and Goetz, T. (2002). A decision-tree-based symbolic rule induction system for text categorization. *IBM Systems Journal, 41*(3), 428-437.

Joho, H., and Sanderson, M. (2007). *Document frequency and term specificity*. Proceedings of the Large Scale Semantic Access to Content (Text, Image, Video, and Sound). May 30 - June 01, 2007. Pittsburgh, PA, USA, 350-359.

Jung, Y., Park, H., and Du, D. (2001). A Balanced term-weighting scheme for improved document comparison and classification. *preprint*.

Ki-moon, B. (2012). The Use Of The Internet For Terrorist Purposes. New York,: United Nations.

Koller, D., and Sahami, M. (1997). *Hierarchically Classifying Documents Using Very Few Words*. Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97). July 8-12, 1997. Nashville, TN, USA, 170-178.

L'Huillier, G., Alvarez, H., Aguilera, F., and Rios, S. A. (2010). *Topic-based Social Network Analysis for Virtual Communities of Interests in the Dark Web*. Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics(ISI-KDD 2010). July 25-28, 2010. Washington, DC, USA, 66-73.

Larkey, L., and Croft, W. B. (1996). *Combining classifiers in text categorization*. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. August 18-22, 1996. Zurich, Switzerland, 289-297.

Larkey, L. S., Ballesteros, L., and Connell, M. E. (2007). Light Stemming for Arabic Information Retrieval. In A. Soudi, A. d. Bosch, and G. Neumann (Eds.). *Arabic Computational Morphology* (Vol. 38, pp. 221-243). Netherlands: Springer

Last, M., Markov, A., and Kandel, A. (2006). Multi-lingual detection of terrorist content on the web. In H. Chen, F.-Y. Wang, C. C. Yang, D. Zeng, M. Chau, and K. Chang (Eds.). *Intelligence and Security Informatics* (pp. 16-30). Berlin, Germany: Springer Berlin Heidelberg.

Lee, L., Wan, C., Rajkumar, R., and Isa, D. (2012). An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization. *Applied Intelligence, 37*(1), 80-99.

Lee, Z.-S., Maarof, M. A., Selamat, A., and Shamsuddin, S. M. (2008). *Enhance Term Weighting Algorithm as Feature Selection Technique for Illicit Web Content Classification*. Proceedings of the Eighth International Conference on Intelligent Systems Design and Applications (ISDA'08). November 26-28, 2008 Kaohsiung, Malaysia, 145-150.

Leopold, E., and Kindermann, J. (2002). Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Machine Learning, 46*(1-3), 423-444.

Lewis, D. D., and Gale, W. A. (1994). *A sequential algorithm for training text classifiers*. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. July 03-06, 1994. Dublin, Ireland, 3-12.

Liu, Y., Loh, H., and Tor, S. (2005). Comparison of Extreme Learning Machine with Support Vector Machine for Text Classification. In M. Ali, and F. Esposito (Eds.). *Innovations in Applied Artificial Intelligence* (Vol. 3533, pp. 390-399). Berlin, Germany: Springer Berlin Heidelberg.

Liu, Y., Loh, H. T., and Sun, A. X. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems with Applications, 36*(1), 690-701.

Author's copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Luying, L., Jianchu, K., Jing, Y., and Zhongliang, W. (2005). *A comparative study on unsupervised feature selection methods for text clustering*. Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on. 30 Oct.-1 Nov. 2005, 597-601.

Man, L., Tan, C. L., Jian, S., and Yue, L. (2009). Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(4), 721-735.

McCallum, A., and Nigam, K. (1998). *A comparison of event models for Naive Bayes text classification*. Proceedings of the AAAI-98 Workshop on Learning for Text Categorization. July 26–27 and 31, 1998. Madison, Wisconsin, USA, 41-48.

Meng, J., Lin, H., and Yu, Y. (2011). A two-stage feature selection method for text categorization. *Computers & Mathematics with Applications, 62*(7), 2793-2800.

Mengle, S. S. R., and Goharian, N. (2009). Ambiguity measure feature-selection algorithm. *Journal of the American Society for Information Science and Technology, 60*(5), 1037-1050.

Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., and Waibel, A. (1990). Machine Learning. *Annual Review of Computer Science, 4*(1), 417-433.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., and Cunningham, P. (2013). *Uncovering the Wider Structure of Extreme Right Communities Spanning Popular Online Networks*. Proceedings of the 5th Annual ACM Web Science Conference. May 02- 04, 2013. Paris, France, 276-285.

Olatunji, S. O., Selamat, A., and Raheem, A. A. A. (2010). *Modeling Permeability Prediction Using Extreme Learning Machines*. Proceedings of the 2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation (AMS). May 26-28, 2010. Kota Kinabalu, Malaysia, 29-33.

Paik, J. H. (2013). *A novel TF-IDF weighting scheme for effective ranking*. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. July 28 - August 01 ,2013. Dublin, Ireland, 343-352.

Peng, H., Fulmi, L., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27*(8), 1226-1238.

Qin, J., Zhou, Y., and Chen, H. (2011). A multi-region empirical study on the internet presence of global extremist organizations. *Information Systems Frontiers, 13*(1), 75-88.

Qin, J., Zhou, Y., Reid, E., and Chen, H. (2008). Studying Global Extremist Organizations' Internet Presence Using the DarkWeb Attribute System. In H. Chen, E. Reid, J. Sinai, A. Silke, and B. Ganor (Eds.). *Terrorism Informatics* (pp. 237-266). USA: Springer

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning, 1*(1), 81-106.

Rafrafi, A., Guigue, V., and Gallinari, P. (2012). *Coping with the Document Frequency Bias in Sentiment Classification*. Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM). June 4–7, 2012. Dublin, Ireland, 314-314.

Ran, L., and Xianjiu, G. (2010). *An Improved Algorithm to Term Weighting in Text Classification*. Proceedings of the International Conference on Multimedia Technology (ICMT). October 29-31, 2010. Ningbo, China, 1-3.

Rennie, J., Shih, L., Teevan, J., and Karger, D. (2003). *Tackling the poor assumptions of Naive Bayes text classifiers*. Proceedings of the Twentieth International Conference on Machine Learning (ICML). August 21-24, 2003. Washington DC, USA, 616-623.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation, 60*(5), 503-520.

Saad, M. K., and Ashour, W. (2010). *OSAC: Open Source Arabic Corpora*. Proceedings of the 6th International Conference on Electrical and Computer Systems. November 25-26, 2010. Lefke, Cyprus, 118-123.

Sahu, B., and Mishra, D. (2012). A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data. *Procedia Engineering, 38*, 27-31.

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Sam, L. Z., Maarof, M. A., and Selamat, A. (2006). *Automated Web Pages Classification with Integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as Feature Reduction.*

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Proceedings of the International Conference on Man-Machine Systems (ICoMM06). September 15-16, 2006. Langkawi, Malaysia.

Sanderson, M., and Ruthven, I. (1996). *Report on the Glasgow IR group (glair4) submission*. Proceedings of the The Fifth Text Retrieval Conference (TREC-5). November 20-22, 1996. Gaithersburg, Maryland, 517-520.

Schapire, R., and Singer, Y. (2000). BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning, 39*(2-3), 135-168.

Selamat, A., and Omatu, S. (2004). Web page feature selection and classification using neural networks. *Information Sciences, 158*(1), 69-88.

Selamat, A., Subroto, I. M. I., and Ng, C.-C. (2009). Arabic script web page language identification using hybrid-KNN method. *International Journal of Computational Intelligence and Applications, 8*(3), 315-343.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications, 33*(1), 1-5.

Song, W., and Park, S. C. (2009). Genetic algorithm for text clustering based on latent semantic indexing. *Computers & Mathematics with Applications, 57*(11–12), 1901-1907.

Sun, D.-Y., Guo, S.-Q., Zhang, H., and Li, B.-X. (2011). *Study on covert networks of terroristic organizations based on text analysis*. Proceedings of the 2011 IEEE International Conference on Intelligence and Security Informatics (ISI 2011). July 10-12,2011. Beijin, China, 373-378.

Tianjun, F., Chun-Neng, H., and Hsinchun, C. (2009). *Identification of extremist videos in online video sharing sites*. Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics (ISI 2009). June 8-11, 2009. Dallas, TX , USA, 179-181.

Ting, S. L., See-To, E. K., and Tse, Y. K. (2013). Web Information Retrieval for Health Professionals. *Journal of Medical Systems, 37*(3), 1-14.

