International Journal of Innovative Computing, Information and Control
Volume **13**, Number **4**, August **2017**

pp. 1385–1398

UTILIZING STANDARD DEVIATION IN TEXT CLASSIFICATION WEIGHTING SCHEMES

Fawaz Shukhier Al-Anzi, Dia AbuZeina and Shatha Hasan
Department of Computer Engineering
Kuwait University
P.O. Box 5969 Safat, 13060, Kuwait
{ fawaz.alanzi; abuzeina }@ku.edu.kw; shatha.hasan@cs.ku.edu.kw

Received February 2017; revised May 2017

Abstract. *The term frequency – inverse document frequency (TF-IDF) weighting scheme is widely used in text classification for weighting the features of the vector space model (VSM). It aims at enhancing words' discriminating capabilities by weighing up the less frequently used words and, at the same time, weighing down the high frequency words (i.e., the common words such as prepositions). This paper attempts to provide an enhanced variant of the well-known TF-IDF method. The TF-IDF is a statistical estimation that computes the weight of each word based on the frequency of the word in both the document and the entire data collection. In this work, we propose considering the word's standard deviation as another factor when computing the word's weight. That is, the common words tend to have larger standard deviations more than the uncommon words. In other words, the more the word appears in documents, the greater the standard deviation is. To investigate the proposed TF-IDF based model, we conducted some experiments for Arabic text classification. We used a training textual data collection that contains 1,750 documents of five categories (250 documents for testing). The experimental results show that the proposed approach is superior to the standard TF-IDF term weighting scheme.*
Keywords: Arabic, Text, Classification, TF-IDF, Singular value decomposition

1. **Introduction.** The significant growth of online textual information has increased the demand for effective content-based text retrieval methods. In fact, information retrieval (IR) systems are characterized by extensive query activities that require speed, accuracy, and rich information. Hence, there has been great interest to promote intelligent algorithms for text mining and document classification. Moving from paper-based to digital solutions such as information systems is dominating the institutional processes since it proficiently automates today's business requirements such as uniqueness, security, productivity, and consistency. In particular, commercial companies such as newspapers highly consider efficient archiving systems for many tasks such as backup, speedy retrieval of information, as well as avoiding human-caused errors. Today, it becomes clear that the huge amount of textual data and the vast flow of information require powerful methods for data analysis, classification, and categorization. The literature review demonstrates some of the applications that employ text classification and the related algorithms such as the well-known weighting scheme term frequency – inverse document frequency (TF-IDF).

Recently, there has been quite a significant research to promote intelligent information retrieval (IR) algorithms for highly gratified results in text mining applications. However, almost all IR algorithms employ the most famous weighting scheme TF-IDF [1,2] as a method to find the degree of importance (i.e., the weight) of each word in the data collection. Utilizing the weights of words is a great idea that plays an important role to enhance the performance in the IR systems. For instance, the study in [3] pointed out that the weighting schemes dominate the performance in text classification task. The main objective of TF-IDF is to alleviate the negative effect of the common words in the classification process. It is unavoidable that documents generally contain some common words such as the propositions (e.g., in, at, and or). The TF-IDF has two parts (TF and IDF) that are when multiplied together produce the weight of each word. TF is the frequency of a particular word in a particular document (i.e., the significance of the word in this document) while the IDF is the logarithm of the total number of documents divided by the total number of documents the word appears (i.e., how infrequent a word is in the corpus). Despite that TF-IDF has been successfully implemented in many IR applications, there is some capacity for further enhancement. For instance, the study in [4] indicated that the most commonly used IT-IDF method is unsupervised and, accordingly, they claimed it is not the best choice for IR.

In this paper, we consider employing the word's standard deviation as a new factor in the TF-IDF model. In fact, the research toward enhanced TF-IDF variants is not new. For instance, the study in [5] has one variant. The intuition behind our work is that the standard deviation is a suitable candidate to capture the words that have large scatter among the data collection. That is, we propose using standard deviation as another factor to include the effect of the word's spreading (i.e., the word's dispersion). From classification point of view, the more the word appears in the documents, the less discriminative power it has. In the same meaning, it is indicated in [6] that the fewer a term appears in categories, the more discriminative the term is for text categorization. For evaluating the proposed method, we implemented the new weighting scheme for a classification task using an Arabic data collection. In this work, we used the vector space model (VSM) [7]. VSM is a semantic loss method that ignores the semantic relationships between words. Therefore, we used the latent semantic indexing (LSI) [8] method that focuses on the semantic meaning of words through the singular value decomposition (SVD) [9] method. Hence, SVD is a mathematical matrix operation that is used to uncover the words relationships as well as to reduce the number of dimensions of the document vectors. The cosine similarity measure was used for classification using a data collection that contains 1,500 documents for training and 250 documents for testing. The data collection has five categories. Despite that the proposed work was evaluated against an Arabic data collection, nevertheless, the extended TF-IDF weighting scheme might have benefits to other languages.

In the next section, we present the literature review. In Section 3, we present some of the Arabic text challenges followed by the proposed method in Section 4. In Section 5, we present the experimental results and the conclusion is shown in Section 6.

