A survey of term weighting schemes for text classification

Abdullah Alsaeedi

Department of Computer Science,
College of Computer Science and Engineering (CCSE),
Taibah University,
Medina, Saudi Arabia
Email: aasaeedi@taibahu.edu.sa

Abstract: Text document classification approaches are designed to categorise documents into predefined classes. These approaches have two main components: document representation models and term-weighting methods. The high dimensionality of feature space has always been a major problem in text classification methods. To resolve high dimensionality issues and to improve the accuracy of text classification, various feature selection approaches were presented in the literature. Besides which, several term-weighting schemes were introduced that can be utilised for feature selection methods. This work surveys and investigates various term (feature) weighting approaches that have been presented in the text classification context.

Keywords: document frequency; supervised term weighting; text classification; unsupervised term weighting.

Reference to this paper should be made as follows: Alsaeedi, A. (2020) 'A survey of term weighting schemes for text classification', *Int. J. Data Mining, Modelling and Management*, Vol. 12, No. 2, pp.237–254.

Biographical notes: Abdullah Alsaeedi received his BSc in Computer Science from the College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia, in 2008, MSc in Advanced Software Engineering from the Department of Computer Science, University of Sheffield, Sheffield, UK, in 2011, and PhD in Computer Science from the University of Sheffield, UK, in 2016. He is currently an Assistant Professor at the Computer Science Department, Taibah University, Madinah, Saudi Arabia. His research interests include software engineering, software model inference, grammar inference, machine learning, data mining, and document processing.

1 Introduction

Text classification (categorisation) may be considered an interesting research point because of the necessity to categorise and organise the growing number of e-texts on the internet. Normally, text categorisation includes a feature-extraction step and a classifier which performs the categorisation process based on labelled data. Text

categorisation has been exploited in some applications such as spam e-mail filtering (Günal et al., 2006; Guzella and Caminhas, 2009), topic detection (Bracewell et al., 2009), web page categorisation (Anagnostopoulos et al., 2004; Chen and Hsieh, 2006; Özel, 2011), sentiment analysis (Vishal and Sheetal, 2016; Jagdale et al., 2019; Alsaeedi and Khan, 2019), and author identification (Cheng et al., 2011; Stamatatos, 2008). For document representation, a multi-dimensional feature vector is utilised. A weighted value such as $TF \cdot IDF$ is used to represent each dimension (Fattah, 2012; Fattah and Ren, 2008). Therefore, many (possibly several thousand) features are created for a certain text collection. An excessive number of features degrades classification accuracy and increases computational time. Hence, in the text classification task, feature selection is an essential step towards improving the accuracy and speeding up the computation. For feature selection, there are three approaches: filters, wrappers, and embedded. The filters methods are computationally fast. In these methods, features with the highest scores are selected first (Guyon and Elisseeff, 2003). The wrapper methods estimate features based on a certain learning model and search algorithm (Gunal et al., 2009; Kohavi and John, 1997). The wrapper approaches are computationally costly when they are compared with filters. The integration of feature selection phase into the training phase of a classifier is established in an embedded method, which needs less computation than the wrappers (Guyon and Elisseeff, 2003; Saeys et al., 2007).

A traditional text categorisation paradigm includes a pre-processing step, an extraction of features, a selection of features, and finally a categorisation phase. The pre-processing step normally includes tokenisation, lower-case conversion, removing of stop words, and stemming. The extraction phase of features normally relies on the vector space model representation using the bag-of-words method (Salton et al., 1975; Joachims, 1997; McCallum et al., 1998). The numbers of pre-processing techniques, such as stemming and stop word removal is intended to minimise the feature vector dimensionality and to enhance the efficiency of the text categorisation task. In the classification phase, models (classifiers) are used. Labelled documents are utilised to train the categorisation model while the learned model is exploited to classify the unlabelled documents (Fattah et al., 2006; Fattah, 2014). Support vector machines (SVMs) (Joachims, 1997; Lewis, 1998) and Naive Bayes (NB) classifiers (Joachims, 1998; Yang and Liu, 1999) have been exploited in the text classification field.

In the literature, various term-weighting schemes have been proposed. Hence, the selection of a reasonable approach may significantly affect the effectiveness of the text categorisation task. Using SVM classifiers, Leopold and Kindermann (2002) have tested various term-weighting approaches and have achieved different accuracies. The term-weighting scheme selection is essential for other text mining tasks such as text categorisation, novelty mining, cross-domain classification, and sentiment analysis. Major IT enterprise like IBM (Papineni, 2001) and Yahoo (Carmel et al., 2014) have exploited different term-weighting schemes. The most widely used term-weighting scheme is $TF \cdot IDF$ that was proposed by Jones (1972) and then Robertson (2004).

Using class information, term-weighting approaches have been investigated in different ways. One of them involves computing term weights based on well-known feature-selection metrics such as the odds ratio (OR), the gain ratio (GR), information gain (IG) and chi-square (χ^2). Another method relies on confidence intervals that are based on prior statistical information in the labelled training data (Soucy and Mineau, 2005). These approaches are expected to be performed well compared to traditional term weighting schemes as they are based on document distribution. A number of

experiments were conducted by Soucy and Mineau (2005) to compare traditional term-weighting approaches. Many researches have utilised different term-weighting approaches in feature selection for text classification problems. This survey paper will present supervised and unsupervised term-weighting approaches.

2 Term-weighting schemes

In general, the term-weighting process focuses on assigning a score to each term during the document representation processes. Text classification methods rely on suitable representation of text documents (Naderalvojoud et al., 2014). There are different models for representing text documents. In these models, the importance of any terms varies for different documents. Thus, assigning a weight (value) associated with each term is essential for representing text documents.

2.1 Unsupervised term-weighting approaches

The unsupervised methods do not rely on prior information about membership of training documents to categories (classes). The drawback of using the unsupervised term weighting approaches is that they do not consider the document distribution.

Let D denote the number of documents (also called the document size), and $d(t_i)$ is the collection number of documents that a term t_i occurs at least once such that $\{d \in D \wedge t_i \in d\}$. Let T_j denotes the set of terms appears in a specific document d_j and $|T|$ is the number of terms appears in the j^{th} document.