Tong, S., and Koller, D. (2002). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research, 2*(1), 45-66.

Tsai, C.-F., and Hsiao, Y.-C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems, 50*(1), 258-269.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24*(7), 1024-1032.

Unler, A., Murat, A., and Chinnam, R. B. (2011). mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Information Sciences, 181*(20), 4625-4641.

Vens, C., Struyf, J., Schietgat, L., Džeroski, S., and Blockeel, H. (2006). Decision trees for hierarchical multilabel classification: A case study in functional genomics. In J. Fürnkranz, T. Scheffe, and M. Spiliopoulou (Eds.). *Knowledge Discovery in Databases: Pkdd 2006* (Vol. 4213, pp. 18-29). Berlin, Germany: Springer Berlin Heidelberg.

Wadhwa, P., and Bhatia, M. (2013). *Tracking on-line radicalization using investigative data mining*. Proceedings of the National Conference on Communications (NCC). February 15-17, 2013. New Delhi, India 1-5.

Weiss, S. M., Apte, C., Damerau, F. J., Johnson, D. E., Oles, F. J., Goetz, T., and Hampp, T. (1999). Maximizing text-mining performance. *IEEE Intelligent Systems and their Applications, 14*(4), 63-69.

Wibowo, W., and Williams, H. E. (2002). *Simple and accurate feature selection for hierarchical categorisation*. Proceedings of the 2002 ACM symposium on Document engineering. November 8-9, 2002. McLean, Virginia, USA, 111-118.

Wu, H., Robert Wing Pong, L., Wong, K., and Kwok, K. (2008). Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems, 26*(3), 1-37.

Xianshan, Z., and Guangzhu, Y. (2012). *Finding criminal suspects by improving the accuracy of similarity measurement*. Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) May 29-31, 2012. Sichuan, China, 1145-1149.

Yang, C. C., Tang, X., and Gong, X. (2011). *Identifying dark web clusters with temporal coherence analysis*. Proceedings of the 2011 IEEE International Conference on Intelligence and Security Informatics (ISI 2011). July 10-12, 2011. Beijing, China, 167-172.

Yang, L., Liu, F., Kizza, J. M., and Ege, R. K. (2009). *Discovering Topics from Dark Websites*. Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Cyber Security (CICS). March 30 - April 2, 2009. Nashville, TN, USA, 175-179.

Authors' copy

In Press, Accepted Manuscript, Available online 3 October 2015
http://www.sciencedirect.com/science/article/pii/S092523121501396X

Yang, Y. (1995). *Noise reduction in a statistical approach to text categorization*. Paper presented at the Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, USA.

Yang, Y., and Liu, X. (1999). *A re-examination of text categorization methods*. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. August 15-19, 1999. Berkeley, California, USA, 42-49.

Yang, Y., and Pedersen, J. O. (1997). *A Comparative Study on Feature Selection in Text Categorization*. Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97). July 8-12, 1997. Nashville, TN, USA, 412-420.

Yanjun, L., Congnan, L., and Chung, S. M. (2008). Text Clustering with Feature Selection by Using Statistical Data. *Knowledge and Data Engineering, IEEE Transactions on, 20*(5), 641-652.

Zhang, P., Bui, T. D., and Suen, C. (2005). *Hybrid feature extraction and feature selection for improving recognition accuracy of handwritten numerals*. Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05). August 29 - September 1, 2005. Seoul, Korea, 136-140 Vol. 131.

Zhang, Y., Ding, C., and Li, T. (2008). Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics, 9*(Suppl 2), S27.

Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology, 57*(3), 378-393.

Zhou, Y., Qin, J., Lai, G., Reid, E., and Chen, H. (2006). *Exploring the dark side of the web: collection and analysis of u.s. extremist online forums*. Proceedings of the 2006 IEEE international conference on Intelligence and Security Informatics (ISI 2006). May 23-24, 2006 San Diego, CA, USA, 621-626.

Zimbra, D., and Chen, H. (2012). *Scalable sentiment classification across multiple dark web forums*. Proceedings of the 2012 IEEE International Conference on Intelligence and Security Informatics (ISI 2012). June 11-14, 2012. Washington, DC, USA, 78-83.

Authors' copy

View publication stats