2. Literature Review.

The TF-IDF has widely been studied in text mining and IR research. In fact, the TF-IDF is an extremely important weighing scheme which many of the IR studies utilize for better performance. In this section, we demonstrate some of the TF-IDF based weighting schemes. However, we emphasize that none of the published studies consider the standard deviation concept. The study in [10] indicates that the TF-IDF does not leverage the information implicitly contained in the categorization task to represent documents. Hence, it introduces a new weighting method based on statistical estimation of the importance of a word for a specific categorization problem. A term weighing measure was proposed in [11], and the measure is based on an information-theoretic view of retrieval events. It is expressed as a product of the occurrence probabilities of terms and their amounts of information. It is indicated in [12] that the conventional TF-IDF might cause high false alarm rate in anomaly detection. Hence, it presents a model that considers the special information between different processes and sessions of computer audit data. An improved TF-IDF approach was proposed in [13] based on confidence, support and characteristic words. Synonyms defined by a lexicon are processed in the improved TF-IDF approach. The research in [14] demonstrates that the shortcoming of traditional weighting schemes is the limitations in extracting semantically exact indexes that represent the semantic content of a document. Hence, it presents an enhanced TF-IDF model in an indexing formalism that considers not only the terms in a document, but also the concepts. To improve the term's discriminating power, [15] proposed a term weighting scheme called term frequency – relevance frequency (tf.rf) that considers the relevant document distribution.

The literature also has many studies that consider adaptation of TF-IDF. For instance, the study in [16] proposed an ontology-based scheme for the semiautomatic annotation of documents and a retrieval system. Weights were computed automatically by an adaptation of the TF-IDF algorithm, based on the frequency of occurrence of the instances in each document. In [17], the researchers indicate that the limitation of the TF-IDF approach is high dimensionality of data and it does not consider the relations among the terms. Therefore, they proposed an approach that is called weighted concept frequency-inverse document frequency (CF-IDF) with background knowledge of domain ontology. The work in [18] presents a weighting scheme that is based on finding the average word occurrences of words in documents. They also use the document centroid vector to remove less significant weights from the documents. A term weighting scheme is proposed in [19] to exploit the semantics of categories and indexing terms. They indicate that the proposed method is to overcome the limitation of the TF-IDF that only exploits statistical information of terms in documents. [20] indicates that the IDF in the TF-IDF is oblivious to the training class labels and naturally scales some features inappropriately. Hence, they replace IDF with bi-normal separation (BNS) with excellent at ranking words for feature selection filtering. The study in [21] demonstrates a supervised framework for extracting keywords from meeting transcripts. They indicate that a feedback loop mechanism is able to outperform both TF-IDF weighting and a keyphrase extraction system known for its satisfying performance on written text.

Recently, there are some studies that utilize enhanced TF-IDF variants in different applications. The study in [22] introduces a recommender system that employs TF-IDuF as a term-weighting scheme that does not require access to the general document corpus and that considers information from the users' personal document collections. [23] provides a probabilistic explanation for the TF-IDF heuristic. It also shows that the ideas behind explanation can help us come up with more TF-IDF variants. The study in [24] generalizes the IDF in the standard TF-IDF by proposing to use joint IDF for a set of terms together, compared with using each term's IDF individually. The above-mentioned studies show that the standard deviation has not yet been implemented in IR weighting schemes that is the motivation of this work. Regarding linguistic applications that utilize the standard TF-IDF, Table 1 shows some of the noted applications.

3. **Text Classification Challenges.**

Text classification is a challenging task especially when the text has many words that are related to different categories or topics. For example, Figure 1 contains a document that belongs to the sport category based on our knowledge of the training data collection in this work. However, it has few words related to the sport category. Instead, it has some words that are related to the economy category such as {رجال الاعمال ➔ businesspersons, جهات المال والاعمال ➔ business & finance centers}. In addition, there are some words that are related to the education category such as {باحثون ➔ researchers, جامعة ➔ a university}. Of course, the text classification task has other challenges such as stop words that have little discriminative power. Examples

Table 1. Utilizing TF-IDF in different IR applications

Reference	Domain
[25]	Classification of email into a user-defined hierarchy
[26]	Document classification
[27]	Discovering frequent sequential patterns and phrases
[28]	Multi-document summarizer
[29]	Intrusion detection
[30]	Measuring the semantic similarity
[31]	Content-based recommendation
[32]	Indexing information framework based on semantics
[33]	Summarization
[34]	Predicting social tags from MP3 for music recommendation
[35]	Automatic mood classification for music systems
[36]	Extractive summarization of microblog posts
[37]	Detecting phishing web sites
[38]	Content recommendation system based on user profile
[39]	Classifying duplicate bug reports