2.1.1 Term frequency

In the text processing methods, the term frequency (TF) $TF\ (t_i, d_j)$ for the i^{th} term (t_i) in the j^{th} document can be calculated as follows:

$$TF\,(t_i, d_j) = \frac{f_{ij}}{\sum_{t \in T_j} f_{tj}} \tag{1}$$

For a given term t_i, the f_{ij} denotes the number of times that t_i occurs in the given j^{th} document.

2.1.2 TF inverse document frequency (TF · IDF)

$TF \cdot IDF$ is a term weighting scheme that is applied widely in various data mining methods like text clustering and categorisation. $TF \cdot IDF$ is created by incorporating TF with inverse document frequency (IDF). It is widely used to measure the importance of terms appearing in documents (Lam, 1999; Tang et al., 2016). For automatic text categorisation (ATC), $TF \cdot IDF$ is the most used term weighting approach that is considered as a baseline in the literature. $TF \cdot IDF$ is calculated as:

$$TF \cdot IDF\,(t_i, d_j) = TF\,(t_i, d_j) \times \log \frac{D}{d(t_i)} \tag{2}$$

where $TF\ (t_i, d_j)$ is the TF of term t_i in document d_j and $\frac{D}{d(t_i)}$ is the *IDF* of term t_i. Haddoud et al. (2016a) claimed that there are various weaknesses of applying unsupervised term-weight approaches like $TF \cdot IDF$. In the context of text categorisation, Chen et al. (2016) claimed that $TF \cdot IDF$ is not effective due to that $TF \cdot IDF$ ignores some training document class labels.

2.1.3 TF probabilistic inverse document frequency ($TF \cdot PIDF$)

Wu and Salton (1981) introduced $TF \cdot PIDF$ term-weighting approach that replaces the *IDF* factor with probabilistic inverse document frequency (*PIDF*). $TF \cdot PIDF$ is calculated as:

$$TF \cdot PIDF\,(t_i, d_j) = TF\,(t_i, d_j) \times \log \frac{D - d(t_i)}{d(t_i)} \tag{3}$$

2.1.4 Weighted inverse document frequency

Tokunaga and Makoto (1994) proposed an IDF variant as a term-weighting approach which is called the weighted inverse document frequency (*WIDF*) and is defined as:

$$WIDF\,(t_i, d_j) = \frac{1}{\sum_{d_m \in D} TF\,(t_i, d_m)} \tag{4}$$

Deisy et al. (2010) claimed that the weakness of $TF \cdot IDF$ is that recalculation of weights to all documents is required as a new document occurs. This is due to that the fact that $TF \cdot IDF$ depends on a number of documents. *WIDF* (Tokunaga and Makoto, 1994) overcomes this issue by weighting terms that sum up to one over the collection of texts. A drawback of WIDF is that when the number of documents becomes large, the terms that have the nearest frequency have almost equal weight, which makes the learning task more difficult.

2.1.5 Modified inverse document frequency

A drawback of *WIDF* is that the terms with the nearest frequency have the same *WIDF* weight (Deisy et al., 2010). In order to overcome these drawbacks, Deisy et al. (2010) proposed a modified inverse document frequency (*MIDF*). This relies on document frequency and *TF*. It is defined as:

$$MIDF\,(t_i, d_j) = \frac{d(t_i)}{\sum_{d_m \in D} TF\,(t_i, d_m)} \tag{5}$$

The results shown in Deisy et al. (2010) demonstrated that the performance of the *SVM* classifier based on the *MIDF* term-weighting scheme is better than those based on $TF \cdot IDF$ and *WIDF* approaches.

2.1.6 Modified TF

Sabbah et al. (2017) proposed term schemes called *mTF*, *mTFIDF*, *TFmIDF*, and *mTFmIDF*. The basic idea behind proposing these term-weighting schemes is to include the missing words during the calculation of weights. The missing terms are those that are included in the feature space, but they do not appear in the document under calculation.

$$mTF(t_i, d_j) = \frac{TF(t_i, d_j) \times \log \frac{\sqrt{T_C}}{T_{t_i}}}{\log\left[\left(\sum_{t=1}^{n} TF(t_i, d_j)^2\right) \times \left(\frac{length_d^2}{\sqrt{T_C}}\right)\right]} \tag{6}$$

$$T_C = \sum_{d=j}^{D} \sum_{t=i} TF(t_i, d_j) \tag{7}$$

$$T_{t_i} = \sum_{d=j}^{D} TF(t_i, d_j) \quad \text{where} \quad TF(t, d) > 0 \tag{8}$$

In the above equation, T_{t_i} denotes the total frequency of a term t_i in the collection of documents, and T_c denotes the number of terms in whole documents. Let d denote a specific document and $length_d$ denote the number of distinctive terms in d_j, also known as the length of the document d_j. The portion $((length_d^2)/\sqrt{(T_C)})$ computes the amount of missing terms in a specific document with respect to the number of terms appearing in the document collection. Hence, the document length is considered to be the number of distinctive terms in the document.

Moreover, Sabbah et al. (2017) introduced a new term-weighting called modified IDF scheme (*mIDF*) and aims to include the number of documents in which a term does not appear during the calculation.

$$mIDF(t_i) = \log\left[\frac{D}{1/(D - d(t_i)) + 1)}\right] \tag{9}$$

where $(D - d(t_i))$ denotes the number of documents that does not contain a term t_i.

2.2 Supervised term-weighting approaches

Many studies have proven that supervised approaches are efficient compared with unsupervised methods (Gu and Gu, 2017). In the literature, various supervised term-weighting metrics were introduced to replace the *IDF* factor of $TF \cdot IDF$ with other static factors. A replaced factor uses prior knowledge about document categories and statistical information of text documents belonging to these categories. Those replaced factors include schemes such as (χ^2), GR, IG, and OR (Gu and Gu, 2017; Lan et al., 2009, 2005; Domeniconi et al., 2016).