An Arabic Document	The translation using Google translator
الاستيقاظ مبكراً شرط للنجاح ديلي ميل – تظهر الأعمال الدرامية رجل الاعمال الناجح في صورة الشخص الذي يسهر طويلاً في المكتب، لكن انجح رجال الاعمال يكذبون هذه الصورة، مؤكدين ان اول اسرار النجاح في الحياة العملية هو الاستيقاظ المبكر. وتنتشر في دوائر المال والاعمال الاميركية عبارة ان «من يستيقظ بعد الخامسة لا يمكن ان ينجح». وهناك قائمة من الانشطة التي يجب على رجل الاعمال الناجح القيام بها قبل الذهاب الى عمله، وهي: قراءة الرسائل الالكترونية والصحف وممارسة الرياضة وتناول الافطار مع العائلة. وخلصت دراسة اجراها باحثون ألمان الى ان «النجاح المهني من نصيب من يستيقظون مبكراً»، كما يقول مشرف الدراسة كريستوف راندلر من جامعة هايدلبرغ الذي يوضح: «يتميز من يستيقظون مبكراً بدرجة أكبر من النشاط وقدرة على الانجاب والاحساس بالمسؤولية.	Waking up early prerequisite for success Daily Mail dramas show a successful entrepreneur in the form of a person who sees a long in the office, but the most successful business people are lying to this image, stressing that the first secrets of success in the process of life is waking up early. And spread in circles and American money is a business that «wake up after the fifth can not succeed.» There is a list of activities that a successful entrepreneur must be done before going to work, namely: read e-mails, newspapers and exercise and eat breakfast with the family. A study conducted by German researchers that «the professional success of a share of the early wake up», says Musharraf Christophe Randler study from the University of Heidelberg, who explains: «characterized by waking up early, the largest of activity and the ability to have children and sense of responsibility degree.

Figure 1. An Arabic article with its translation using Google translator

include the prepositions {من ➔ from, الى ➔ to, في ➔ in, على ➔ on}. Intuitively, since the discriminative words play a vital role in the classification process, the variety of document's contents might increase the misclassification rate. The presented document, in Figure 1, raises the importance of choosing the best discriminative words especially when the corpus documents have such contents diversity.

For further illustration of the challenges in text classification, we consider the following case that demonstrates an article of different categories. The title and a part of the body of the article is shown in Figure 2. The article is related to an interview with the ambassador of Senegal in Kuwait. No doubt, these kinds of articles generally contain different topics of different categories such as politics, economy, and education. Figure 2 also shows some of keywords that appear in the article. The keywords clearly belong to different categories and, therefore, it will be hard to categorize the document in question.

An Arabic article	The translation using Google translator
إمباكي أكد أن السنغال من أكبر الدول الديمقراطية في أفريقيا السفير السنغالي لـ «الأنباء»: تعاون أمني استخباراتي وتنسيق سياسي مع الكويت - إمباكي أكد أن السنغال من أكبر الدول الديمقراطية في أفريقيا. - السفير السنغالي لـ «الأنباء»: تعاون أمني استخباراتي وتنسيق سياسي مع الكويت. - نأمل توقيع اتفاقية الإعفاء من التأشيرات لحاملي الجوازات الدبلوماسية قبل انعقاد الدورة المشتركة في 2017. - تواجد المستثمر الكويتي بالسنغال دون الطموح والأجانب يستطيعون إنشاء شركة خلال 48 ساعة. - استثمارات بقيمة 65 مليون دولار لمجموعة الخرافي في «دكار». - قضية الصحراء الغربية تحلّ عن طريق الأمم المتحدة دون تدخلات خارجية. - فرنسا الأولى في إعادة الأمن إلى مالي ونحن لعبنا دوراً مهماً في هذا الإطار. ...	**Mbeki stressed that Senegal is one of the largest democratic countries in Africa, the Senegalese Ambassador to Al-Anbaa: intelligence cooperation and political coordination with Kuwait** - Mbaki stressed that Senegal is one of the largest democratic countries in Africa. - Senegalese ambassador to Al-Anbaa: intelligence cooperation and political coordination with Kuwait. -We hope to sign the visa exemption agreement for diplomatic passport holders before the 2017 joint session. The presence of Kuwaiti investor in Senegal without ambition and foreigners can establish a company within 48 hours. -Investments of $ 65 million for Al-Kharafi group in Dakar. -The issue of Western Sahara is resolved through the United Nations without external interference. - France was the first to restore security to Mali and we played an important role in this regard. ...
Some of the keywords of the article	
السنغال، الديموقراطية، أفريقيا، السفير، أمني، استخباراتي، سياسي، الكويت، اتفاقية، التأشيرات، الجوازات، الدبلوماسية، المستثمر، شركة، استثمارات، مليون، دولار، الصحراء، الأمم المتحدة، فرنسا، الأمن، السلك الديبلوماسي، سفير، الأمير، العالم العربي، الشرق الأوسط، تبادل المعلومات، التنمية الاقتصادية، المشاريع، الضمانات، الحكومة، المباحثات، خارجية، المملكة العربية السعودية، الوضع السوري، المناطق المشتعلة، العراق، الخليجية، مجلس الوزراء، وزير، الخارجية، الصداقة، النقل الجوي، حركات إرهابية، السياسي، القضية الفلسطينية، الجمعيات الخيرية، باخرة محملة الأسلحة، نيجيريا، التجار، الأحزاب الدينية، البرلمان،	Senegal, Democracy, Africa, Ambassador, Security, Intelligence, Politician, Kuwait, agreement, Visas, Passports, Diplomacy, Investor, a company, Investments, Million, Dollars, the desert, United nations, France, Security, Diplomatic corps, Ambassador, the prince, Arab world, Middle east, Exchange of information, economical development, Projects, Guarantees, the government, Discussions, External, Kingdom of Saudi Arabia, The Syrian situation, Flaming areas, Iraq, Gulf Cooperation Council, Council of Ministers, minister, Foreign Affairs, the friendship, Air transport, Terrorist movements, Political, The Palestinian cause, Charities, Steamer loaded with weapons, Nigeria, Merchants, Religious parties, Parliament,

FIGURE 2. An Arabic article with some different keywords

From linguistic point of view, Arabic is a rich language that requires effective text classification algorithms in order to handle different aspects of the language such as morphology, vocabulary, and syntax. [40] highlights some of the challenges of the Arabic language. In addition, [41] implemented the conventional TF-IDF for Arabic text using different machine learning classifiers.

4. **The Proposed Method.**

The proposed method is based on the standard TF-IDF weighting scheme that is one of the popular term weighting methods in various text domains. TF-IDF is very effective for selecting important words that assigns large weights to the high frequency terms in individual documents, but is at the same time relatively rare in the entire corpus. The classical formula of standard TF-IDF is shown in the following Formula (1):

$$w_{i,j} = tf_{i,j} \log\left(\frac{N}{df_i}\right) \tag{1}$$

where $w_{i,j}$ is the weight for word i in document j, $tf_{i,j}$ is the frequency of word i in document j, N is the number of documents in the collection, and df_i is the number of documents that contains the word i. The proposed method suggests to combine the standard deviation (STD) of each word along with the standard TF-IDF weighting scheme. [42] defines the standard deviation as shown in the following Formula (2):

$$\text{standard deviation } (STD) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(xi-\mu)^2} \tag{2}$$

where N is the total number of observation, xi is an observation, and μ is the mean. Hence, standard deviation is the positive square root of the variance that quantifies the spread of observations. [43] indicated that standard deviation is a dispersion measure that summarizes the extent of spread of observations in a sample. Accordingly, it is expected that the high standard deviation produces a high word's dispersion. Therefore, we used this measure to penalize the words that have high standard deviation, and reinforce the words with low standard deviation. As above indicated, we used LSI method for feature extraction. Hence, the standard deviation will be considered when creating the LSI term-by-document matrix. For illustration, Table 2 shows an example of a typical term-by-document matrix that is partially filled by some random numbers (i.e., the row of the *word1* and the row of the *word4*).

Table 2. A typical partially filled term-by-document matrix

	doc1	doc2	doc3	doc4	doc5	doc6	doc7
word1	0	1	0	0	0	1	0
word2							
word3							
word4	2	2	0	1	1	0	1
word5							
word6							

Word1 appears in doc2 and doc6 (i.e., at position 2 and position 6 of the word's row). In this work, we used Python, hence, the index starts at 0 when creating the term-by-document matrix, and the position of the word1 will be 1 and 5. The standard deviation of the two numbers 1 and 5 is 2. Similarly, word4 appears in the position (1, 2, 4, 5, and 7) that becomes (0, 1, 3, 4, and 6). The standard deviation of these numbers is 2.13. Accordingly, as the word appears in more document, its standard deviation gets larger. This is the characteristic of the common words that are generally distributed in a large number of documents. For control effect of this measure (i.e., the STD), we propose using normalizing to scale the values between 0 and 1 before using it. To have the STD included in the standard TF-IDF weighting scheme, we propose the following new weight Formula (3):

$$\text{Weight}\,(w) = TF\text{-}IDF * (1 - \text{normalized}\,(STD)) \tag{3}$$

In this case, the standard deviation will increase the weight of the words that have low standard deviation; and at the same time, it will decrease the weights of words that have high standard deviation. To investigate the proposed method, we prepared an Arabic text corpus that contains 1,750 documents for training and 250 documents for testing. The training set contains 929,205 words with 80,156 unique words. The collected documents belong to five categories as shown in Table 3. The corpus was prepared with the help by Alqabas newspaper in Kuwait [44]. Table 3 shows the statistical information of the corpus.

Before using the corpus, we performed a preprocessing stage that includes deleting all out of the Arabic characters. It also includes deleting numbers, commas, full stops, and all other symbols. The normalization process includes changing some of Arabic characters such as (أ⟶ا) and (إ⟶ا). In the following algorithm, we present all necessary steps to implement the proposed method.

- Create the term-by-document matrix. It is the first step in LSI method. The training data set is used to create a matrix with rows to represent the words and the columns to represent the documents. However, three issues have to be considered before generating the matrix as the following.

– Document frequency (DF) is set. DF is a threshold that indicates how many different documents a selected word appears. This measure is set according to the experience of the training data set. For instance, if DF is set to 5, then only the words that appear in at least five different documents will be included in the term-by-document matrix.
– Ignore characters declared. Sometime it is important to discard some characters that do not help in the classification process. For instance, all English characters might be discarded since our system is for the Arabic text.
– Term-by-document matrix is weighed using the proposed weighting scheme.

- SVD is employed to generate the K rank feature vectors. The K rank gives the number of the dimensions of the generated document vectors. For instance, if K is chosen as 2, then the document vectors will be of two dimensions. Again, choosing K is related to the experience to find the optimal performance as well as to the size of the data.
- The cosine classifier is implemented. However, other distance measures can be used such as the Euclidean distance. We used the cosine measure according to its popularity to find the similarity between two vectors in n-dimensional space.
- Performance is evaluated. We used accuracy measure that is the total number of correctly classified documents (in a category) divided by the total number of example in the corresponding category. Since we measure the accuracy for different cases such as the DF and K rank, we used the average accuracy as an indication to the overall performance of the proposed method.