In the literature, various studies of text classification feature selection approaches have exploited the four essential information components shown in Table 1. *TP* denotes the number of documents that belong to category c_i on the condition that the term

t_k occurs at least once such that $|\{\forall d \in D : d \in c_i \wedge t_k \in d\}|$. *FP* denotes the number of documents that do not belong to category c_i whereas the term t_k occurs at least once such that $|\{\forall d \in D : d \notin c_i \wedge t_k \in d\}|$. *FN* denotes the number of documents that belong to category c_i and the term t_k does not occur such that $|\{\forall d \in D : d \in c_i \wedge t_k \notin d\}|$. *TN* denotes the number of documents that do not belong to category c_i whereas the term t_k does not occur such that $|\{\forall d \in D : d \notin c_i \wedge t_k \notin d\}|$. Let N denote the total number of documents in the training data and it is computed using the sum of the four elements which represent the total number of training documents.

Table 1 Four fundamental classes of information used in supervised term-weighting schemes in the text classification

	c_i	$\overline{c_i}$
t_k	*TP*	*FP*
$\overline{t_k}$	*FN*	*TN*

2.2.1 Chi-square term weighting scheme

Debole and Sebastiani (2004) proposed a chi-square (χ^2) term weighting scheme that is utilised to compute how independent t_k and c_i are. This term-weighting is intended to select the terms with highest (χ^2) scores. The $TF \cdot \chi^2$ is computed as shown in equation (11). This method is quite expensive to run and is better for classifiers such as neural networks (Yang and Pedersen, 1997).

$$\chi^2 = N \times \frac{(TP \cdot TN - FP \cdot FN)^2}{(TP + FN) \cdot (FP + TN) \cdot (TP + FP) \cdot (FN + TN)} \tag{10}$$

$$TF \cdot \chi^2 = TF(t_k, d_j) \times N \times \frac{(TP \cdot TN - FP \cdot FN)^2}{(TP + FN) \cdot (FP + TN) \cdot (TP + FP) \cdot (FN + TN)} \tag{11}$$

2.2.2 Correlation coefficient

Liu et al. (2009) presented the correlation coefficient (*CC*) as term-weighting scheme in the text classification. It is considered to be a variant of χ^2 metric (Ng et al., 1997) and is defined as follows:

$$CC = nTF(t_k, d_j) \times \left[\frac{\sqrt{N}(TP \cdot TN - FP \cdot FN)}{\sqrt{(TP + FN) \cdot (FP + TN) \cdot (TP + FP) \cdot (FN + TN)}} \right] \tag{12}$$

where *nTF* is the normalised *TF* and is computed using the following equation: $nTF(t_k, d_j) = TF(t_k, d_j)/max\ [TF\,(d_j)]$. It is important to highlight that $max\ [TF\ (d_j)]$ denotes the maximum frequency of a term for the document d_j. Liu et al. (2009) showed that *CC* attained a higher accuracy compared to χ^2.

2.2.3 Information gain

In the machine-learning domain, *IG* can be used as a static criterion (weight). It relies upon knowing whether the given term is presence or absence in a document for category prediction (Yang and Pedersen, 1997). Joachims (1998) used the *IG* to elect a subset of features to eliminate the irrelevant ones. Subsequent works by Domeniconi et al. (2016), Debole and Sebastiani (2004) and Deng et al. (2004) showed that the *IG* can be used as term-weighting during the text categorisation using the following equations.

$$IG = -\frac{TP+FP}{N} \cdot \log\left(\frac{TP+FP}{N} + \frac{TP}{N}\right) \cdot \log\left(\frac{TP}{TP+FN} + \frac{FP}{N}\right) \cdot \log\left(\frac{FP}{FP+TN}\right) \tag{13}$$

$$TF \cdot IG = TF(t_k, d_j) \times -\frac{TP+FP}{N} \cdot \log\left(\frac{TP+FP}{N} + \frac{TP}{N}\right) \cdot \log\left(\frac{TP}{TP+FN} + \frac{FP}{N}\right) \cdot \log\left(\frac{FP}{FP+TN}\right) \tag{14}$$

2.2.4 Odd ratio

The *OR* (Mladenic and Grobelnik, 1998) is used to classify documents based on the relative ratio to the positive category and this ratio is computed using occurrence of features (words). It assigns higher scores to words (terms) that frequently occur in one class, but they rarely occur in the other class. Mladenic and Grobelnik (1998) investigated the scoring (the given weight to each feature) of features and their impact on document categorisation. In their experiments, six scoring features were addressed. The best performing scoring method was the *OR* compared to other methods.

$$OR = \log\left(\frac{TP \cdot TN}{FP \cdot FN}\right) \tag{15}$$

$$TF \cdot OR = TF(t_k, d_j) \times \log\left(\frac{TP \cdot TN}{FP \cdot FN}\right) \tag{16}$$

2.2.5 Gain ratio

Debole and Sebastiani (2004) introduced the *GR* function in the idea of term-weighting methods, which is computed as follows (Lan et al., 2009):

$$GR = \frac{IG}{-\frac{(TP+FP)}{N} \cdot \log\frac{(TP+FP)}{N} - \frac{(FN+TN)}{N} \cdot \log\frac{(FN+TN)}{N}} \tag{17}$$

$$TF \cdot GR = TF(t_k, d_j) \times \frac{IG}{-\frac{(TP+FP)}{N} \cdot \log \frac{(TP+FP)}{N} - \frac{(FN+TN)}{N} \cdot \log \frac{(FN+TN)}{N}} \tag{18}$$

2.2.6 Bio-normal separation

In the text classification domain, Forman (2003) proposed a feature-selection metric called bio-normal separation (*BNS*) for the *SVM* classifier. The proposed metric was compared 12 different metrics such as χ^2, *IG*, *OR*, etc. The conducted experiments revealed that the best performing feature selection methods was *BNS*. Forman (2008) used the *BNS* metric to scale the feature values magnitude. That is, the *IDF* factor is replaced by *BNS* to calculate the $TF \cdot BNS$ score is for each feature value.

$$BNS(t_k, d_j) = \left| F^{-1}\Big(\frac{TP}{TP+FN}\Big) - F^{-1}\Big(\frac{FP}{FP+TN}\Big) \right| \tag{19}$$

$$TF \cdot BNS(t_k, d_j) = TF(t_k, d_j) \times \left| F^{-1}\Big(\frac{TP}{TP+FN}\Big) - F^{-1}\Big(\frac{FP}{FP+TN}\Big) \right| \tag{20}$$

where F^{-1} is the inverse cumulative probability function.