TABLE 3. The corpus information

The training data collection				
#	Category	Number of documents	Number of words	Number of unique words
1	Economy	350	225,659	31,942
2	Health	350	151,912	25,835
3	Education	350	224,078	35,294
4	Sports	350	135,473	24,056
5	Tourism	350	192,083	28,218
	Total	1,750	929,205	80,156*
The testing data collection				
1	Economy	50	22,031	6,959
2	Health	50	29,722	9,386
3	Education	50	35,338	8,961
4	Sports	50	15,168	5,482
5	Tourism	50	20,268	6,794
	Total	250	122,527	24,496*

* It is not algebraic summation since the common words are not counted.

To clarify the proposed method, we used a small corpus that contains five English quotes to build a small term-by-document matrix. The example shows how to employ the standard deviation along with the standard TF-IDF. The small corpus includes the following short sentences:

1) "window of opportunity will not open itself"

2) "in the middle of difficulty lies opportunity"
3) "luck is a matter of preparation meeting opportunity"
4) "imagination rules the world"
5) "the man who has no imagination has no wings"

In this example, the created term-by-document matrix ignores all words that are less than four characters such as {a, of, is, who, has}. Hence, the dictionary contains 16 words as follows: ['difficulty', 'imagination', 'itself', 'lies', 'luck', 'matter', 'meeting', 'middle', 'open', 'opportunity', 'preparation', 'rules', 'will', 'window', 'wings', 'world']. Table 4 shows the term-by-document matrix using terms counts for each dimension (i.e., for each feature). In the table, d is the shorthand for document.

Table 4. The dictionary and the term-by-document matrix

	Dictionary	Term-by-Document Matrix (no weight, only term count)				
#	Word (dimension)	$d1$	$d2$	$d3$	$d4$	$d5$
1	difficulty	0	1	0	0	0
2	imagination	0	0	0	1	1
3	itself	1	0	0	0	0
4	lies	0	1	0	0	0
5	luck	0	0	1	0	0
6	matter	0	0	1	0	0
7	meeting	0	0	1	0	0
8	middle	0	1	0	0	0
9	open	1	0	0	0	0
10	opportunity	1	1	1	0	0
11	preparation	0	0	1	0	0
12	rules	0	0	0	1	0
13	will	1	0	0	0	0
14	window	1	0	0	0	0
15	wings	0	0	0	0	1
16	world	0	0	0	1	0

The information provided in Table 4 is used to create the weighted term-by-document matrix using the standard TF-IDF scheme. The weighted words are shown in Table 5. For example, the weight of the word "imagination" in document number 4 (d4) is (1/3)*ln(5/2) = 0.333*0.916 = 0.31, according to Formula (1). The word of the weight zero means that this word does not appear in the corresponding document.

Table 6 shows the weight of each word using the proposed method that uses the standard deviation along with the standard TF-IDF (i.e., TF-IDF*(1−STD)). Intuitively, the standard deviation has been calculated for teach word at the first place. Therefore, we found the normalized standard deviation as follows { difficulty, 0; imagination, 0.61; itself, 0; lies, 0; luck, 0; matter, 0; meeting, 0; middle, 0; open, 0; opportunity, 1; preparation, 0; rules, 0; will, 0; window, 0; wings, 0; world, 0 }. For clarification, to find the standard deviation of the word "imagination", it was indicated in Table 4 that this word appears in document 4 and document 5. However, Python starts indexing at zero, so the word "imagination" appears in the position 3 and 4. The mean of these two values is 3.5. The standard deviation is the square root of $((3-3.5)$^2 $+ (4-3.5)^2)/2 = 0.5$. For all calculated standard deviations, the minimum value is 0 while the maximum value is 0.816, and the maximum value belongs to the word "opportunity". Hence, the normalized

Table 5. TF-IDF weight term-by-document matrix

	Dictionary	Term-by-Document Matrix (TF-IDF weight)				
#	Word (dimension)	*d*1	*d*2	*d*3	*d*4	*d*5
1	difficulty	0	0.40	0	0	0
2	**imagination**	**0**	**0**	**0**	**0.31**	**0.46**
3	itself	0.32	0	0	0	0
4	lies	0	0.40	0	0	0
5	luck	0	0	0.32	0	0
6	matter	0	0	0.32	0	0
7	meeting	0	0	0.32	0	0
8	middle	0	0.40	0	0	0
9	open	0.32	0	0	0	0
10	opportunity	0.10	0.13	0.10	0	0
11	preparation	0	0	0.32	0	0
12	rules	0	0	0	0.54	0
13	will	0.32	0	0	0	0
14	window	0.32	0	0	0	0
15	wings	0	0	0	0	0.80
16	world	0	0	0	0.54	0

standard deviation value of the word "imagination" is $(0.5-0)/0.816-0$ ➔ 0.61 based on the following general Formula (4):

$$\text{Normalized}\, xi = (xi - \min(X))/(\max(X) - \min(X)) \tag{4}$$

Similarly, the normalized standard deviation of the word "opportunity" is 1 since it is the most distributed word in this small corpus. According to the proposed method, TF-IDF*(1−STD), the weight of the "opportunity" is 0 as shown in Table 6. The weight of the word "imagination" in document 4 (i.e., d4) is $0.31 * (1 - 0.61) = 0.12$. Similarly, The weight of the word "imagination" in document 5 (i.e., d5) is $0.46 * (1 - 0.61) = 0.1794$ which is approximately 0.18 as shown in Table 6.

Likewise, the testing set feature vectors created using the same information in the dictionary. For instance, to create the feature vector of the word "imagination", it is simply a vector of 16 dimensions (according to the dictionary size). All dimensions are zero except the location of the word "imagination", the second position in this case, as follows: [0, 0.26, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0].

5. **The Experimental Results.** A number of experiments were conducted to evaluate the proposed method against text classification task. A term-by-document matrix was generated for different experimental cases. The first experimental case had no words weights, and only the words counts were used in the matrix. The second case implemented the standard TF-IDF, and the third case utilized the proposed method that used TF-IDF with the standard deviation. We used different low rank approximates of the term-by-document matrix. Hence, by K rank approximation, we mean the number of eigenvalues used in the reduced feature vectors of the documents. Since we do not know, in advance, the optimal DF and K rank approximation, various DF and K rank values were used to evaluate the performance. We did not use stoplist, as the proposed method already panelizes the words that spread out through the corpus. In our experiments, we discarded all words that are less than four characters length. Table 7 presents the results without weighting, only words count.

Table 6. The term-by-document matrix using the proposed method

	Dictionary	Term-by-Document Matrix (TF-IDF*(1−STD) weight)				
#	Word (dimension)	*d*1	*d*2	*d*3	*d*4	*d*5
1	difficulty	0	0.40	0	0	0
2	**imagination**	**0**	**0**	**0**	**0.12**	**0.18**
3	itself	0.32	0	0	0	0
4	lies	0	0.40	0	0	0
5	luck	0	0	0.32	0	0
6	matter	0	0	0.32	0	0
7	meeting	0	0	0.32	0	0
8	middle	0	0.40	0	0	0
9	open	0.32	0	0	0	0
10	**opportunity**	**0**	**0**	**0**	**0**	**0**
11	preparation	0	0	0.32	0	0
12	rules	0	0	0	0.54	0
13	will	0.32	0	0	0	0
14	window	0.32	0	0	0	0
15	wings	0	0	0	0	0.80
16	world	0	0	0	0.54	0

Table 7. The results without TF-IDF weight

DF	*K* rank	Accuracy (%)
10	10	83.2
10	20	84.0
10	30	81.2
10	40	82.8
15	10	83.2
15	20	84.0
15	30	80.0
15	40	82.4
20	10	84.0
20	**20**	**84.4**
20	30	81.6
20	40	81.2
25	10	84.0
25	20	83.2
25	30	80.0
25	40	81.2
Average➔		82.5 (%)

Table 7 shows that the maximum accuracy is 84.4% that is found at DF = 20 and $K = 20$. The average of the accuracies for randomly selected DF and the K rank values found to be 82.5%. No doubt, the performance is better when using the standard TF-IDF as shown in Table 8. The maximum score was 90.4% at DF = 10 and $K = 40$. The average of the accuracies is 89.5%.

Table 8. The results using the standard TF-IDF weight

DF	K rank	Accuracy (%)
10	10	89.6
10	20	89.2
10	30	90.0
10	**40**	**90.4**
15	10	89.2
15	20	87.6
15	30	90.0
15	40	90.4
20	10	89.6
20	20	87.6
20	30	89.6
20	40	90.0
25	10	90.0
25	20	89.6
25	30	89.6
25	40	89.6
Average➔		89.5 (%)

Table 9. The results of the proposed method (TF-IDF and STD)

DF	K rank	Accuracy (%)
10	10	89.6
10	**20**	**94.0**
10	30	92.8
10	40	92.4
15	10	92.0
15	20	92.4
15	30	91.6
15	40	89.2
20	10	91.6
20	20	93.6
20	30	91.2
20	40	90.0
25	10	92.0
25	20	93.6
25	30	92.0
25	40	90.8
Average➔		91.8 (%)

The proposed method results are shown in Table 9. The table shows that the maximum accuracy is found to be 94.0% at DF = 10 and $K = 20$. The average of accuracies was 91.8%. Hence, the proposed method outperforms the standard TF-IDF by 2.3%.

The information provided in the previous tables is demonstrated in the chart shown in Figure 3. The chart shows that the proposed method is performing better than the standard TF-IDF. However, few points show close accuracy, but on the average, the proposed method has better performing than the standard TF-IDF.

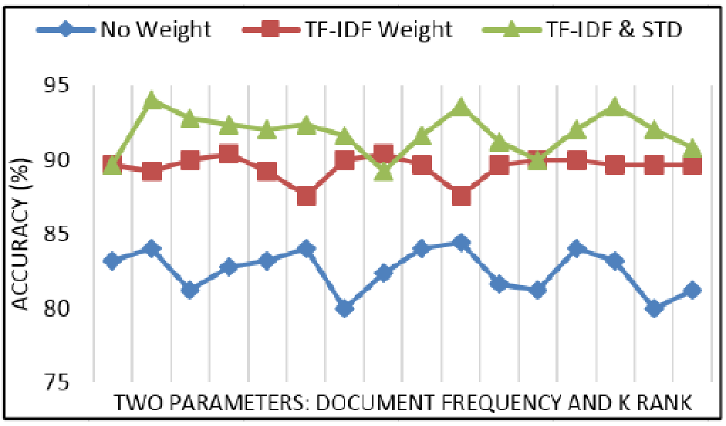

FIGURE 3. The performance enhancement of the proposed method

$$\mathcal{E}_l = \frac{N}{N+z^2}\left(\hat{\mathcal{E}} + \frac{z^2}{2N} - z\sqrt{\frac{\hat{\mathcal{E}}(1-\hat{\mathcal{E}})}{N} + \frac{z^2}{4N^2}}\right)$$

$$\mathcal{E}_u = \frac{N}{N+z^2}\left(\hat{\mathcal{E}} + \frac{z^2}{2N} + z\sqrt{\frac{\hat{\mathcal{E}}(1-\hat{\mathcal{E}})}{N} + \frac{z^2}{4N^2}}\right)$$