2.2.7 Relevance frequency

$TF \cdot RF$ is a term-weighting approach that combines *TF* and *RF* functions (Lan et al., 2009, 2006). The basic idea of this term-weighting scheme is that a term with the highest frequency and more concentration in a positive class compared to the negative one, the more likely it detect positive samples from negative ones and vice versa (Lan et al., 2009). The relevance frequency (*RF*) factor is introduced to increase the discrimination power between various terms in cases where the $TF \cdot IDF$ fails to discriminate between the positive and negative documents. The conducted experiments in Lan et al. (2009, 2006) showed that $TF \cdot RF$ performed better than $TF \cdot \chi^2$, $TF \cdot OR$, and $TF \cdot IG$.

$$RF = \log\left(2 + \frac{TP}{max\ (1, FN)}\right) \tag{21}$$

$$TF \cdot RF = TF\,(t_k, d_j) \times \log\left(2 + \frac{TP}{max\ (1, FN)}\right) \tag{22}$$

2.2.8 Mutual information

In formation theory, the mutual information (*MI*) measures the association between words and classes (Yang and Pedersen, 1997). Yang and Pedersen (1997) showed that the *MI* can be computed as follows:

$$MI = \log\left(\frac{TP \cdot N}{(TP + FP) \cdot (TP + FN)}\right) \tag{23}$$

$$TF \cdot MI = TF(t_k, d_j) \times \log\left(\frac{TP \cdot N}{(TP + FP) \cdot (TP + FN)}\right) \tag{24}$$

The MI equation can be written as follows:

$$TF \cdot MI(t_k, c_i) = TF(t_k, d_j) \times \log\left(\frac{p(t_k|c_i)}{p(t_k)p(c_i)}\right) \tag{25}$$

2.2.9 Mutual information using sample variance (MIUSV)

The *MI* assigns higher scores to the terms that have a strong influence (rare terms) compared to the common terms. To reduce this problem, a new term selection called *MI* using sample variance (*MIUSV*) was proposed (Agnihotri et al., 2017).

$$MIUSV(t_k) = \max_{0 \le j \le 1} \frac{maxf(t_k, c_j) \times TF \cdot MI(t_k, c_i)}{V(t_k, c_j)} \tag{26}$$

where $maxf(t_k, C_j)$ denote the maximum frequency of term t_k in class C_j.

$$V(t_k, c_j) = \frac{1}{N-1} \times \left(a_j - \frac{a_j}{a_j + c_j}\right)^2 + \in \tag{27}$$

2.2.10 Delta TF · IDF

Martineau and Finin (2009) proposed a new supervised term-weighting approach called *Delta TF · IDF*. They proved that the *Delta TF · IDF* improves the accuracy for sentiment analysis.

$$Delta\ TF \cdot IDF(t_k, c_j) = TF(t_k, d_j) \times \log_2\left(\frac{N_{c_i} \cdot df_{\overline{c}_i}}{df_{c_i} \cdot N_{\overline{c}_i}}\right) \tag{28}$$

$$Delta\ TF \cdot IDF(t_k, c_j) = TF(t_k, d_j) \times \log_2\left(\frac{(c+d) \cdot a}{(a+b) \cdot c}\right) \tag{29}$$

where N_{c_i} and $N_{\overline{c}_i}$ represent the number of training documents in the positive class and the negative category respectively. In addition, df_{c_i} and $df_{\overline{c}_i}$ denote the number of training documents containing a term t_k in the positive and negative classes respectively. The results produced by the conducted experiments showed that the *Delta TF · idf* performed better than the *TF · IDF* method (Martineau and Finin, 2009; Paltoglou and Thelwall, 2010).

2.2.11 inverse gravity moment and TF (TF · IGM)

Chen et al. (2016) proposed a *TF · IGM* term-weighting scheme, which combines the *TF* with the *IGM* factor. A term should have more concentrated inter-class distribution in order to have better class distinguishing power compared to others. An inverse gravity moment of the inter-class distribution of term t_k is denoted as *igm* (t_k) denotes and is computed as follows:

$$igm\,(t_k) = \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr} \cdot r} \tag{30}$$

where the f_{kr} denotes the number of documents belong to the r^{th} class containing the t_k term and f_{k1} denotes the frequency of t_k in the 1st class.

$$w_g\,(t_k) = 1 + \lambda \cdot igm\,(t_k) \tag{31}$$

where $w_g\,(t_k)$ is the *IGM*-based global weighting factor for a term t_k and λ is an adaptable coefficient.

$$TF \cdot IGM\,(t_k, d_j) = TF\,(t_k, d_j) \times \left(1 + \lambda \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr} \cdot r}\right) \tag{32}$$

It is important to highlight that *RTF · IGM* is an improved modification of *TF · IGM*.

$$RTF \cdot IGM\,(t_k, d_j) = \sqrt{TF\,(t_k, d_j)} \times \left(1 + \lambda \frac{f_{k1}}{\sum_{r=1}^{m} f_{kr} \cdot r}\right) \tag{33}$$

The performance of *TF · IGM* was evaluated by experiments on three commonly used corpora and *SVM* and *k-NN* classifiers. The conducted experiments showed that *TF · IGM* outperformed *TF · IDF* and other supervised term-weighting approaches.

2.2.12 TF · ICF

Deqing and Zhang (2013) proposed inverse category frequency (*ICF*) to be a part of the term-weighting scheme. Before defining *ICF*, let class frequency (*CF*) denote the number of classes in which a term t_k appears, $\mid C \mid$ denotes the number of categories in training documents. Let *ICF*, which is defined in the same way as *IDF* and it is calculated as:

$$ICF\,(t_k) = \log\left(\frac{\mid C \mid}{cf(t_k)}\right) \tag{34}$$

$$TF \cdot ICF\,(t_k, d_j) = TF\,(t_k, d_j) \times \log\left(\frac{\mid C \mid}{cf(t_k)}\right) \tag{35}$$

$$ICF\text{-}based\,(t_k, d_j) = TF\,(t_k, d_j) \times \log\left(2 + \frac{a}{max\,(1, c)} \times \frac{\mid C \mid}{cf(t_k)}\right) \tag{36}$$

2.2.13 Inverse term entropy

The majority of supervised feature weighting schemes were designed for binary text classification. In other words, supervised term-weighting approaches cannot be used for multi-class classification (Gu and Gu, 2017). In addition, Gu and Gu (2017) introduced a novel term-weighting method called inverse term entropy (*ITE*) for multi-class categorisation.

$$g_k = b_0 + (1 - b_0)\left[log_2 \mid C \mid + \sum_{i \in [1,|C|]} \frac{(TP+1)/N_i}{\sum_{i \in [1,|C|]}(TP+1)/N_i} \log_2 \frac{(TP+1)/N_i}{\sum_{i \in [1,|C|]}(TP+1)/N_i}\right] \tag{37}$$

where N_i is the number of documents belong to class c_i such that

$$N_i = TP + FN \tag{38}$$

The proposed method by Gu and Gu (2017) has been proven to reduce the problem of over-weighting. The conducted experiments proved that *ITE* outperforms other supervised weighting schemes.