FIGURE 4. Confidence interval calculation formulas

Finally, we investigated whether the proposed method significantly outperforms the standard TF-IDF method. We used the tests of significance method that was proposed by Plötz in [45] to detect the significance of the obtained enhancement. We used 95% as a level of confidence. We also used the average error rate of the proposed method to be 8.2% (100 – 91.8, see Table 9). The first step is to compute the confidence interval $[l, u]$ based on the formulas that are shown in Figure 4, where $\hat{\varepsilon}$ is the average error rate and N is the total number of documents in the testing set (i.e., 250 documents). Since we used 95% as a level of confidence, z is equal to 1.96 from the standard normal distribution. Level of confidence might be interpreted as a 95% probability that a standard normal variable, z, will fall between -1.96 and 1.96.

The confidence interval is computed and found to be $[9.15\%, 12.02\%]$. The average error rate of the proposed method is 8.2%. Since 8.2% is outside the confidence interval, we conclude that the proposed method significantly outperforms the standard TF-IDF.

6. **Conclusion.** The TF-IDF weighting scheme is a popular method that is widely used to enhance the performance in text classification systems. This paper presents a new improved TF-IDF variant that is based on the word's standard deviation. The experimental results show that the proposed method significantly outperforms the conventional TF-IDF. The paper's contribution boosts the research to find more variants of the TF-IDF. As a future work, we propose to utilize the proposed method for extracting the stop words list. We also propose to investigate the modified weighting scheme for other text mining applications such as document clustering such as in [46].

Acknowledgment. This work is supported by Kuwait Foundation of Advancement of Science (KFAS), Research Grant Number P11418EO01. The authors would like also to thank Kuwait University – Research Administration for its support of this research work.

REFERENCES

[1] G. Salton and C. Buckley, Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, vol.24, no.5, pp.513-523, 1988.

[2] K. S. Jones, A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, vol.28, no.1, pp.11-21, 1972.

[3] E. Leopold and J. Kindermann, Text categorization with support vector machines – How to represent texts in input space, *Machine Learning*, vol.46, pp.423-444, 2002.

[4] P. Soucy and G. W. Mineau, Beyond TFIDF weighting for text categorization in the vector space model, *IJCAI*, vol.5, pp.1130-1135, 2005.

[5] Y. Yang and X. Liu, A re-examination of text categorization methods, *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[6] D. Wang and H. Zhang, Inverse-category-frequency based supervised term weighting schemes for text categorization, *Journal of Information Science and Engineering*, vol.29, no.2, pp.209-225, 2013.

[7] G. Salton, A. Wong and C.-S. Yang, A vector space model for automatic indexing, *Communications of the ACM*, vol.18, no.11, pp.613-620, 1975.

[8] S. Deerwester et al., Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, vol.41, no.6, p.391, 1990.

[9] M. W. Berry, S. T. Dumais and G. W. O'Brien, Using linear algebra for intelligent information retrieval, *SIAM Review*, vol.37, no.4, pp.573-595, 1995.

[10] P. Soucy and G. W. Mineau, Beyond TFIDF weighting for text categorization in the vector space model, *IJCAI*, vol.5, pp.1130-1135, 2005.

[11] A. Aizawa, An information-theoretic perspective of TF-IDF measures, *Information Processing & Management*, vol.39, no.1, pp.45-65, 2003.

[12] Z. Zhang and H. Shen, Application of online-training SVMs for real-time intrusion detection with different considerations, *Computer Communications*, vol.28, no.12, pp.1428-1442, 2005.

[13] Y. T. Zhang, L. Gong and Y. C. Wang, An improved TF-IDF approach for text classification, *Journal of Zhejiang University Science A*, vol.6, no.1, pp.49-55, 2005.

[14] B. Y. Kang and S. J. Lee, Document indexing: A concept-based approach to term weight estimation, *Information Processing & Management*, vol.41, no.5, pp.1065-1080, 2005.

[15] M. Lan, S. Y. Sung, H. B. Low and C. L. Tan, A comparative study on term weighting schemes for text categorization, *Proc. of IEEE International Joint Conference on Neural Networks*, vol.1, pp.546-551, 2005.

[16] P. Castells, M. Fernandez and D. Vallet, An adaptation of the vector-space model for ontology-based information retrieval, *IEEE Trans. Knowledge and Data Engineering*, vol.19, no.2, pp.261-272, 2007.

[17] S. Agarwal, A. Singhal and P. Bedi, Classification of RSS feed news items using ontology, *The 12th International Conference on Intelligent Systems Design and Applications (ISDA)*, 2012.

[18] O. Ibrahim and D. Landa-Silva, A new weighting scheme and discriminative approach for information retrieval in static and dynamic document collections, *The 14th UK Workshop on Computational Intelligence (UKCI)*, pp.1-8, 2014.

[19] Q. Luo, E. Chen and H. Xiong, A semantic term weighting scheme for text categorization, *Expert Systems with Applications*, vol.38, no.10, pp.12708-12716, 2011.

[20] G. Forman, BNS feature scaling: An improved representation over TF-IDF for SVM text classification, *Proc. of the 17th ACM Conference on Information and Knowledge Management*, pp.263-270, 2008.

[21] F. Liu, F. Liu and Y. Liu, A supervised framework for keyword extraction from meeting transcripts, *IEEE Trans. Audio, Speech, and Language Processing*, vol.19, no.3, pp.538-548, 2011.

[22] J. Beel, S. Langer and B. Gipp, TF-IDuF: A novel term-weighting scheme for user modeling based on users' personal document collections, *Proc. of the 12th iConference*, 2017.

[23] L. Havrlant and V. Kreinovich, A simple probabilistic explanation of term frequency-inverse document frequency (TF-IDF) heuristic (and variations motivated by this explanation), *International Journal of General Systems*, vol.46, no.1, pp.27-36, 2017.

[24] P. Viswanath, J. Rohini and Y. C. A. P. Reddy, Query performance prediction using joint inverse document frequency of multiple terms, *Emerging Trends in Electrical, Communications and Information Technologies: Proceedings of ICECIT-2015*, Springer, Singapore, 2017.

[25] J. D. Brutlag and C. Meek, Challenges of the email domain for text classification, *ICML*, pp.103-110, 2000.

[26] C. H. Caldas and L. Soibelman, Automating hierarchical document classification for construction management information systems, *Automation in Construction*, vol.12, no.4, pp.395-406, 2003.
[27] S.-T. Wu, Y. Li, Y. Xu, B. Pham and P. Chen, Automatic pattern-taxonomy extraction for web mining, *Proc. of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pp.242-248, 2004.
[28] D. R. Radev, H. Jing, M. Styś and D. Tam, Centroid-based summarization of multiple documents, *Information Processing & Management*, vol.40, no.6, pp.919-938, 2004.
[29] W. H. Chen, S. H. Hsu and H. P. Shen, Application of SVM and ANN for intrusion detection, *Computers & Operations Research*, vol.32, no.10, pp.2617-2634, 2005.
[30] R. Mihalcea, C. Corley and C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity, *AAAI*, vol.6, pp.775-780, 2006.
[31] M. J. Pazzani and D. Billsus, *Content-Based Recommendation Systems*, The Adaptive Web, Springer Berlin Heidelberg, 2007.
[32] C. R. Valêncio, R. D. Martins, M. H. Marioto, P. L. P. Correa and M. Babini, Automatic knowledge extraction supported by semantic enrichment in medical records, *International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, pp.79-83, 2013.
[33] V. Gupta and G. S. Lehal, A survey of text summarization extractive techniques, *Journal of Emerging Technologies in Web Intelligence*, vol.2, no.3, pp.258-268, 2010.
[34] D. Eck, P. Lamere, T. Bertin-Mahieux and S. Green, Automatic generation of social tags for music recommendation, *Advances in Neural Information Processing Systems*, pp.385-392, 2008.
[35] M. V. Zaanen and P. Kanters, Automatic mood classification using TF*IDF based on lyrics, *ISMIR*, pp.75-80, 2010.
[36] D. Inouye and J. K. Kalita, Comparing twitter summarization algorithms for multiple post summaries, *IEEE the 3rd International Conference on Privacy, Security, Risk and Trust and IEEE the 3rd International Conference on Social Computing*, pp.298-306, 2011.
[37] Y. Zhang, J. I. Hong and L. F. Cranor, Cantina: A content-based approach to detecting phishing web sites, *Proc. of the 16th International Conference on World Wide Web*, pp.639-648, 2007.
[38] T. Chen, W. L. Han, H. D. Wang, Y. X. Zhou, B. Xu and B. Y. Zang, Content recommendation system based on private dynamic user profile, *International Conference on Machine Learning and Cybernetics*, vol.4, pp.2112-2118, 2007.
[39] N. Jalbert and W. Weimer, Automated duplicate detection for bug tracking systems, *IEEE International Conference on Dependable Systems and Networks with FTCS and DCC*, pp.52-61, 2008.
[40] F. S. Al-Anzi and D. AbuZeina, Stemming impact on Arabic text categorization performance: A survey, *The 5th International Conference on Information & Communication Technology and Accessibility (ICTA)*, 2015.
[41] F. S. Al-Anzi and D. AbuZeina, Toward an enhanced Arabic text classification using cosine similarity and latent semantic indexing, *Journal of King Saud University – Computer and Information Sciences*, 2016.
[42] *http://mathworld.wolfram.com/StandardDeviation.htm*.
[43] R. Chattamvelli, *Data Mining Algorithms*, Alpha Science International, 2011.
[44] *http://www.alqabas.com.kw/Default.aspx*.
[45] T. Plötz, *Advanced Stochastic Protein Sequence Analysis*, Ph.D. Thesis, Bielefeld University, 2005.
[46] A. K. Murugesan and B. J. Zhang, A new term weighting scheme for document clustering, *The 7th Int. Conf. Data Min. (DMIN 2011-WORLDCOMP 2011)*, Las Vegas, NV, USA, 2011.