2.2.14 Inverse document frequency excluding category

Inverse document frequency excluding category (*IDFEC*) is a variant of $TF \cdot IDF$ that has been proposed in Domeniconi et al. (2016). The basic idea is that weights assigned to terms appear in documents that belong to the same class should be increased; therefore, these terms are not poor.

$$TF \cdot IDFEC\,(t_k, d_j) = TF\,(t_k, d_j) \times \log\left(\frac{FN + TN}{max\ (1, FN)}\right) \tag{39}$$

It is clear that $TF \cdot IDFEC$ takes only negative examples. Domeniconi et al. (2016) proposed to take into consideration documents that belong to a category c_k and contain the term. Hence, they introduced a new term-weighting scheme called $TF \cdot IDFEC$-*based* to include both negative and positive documents.

$$TF \cdot IDFEC\text{-}based\,(t_k, d_j) = TF\,(t_k, d_j) \times \log\left(2 + \frac{TP + FN + TN}{max\ (1, FN)}\right) \tag{40}$$

2.2.15 SW term weighting scheme

Alsmadi and Hoon (2018) proposed a *SW* term-weighting method to deal with the short text classification. The *SW* term-weighting method is computed using equation (41). The conducted experiment showed that *SVM* classifier with *SW* term-weighting method attained the best accuracy compared to $TF \cdot IDF$, $TF \cdot RF$, $TF \cdot \chi^2$, and $TF \cdot IG$.

$$SW(t_k, d_j) = \frac{TF\,(t_k, d_j) + 1}{\sum_{i=1}^{|T|} TF\,(t_k, d_j) + \mid T \mid} \times \log\left(1 + \frac{TP}{FP + FN + 1}\right) \tag{41}$$

2.2.16 Log TF · TRR term-weighting scheme

Youngjoong (2015) proposed a new term-weighting by replacing the *IDF* factor by term relevance ratio (*TRR*). This computes the probabilities of negative and positive classes. A negative class is denoted by ($\overline{cl}$) and a positive one is indicated by cl.

$$log\ TF \cdot TRR = \left(\log \left(TF\left(t_k, d_j\right) \right) + 1 \right) \cdot \log \left[\frac{P(t_k \mid cl)}{P(t_k \mid \overline{cl})} + \alpha \right] \tag{42}$$

where α is used as a constant value to make the logarithmic value a positive value. Youngjoong (2015) suggested using the same notation proposed by Lan et al. (2009).

$$P\left(t_k \mid cl\right) = \frac{TP}{TP + FP} \tag{43}$$

$$P\left(t_k \mid \overline{cl}\right) = \frac{FN}{FN + TN} \tag{44}$$

2.2.17 Prob-based term-weighting scheme

Liu and Loh (2007) proposed a term-weighting scheme based on the probability that relies on relevance indicators using probability estimations. The ratio TP/FP tends to be high if the term t_k is relevant to the category c_i more than others categories. Let t_l, t_k denote two terms and c_i refers to a specific class, the term which has a higher value of TP/FN will be a representative feature for c_i; this means that a larger part of the feature appears in class c_i. The strength of a term can be computed using TP/FP and TP/FN which are considered to be relevant indicators.

$$Prob\text{-}based = TF\left(t_k, d_j\right) \cdot \left(1 + \frac{TP}{FP} \cdot \frac{TP}{FN} \right) \tag{45}$$

2.2.18 $logtf \cdot rf_{max}$ term weighting approach

Xuan and Quang (2014) proposed a term-weighting method called $logtf \cdot rf_{max}$. The OneVsAll approach was used to transform the multi-label categorisation task of N classes into N binary categorisation tasks Lan et al. (2009). For each given term, N rf values are required for each category for each binary classification. The $logtf \cdot rf_{max}$ term weighting uses OneVsAll approach and assigns a single rf_{max} for each term for all binary classification.

$$logtf \cdot rf_{max} = log_2\left(1.0 + tf\right) \times \max_{i=1}^{N} \left\{ RF\left(C_i\right) \right\} \tag{46}$$

3 Discussion and summary

Deng et al. (2004) used the $TF \cdot \chi^2$ to weight terms in their experiments with the *SVM* classifier and showed that $TF \cdot \chi^2$ outperformed $TF \cdot IDF$ and $TF \cdot OR$. However, Debole and Sebastiani (2004) assigned term weights using *GR*, χ^2, and *IG* methods.

These approaches did not show any clear improvement over the traditional $TF \cdot IDF$ term-weighting.

Batal and Hauskrecht (2009) have demonstrated that the performance of the *KNN* model may be significantly enhanced, as IG and χ^2 are used to weight terms during the feature selection step. Haddoud et al. (2016b) proposed combining multiple predictions obtained from *SVM* classifiers, where each of them used one of the 96 metrics. The collected predictions were used as inputs for the final *SVM* classifier. This aimed to enhance the classification accuracy. The *SVM* model performed well in all combinations of multiple metrics.

Deqing and Zhang (2013) investigated the shortness of $TF \cdot RF$ and *prob-based* schemes. Terms distribution among categories disappeared when dividing the training corpus into positive and negative categories. The experiments conduced by Deqing and Zhang (2013) and Domeniconi et al. (2015) showed that $TF \cdot ICF$ and *ICF-based* term-weighting schemes outperformed other existing approaches such as $TF \cdot RF$ and $TF \cdot IDF$ (Deqing and Zhang, 2013; Domeniconi et al., 2015). It is clear that $TF \cdot ICF$ is suitable for the multi-class classification tasks, and *ICF-based* is appropriate for the binary classification tasks (Deqing and Zhang, 2013). Guru et al. (2019) used the KNN classifier to classify Arabic texts and showed that $TF \cdot ICF$ supervised term-weighting performed better than $TF \cdot IDF$. This proves the applicability of applying term weighting schemes to different languages.

According to Lan et al. (2009), $TF \cdot RF$ performed consistently better than $TF \cdot \chi^2$, $TF \cdot OR$, $TF \cdot IDF$, and $TF \cdot IG$ on Reuters and 20 Newsgroups corpora. Moreover, $TF \cdot \chi^2$ and $TF \cdot IG$ performed worse than any other schemes, while $TF \cdot OR$ outperformed $TF \cdot \chi^2$ and $TF \cdot IG$ (Lan et al., 2009). On the other hand, the experiments conducted by Domeniconi et al. (2016) showed that the performance of $TF \cdot IDFEC\text{-}based$ and $TF \cdot RF$ were very similar without one clearly outperforming the other. However, the performance of $TF \cdot RF$ was more stable as long as there was a high number of features, but $TF \cdot IDFEC\text{-}based$ term-weighting method were slightly better when the considered number of features was small Domeniconi et al. (2016). Based on the results presented in Domeniconi et al. (2016), both $TF \cdot IDFEC\text{-}based$ and $TF \cdot IDFEC$ outperformed $TF \cdot IDF$ and $TF \cdot ICF\text{-}based$ on Reuters-52 and 20 Newsgroups datasets.

Chen et al. (2016) selected *SVM* and *kNN* classifiers for measuring the effectiveness of $TF \cdot IGM$ and $RTF \cdot IGM$. The performance of $TF \cdot IGM$ and $RTF \cdot IGM$ (Chen et al., 2016) were proven to outperform other term-weighting approaches, especially with the multi-class classification problems. The $TF \cdot IGM$ and $RTF \cdot IGM$ were shown to perform better than $TF \cdot RF$.

Based on the experimental results shown in Youngjoong (2015), the performance of $\log TF \cdot TRR$ was consistently better than other weighting schemes over all the classifiers and datasets. Besides which, the $\log TF \cdot TRR$ term-weighting approaches performed well and outperformed $TF \cdot IDF$, $TF \cdot RF$, and *Delta* $TF \cdot idf$. On the other hand, the conducted experiments in Xuan and Quang (2014) showed that $logtf \cdot rf_{max}$ outperformed $TF \cdot RF$.

In conclusion, based on the summary of term-weighting performance, it is obvious that $TF \cdot IDFEC\text{-}based$ and $TF \cdot IDFEC$ performed well for binary classification. In addition, $TF \cdot ICF$, $TF \cdot IGM$ and $RTF \cdot IGM$ term weighting schemes are preferable for multi-label classifications. It is important to highlight that term-weighting approaches can be applied for real-life applications such as sentiment analysis (Alsmadi and Hoon,

2018). Zin et al. (2018) studied the effect of term-weighting on the performance of SVM classifier in sentiment analysis of movie reviews. In addition, Parlak and Uysal (2018) used TF and $TF \cdot IDF$ as term-weighting to classify medical documents.

4 Conclusions

In this work, unsupervised and supervised term-weighting schemes have been investigated for text classification tasks. Based on this survey, it can be said that that supervised term weighting approaches such as $TF \cdot ICF$, $TF \cdot IDFEC$-*based* and $TF \cdot IDFEC$ are superior to the traditional unsupervised term-weighting approaches such as $TF \cdot IDF$ in terms of the total system performance. However, unsupervised term-weighting approaches are simple and fast in general. On the other hand, supervised term weighting approaches require annotated training data. Potentially interesting future research could include comparing all the term weighting schemes presented in this survey on different corpora.

References

Agnihotri, D., Verma, K. and Tripathi, P. (2017) 'Mutual information using sample variance for text feature selection', in *Proceedings of the 3rd International Conference on Communication and Information Processing, ICCIP'17*, ACM, New York, NY, USA, pp.39–44.

Alsaeedi, A. and Khan, M.Z. (2019) 'A study on sentiment analysis techniques of twitter data', *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 2, pp.361–374.

Alsmadi, I. and Hoon, G.K. (2018) 'Term weighting scheme for short-text classification: Twitter corpuses', *Neural Computing and Applications*, January, pp.1–13.

Anagnostopoulos, I., Anagnostopoulos, C., Loumos, V. and Kayafas, E. (2004) 'Classifying web pages employing a probabilistic neural network', *IEE Proceedings-Software*, Vol. 151, No. 3, pp.139–150.

Batal, I. and Hauskrecht, M. (2009) 'Boosting knn text classification accuracy by using supervised term weighting schemes', in *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM'09*, ACM, New York, NY, USA, pp.2041–2044.

Bracewell, D.B., Yan, J., Ren, F. and Kuroiwa, S. (2009) 'Category classification and topic discovery of Japanese and English news articles', *Electronic Notes in Theoretical Computer Science, Proceedings of the Irish Conference on the Mathematical Foundations of Computer Science and Information Technology (MFCSIT 2006)*, Vol. 225, pp.51–65.

Carmel, D., Mejer, A., Pinter, Y. and Szpektor, I. (2014) 'Improving term weighting for community question answering search using syntactic analysis', in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM'14*, ACM, New York, NY, USA, pp.351–360.

Chen, R-C. and Hsieh, C-H. (2006) 'Web page classification based on a support vector machine using a weighted vote schema', *Expert Systems with Applications*, Vol. 31, No. 2, pp.427–435.

Chen, K., Zhang, Z., Long, J. and Zhang, H. (2016) 'Turning from TF-IDF to TF-IGM for term weighting in text classification', *Expert Systems with Applications*, Vol. 66, pp.245–260.

Cheng, N., Chandramouli, R. and Subbalakshmi, K.P. (2011) 'Author gender identification from text', *Digital Investigation*, Vol. 8, No. 1, pp.78–88.

Debole, F. and Sebastiani, F. (2004) 'Supervised term weighting for automated text categorization', in Sirmakessis, S. (Ed.): *Text Mining and Its Applications*, pp.81–97, Springer, Berlin, Heidelberg.

Deisy, C., Gowri, M., Baskar, S., Kalaiarasi, S.M.A. and Ramraj, N. (2010) 'A novel term weighting scheme midf for text categorization', *Journal of Engineering Science and Technology*, Vol. 5, No. 1, pp.94–107.

Deng, Z-H., Tang, S-W., Yang, D-Q., Li, M.Z., Li, Y. and Xie, K-Q. (2004) 'A comparative study on feature weight in text categorization', in Yu, J.X., Lin, X., Lu, H. and Zhang, Y. (Eds.): *Advanced Web Technologies and Applications*, pp.588–597, Springer, Berlin, Heidelberg.

Deqing, W. and Zhang, H. (2013) 'Inverse-category-frequency based supervised term-weighting schemes for text categorization', *Journal of Information Science and Engineering*, Vol. 29, No. 2, pp.209–225.

Domeniconi, G., Moro, G., Pasolini, R. and Sartori, C. (2015) 'A study on term weighting for text categorization: a novel supervised variant of tf.idf', in *Proceedings of 4th International Conference on Data Management Technologies and Applications, Data 2015*, SCITEPRESS – Science and Technology Publications, LDA, Portugal, pp.26–37.

Domeniconi, G., Moro, G., Pasolini, R. and Sartori, C. (2016) 'A comparison of term weighting schemes for text classification and sentiment analysis with a supervised variant of tf.idf', in Helfert, M., Holzinger, A., Belo, O. and Francalanci, C. (Eds.): *Data Management Technologies and Applications*, pp.39–58, Springer International Publishing, Cham.

Fattah, M.A. (2012) 'The use of MSVM and hmm for sentence alignment', *Journal of Information Processing Systems*, Vol. 8, No. 2, pp.301–314.

Fattah, M.A. (2014) 'A hybrid machine learning model for multi-document summarization', *Applied Intelligence*, June, Vol. 40, No. 4, pp.592–600.

Fattah, M.A. and Ren, F. (2008) 'Probabilistic neural network based text summarization', in *2008 International Conference on Natural Language Processing and Knowledge Engineering*, October, pp.1–6.

Fattah, M.A., Ren, F. and Kuroiwa, S. (2006) 'Effects of phoneme type and frequency on distributed speaker identification and verification', *IEICE Transactions on Information and Systems*, Vol. 89, No. 5, pp.1712–1719.

Forman, G. (2003) 'An extensive empirical study of feature selection metrics for text classification', *J. Mach. Learn. Res.*, March, Vol. 3, pp.1289–1305.

Forman, G. (2008) 'BNS feature scaling: an improved representation over TF-IDF for SVM text classification', in *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM'08*, ACM, New York, NY, USA, pp.263–270.

Günal, S., Ergin, S., Bilginer Gülmezoğlu, M. and Gerek, Ö. (2006) 'On feature extraction for spam e-mail detection', in *International Workshop on Multimedia Content Representation, Classification and Security*, Springer, pp.635–642.

Gu, Y. and Gu, X. (2017) 'A supervised term weighting scheme for multi-class text categorization', in Huang, D-S., Hussain, A., Han, K. and Gromiha, M.M. (Eds.): *Intelligent Computing Methodologies*, pp.436–447, Springer International Publishing, Cham.

Gunal, S., Gerek, O.N., Ece, D.G. and Edizkan, R. (2009) 'The search for optimal feature set in power quality event classification', *Expert Systems with Applications*, Vol. 36, No. 7, pp.10266–10273.

Guru, D.S., Ali, M., Suhil, M. and Hazman, M. (2019) 'A study of applying different term weighting schemes on Arabic text classification', in *Data Analytics and Learning*, pp.293–305, Springer, Singapore.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, March, Vol. 3, pp.1157–1182.

Guzella, T.S. and Caminhas, W.M. (2009) 'A review of machine learning approaches to spam filtering', *Expert Systems with Applications*, Vol. 36, No. 7, pp.10206–10222.

Haddoud, M., Mokhtari, A., Lecroq, T. and Abdeddaïm, S. (2016a) 'Supervised term weights for biomedical text classification: improvements in nearest centroid computation', in Angelini, C., Rancoita, P.M.V. and Rovetta, S. (Eds.): *Computational Intelligence Methods for Bioinformatics and Biostatistics*, pp.98–113, Springer International Publishing, Cham.

Haddoud, M., Mokhtari, A., Lecroq, T. and Abdeddaïm, S. (2016b) 'Combining supervised term-weighting metrics for SVM text classification with extended term representation', *Knowledge and Information Systems*, December, Vol. 49, No. 3, pp.909–931.

Jagdale, R.S., Shirsat, V.S. and Deshmukh, S.N. (2019) 'Sentiment analysis on product reviews using machine learning techniques', in Mallick, P.K., Balas, V.E., Bhoi, A.K. and Zobaa, A.F. (Eds.): *Cognitive Informatics and Soft Computing*, pp.639–647, Springer, Singapore.

Joachims, T. (1997) 'A probabilistic analysis of the rocchio algorithm with tfidf for text categorization', in *Proceedings of the 14th International Conference on Machine Learning, ICML'97*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp.143–151.

Joachims, T. (1998) 'Text categorization with support vector machines: learning with many relevant features', in Nédellec, C. and Rouveirol, C. (Eds.): *Machine Learning: ECML-98*, Springer, Berlin, Heidelberg, pp.137–142.

Jones, K.S. (1972) 'A statistical interpretation of term specificity and its application in retrieval', *Journal of Documentation*, Vol. 28, No. 1, pp.11–21.

Kohavi, R. and John, G.H. (1997) 'Wrappers for feature subset selection', *Artificial Intelligence*, Vol. 97, No. 1, pp.273–324, Relevance.

Lam, S.L.Y. (1999) 'Feature reduction for neural network based text categorization', in *Proceedings. 6th International Conference on Advanced Systems for Advanced Applications*, April, pp.195–202.

Lan, M., Sung, S-Y., Low, H-B. and Tan, C-L. (2005) 'A comparative study on term weighting schemes for text categorization', in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, July, Vol. 1, pp.546–551.

Lan, M., Tan, C-L. and Low, H-B. (2006) 'Proposing a new term weighting scheme for text categorization', in *Proceedings of the 21st National Conference on Artificial Intelligence, AAAI'06*, AAAI Press, Vol. 1, pp.763–768.

Lan, M., Tan, C.L., Su, J. and Lu, Y. (2009) 'Supervised and traditional term weighting methods for automatic text categorization', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, April, Vol. 31, No. 4, pp.721–735.

Leopold, E. and Kindermann, J. (2002) 'Text categorization with support vector machines. how to represent texts in input space?', *Machine Learning*, January, Vol. 46, No. 1, pp.423–444.

Lewis, D.D. (1998) 'Naive (Bayes) at forty: the independence assumption in information retrieval', in Nédellec, C. and Rouveirol, C. (Eds.): *Machine Learning: ECML-98*, Springer, Berlin, Heidelberg, pp.4–15.

Liu, Y. and Loh, H.T. (2007) 'A simple probability based term weighting scheme for automated text classification', in Okuno, H.G. and Ali, M. (Eds.): *New Trends in Applied Artificial Intelligence*, pp.33–43, Springer, Berlin, Heidelberg.

Liu, Y., Loh, H.T. and Sun, A. (2009) 'Imbalanced text classification: a term weighting approach', *Expert Systems with Applications*, Vol. 36, No. 1, pp.690–701.

Martineau, J. and Finin, T. (2009) 'Delta TFIDF: an improved feature space for sentiment analysis', *ICWSM*, Vol. 9, p.106.

McCallum, A., Nigam, K. et al. (1998) 'A comparison of event models for Naive Bayes text classification', in *AAAI-98 Workshop on Learning for Text Categorization*, Citeseer, Vol. 752, pp.41–48.

Mladenic, D. and Grobelnik, M. (1998) 'Feature selection for classification based on text hierarchy', in *Text and the Web, Conference on Automated Learning and Discovery CONALD-98*.

Naderalvojoud, B., Bozkir, A.S. and Sezer, E.A. (2014) 'Investigation of term weighting schemes in classification of imbalanced texts', in *Proceedings of European Conference on Data Mining (ECDM)*, Lisbon, pp.15–17.

Ng, H.T., Goh, W.B. and Low, K.L. (1997) 'Feature selection, perceptron learning, and a usability case study for text categorization', in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'97*, ACM, New York, NY, USA, pp.67–73.

Özel, S.A. (2011) 'A web page classification system based on a genetic algorithm using tagged-terms as features', *Expert Systems with Applications*, Vol. 38, No. 4, pp.3407–3415.

Paltoglou, G. and Thelwall, M. (2010) 'A study of information retrieval weighting schemes for sentiment analysis', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL'10*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp.1386–1395.

Papineni, K. (2001) 'Why inverse document frequency?', in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, NAACL'01*, Association for Computational Linguistics, Stroudsburg, PA, USA, pp.1–8.

Parlak, B. and Uysal, A.K. (2018) 'On feature weighting and selection for medical document classification', in Rocha, Á. and Reis, L.P. (Eds.): *Developments and Advances in Intelligent Systems and Applications*, pp.269–282, Springer International Publishing, Cham.

Robertson, S. (2004) 'Understanding inverse document frequency: on theoretical arguments for IDF', *Journal of Documentation*, Vol. 60, No. 5, pp.503–520.

Sabbah, T., Selamat, A., Selamat, M.H., Al-Anzi, F.S., Viedma, E.H., Krejcar, O. and Fujita, H. (2017) 'Modified frequency-based term weighting schemes for text classification', *Applied Soft Computing*, Vol. 58, pp.193–206.

Saeys, Y., Inza, I. and Larrañaga, P. (2007) 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, Vol. 23, No. 19, pp.2507–2517.

Salton, G., Wong, A. and Yang, C.S. (1975) 'A vector space model for automatic indexing', *Commun. ACM*, November, Vol. 18, No. 11, pp.613–620.

Soucy, P. and Mineau, G.W. (2005) 'Beyond TFIDF weighting for text categorization in the vector space model', in *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI'05*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp.1130–1135.

Stamatatos, E. (2008) 'Author identification: Using text sampling to handle the class imbalance problem', *Information Processing & Management, Evaluating Exploratory Search Systems Digital Libraries in the Context of Users' Broader Activities*, Vol. 44, No. 2, pp.790–799.

Tang, B., Kay, S. and He, H. (2016) 'Toward optimal feature selection in Naive Bayes for text categorization', *IEEE Transactions on Knowledge and Data Engineering*, September, Vol. 28, No. 9, pp.2508–2521.

Tokunaga, T. and Makoto, I. (1994) 'Text categorization based on weighted inverse document frequency', in *Special Interest Groups and Information Process Society of Japan (SIG-IPSJ)*, pp.33–39.

Vishal, A.K. and Sheetal, S. (2016) 'Sentiment analysis of twitter data: a survey of techniques', *CoRR*, DOI: abs/1601.06971.

Wu, H. and Salton, G. (1981) 'A comparison of search term weighting: term relevance vs. inverse document frequency', *SIGIR Forum*, May, Vol. 16, No. 1, pp.30–39.

Xuan, N.P. and Quang, H.L. (2014) 'A new improved term weighting scheme for text categorization', in Huynh, V.N., Denoeux, T., Tran, D.H., Le, A.C. and Pham, S.B. (Eds.): *Knowledge and Systems Engineering*, pp.261–270, Springer International Publishing, Cham.

Yang, Y. and Liu, X. (1999) 'A re-examination of text categorization methods', in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'99*, ACM, New York, NY, USA, pp.42–49.

Yang, Y. and Pedersen, J.O. (1997) 'A comparative study on feature selection in text categorization', in *Proceedings of the 14th International Conference on Machine Learning, ICML'97*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp.412–420.

Youngjoong, K. (2015) 'A new term-weighting scheme for text classification using the odds of positive and negative class probabilities', *Journal of the Association for Information Science and Technology*, Vol. 66, No. 12, pp.2553–2565.

Zin, H.M., Mustapha, N., Murad, M.A.A. and Sharef, N.M. (2018) 'Term weighting scheme effect in sentiment analysis of online movie reviews', *Advanced Science Letters*, Vol. 24, No. 2, pp.933–937.