Journal of Informetrics 14 (2020) 101076

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

A novel term weighting scheme for text classification: TF-MONO

Turgut Dogan [a,*], Alper Kursat Uysal [b]

[a] Department of Computer Engineering, Trakya University, Edirne, Turkey
[b] Department of Computer Engineering, Eskişehir Technical University, Eskisehir, Turkey

ARTICLE INFO

Article history:
Received 7 February 2020
Received in revised form 6 July 2020
Accepted 8 July 2020

Keywords:
Text classification
Supervised term weighting
Max-occurrence
Non-occurrence

ABSTRACT

The effective representation of the relationship between the documents and their contents is crucial to increase classification performance of text documents in the text classification. Term weighting is a preprocess aiming to represent text documents better in Vector Space by assigning proper weights to terms. Since the calculation of the appropriate weight values directly affects performance of the text classification, in the literature, term weighting is still one of the important sub-research areas of text classification. In this study, we propose a novel term weighting (MONO) strategy which can use the non-occurrence information of terms more effectively than existing term weighting approaches in the literature. The proposed weighting strategy also performs intra-class document scaling to supply better representations of distinguishing capabilities of terms occurring in the different quantity of documents in the same quantity of class. Based on the MONO weighting strategy, two novel supervised term weighting schemes called TF-MONO and SRTF-MONO were proposed for text classification. The proposed schemes were tested with two different classifiers such as SVM and KNN on 3 different datasets named Reuters-21578, 20-Newsgroups, and WebKB. The classification performances of the proposed schemes were compared with 5 different existing term weighting schemes in the literature named TF-IDF, TF-IDF-ICF, TF-RF, TF-IDF-ICSDF, and TF-IGM. The results obtained from 7 different schemes show that SRTF-MONO generally outperformed other schemes for all three datasets. Moreover, TF-MONO has promised both Micro-F1 and Macro-F1 results compared to other five benchmark term weighting methods especially on the Reuters-21578 and 20-Newsgroups datasets.

1. Introduction

Text classification/categorization is one of the most widely used research areas for easy access, organization and categorization of electronic documents that have considerably increased in the digital environment. In text classification, textual documents are assigned to predetermined categories according to their contents. Author recognition (Fourkioti, Symeonidis, & Arampatzis, 2019), sentiment analysis (Peng, Ma, Li, & Cambria, 2018; Yang, Zhang, Jiang, & Li, 2019; Yoon, Kim, Kim, & Song, 2016), SMS spam filtering (Abayomi-Alli, Misra, Abayomi-Alli, & Odusami, 2019) and short text classification (Alsmadi & Hoon, 2019) can be shown as different sub domains of text classification in literature. In addition to textual data, bibliometric data also have been frequently used by researchers in recent years for various research studies in which machine learning

* Corresponding author.
E-mail addresses: turgutdogan@trakya.edu.tr (T. Dogan), akuysal@eskisehir.edu.tr (A.K. Uysal).

https://doi.org/10.1016/j.joi.2020.101076
1751-1577/

and text classification methods are employed. Term weighting methods, one of the significant text classification steps, are employed in corresponding studies. In other words, in the literature, there exists also studies (Zhang, Wang, Gottwalt, Saberi, & Chang, 2019) where term weighting is proposed for bibliometric data. The text classification process generally contains four stages that are feature extraction and selection, feature (term) weighting and classification. In the feature extraction stage, text documents are mostly represented with bag-of-words technique after extracting the numeric information from the raw text documents (Joachims, 1996). This technique ignores the order of terms in the text documents while representing each text document. During feature extraction, some preprocessing methods such as tokenization, stop-word removal, lowercase conversion and stemming are applied to the text contents depending on the type of work (Uysal & Gunal, 2014). Feature selection is the stage in which feature selection methods are used to discover distinctive features, especially when working with high-dimensional text datasets (Feng, Guo, Jing, & Sun, 2015; Huang et al., 2019; Parlak & Uysal, 2019; Uysal, 2016; Wang & Hong, 2019). The relationships between selected/extracted features and documents are determined in the term weighting stage. In the Vector Space Model, each document is denoted as feature vector (Dogan & Uysal, 2019). The concept of weight states the contribution of each features to the distinctiveness of each document in whole collection. Term weighting is a crucial stage because assigning suitable weights to terms effects the classification performance positively. Therefore, studies aiming to develop term weighting methods (schemes) are still one of the vital parts of the text classification/categorization (Badawi & Altınçay, 2014; Deisy, Gowri, Baskar, Kalaiarasi, & Ramraj, 2010; Dogan & Uysal, 2019; Escalante et al., 2015; Feng, Li, Sun, & Zhang, 2018; Khreisat, 2009; Kim & Kim, 2016; Lan, Tan, Su, & Lu, 2009; Wu, Gu, & Gu, 2017; Zhang et al., 2019).

Term weighting schemes generally can be expressed with two groups which are unsupervised and supervised schemes. Unsupervised schemes (Sparck Jones, 2004) do not use the class information of documents on the weighting process while supervised schemes try to effectively use class information of the documents on the weight calculating phase of terms (Debole & Sebastiani, 2004). When looking at the history of the term weighting methods from past to present, it is possible to say that supervised methods show better classification performance than unsupervised methods (Chen, Zhang, Long, & Zhang, 2016; Dogan & Uysal, 2019; Ren & Sohrab, 2013). Therefore, recent researches in term weighting methods do not only focus on using or not using class information of terms. They mostly focus on using class information of terms more effectively.

Some of the term weighting methods calculate the innumerable (class-specific weight for each term) weight values for each term, while others compute a single weight score (showing importance of term over all classes in the collection) for each term. The methods generating class-based scores for each term also have binary classification approach to weighting (Lan et al., 2009; Liu, Loh, & Sun, 2009). In other words, they calculate the class-based weights of any term according to their distribution inside the positive and negative classes. However, since only one weight score can be used for each feature in the Vector Space Model, these types of weighting methods have various globalization policies (Dogan & Uysal, 2018) to convert generated class-based scores into single score. It can be stated that the weighting methods computing a single weight score are suitable for multi-class classification such as TF-IGM (Chen et al., 2016). Concerning the successes of the recent term weighting schemes, it seems that the term weighting methods, which are suitable for multi-class classification, are generally more successful than the weighting methods applying binary classification approach.

The supervised term weighting schemes existing in the literature targets at effective use of the class information of the documents in the term weighting process. For this purpose, document distributions in classes in which terms occurs are generally focal point in proposed supervised schemes. While calculating the weight of any term, indeed, the documents' distribution information in the class (or classes) where related term do not occur is important as the documents' distribution information in the class (or classes) where related term occurs. This hypothesis could be explained as following: If a term occurs in all documents (if possible) belonging to same class and slightly occur in all documents (if possible) belonging to rest of the classes, this term is distinctive. On the other hand, if a term occurs in all documents (if possible) belonging to same class (or classes) and does not occur in all documents (if possible) belonging to rest of the classes, this term is highly distinctive. Thus, the information of document-distributions belonging to all classes where the terms do not occur must also be part of the weighting process of terms.

The aim of this study is to present a new weighting strategy including document-distributions of the classes in which the terms do not occur as well as the class-information of the documents in which the terms do not occur. This strategy aims to assign more effective weights to terms in the supervised weighting process. Our weighting strategy is based on the calculation of two rates called maximum-occurrence (MO) and non-occurrence (NO). The MO ratio is based on the document-distribution of the class where the term occurs most, whereas the NO ratio is based on the document-distributions in the remaining classes where term does not occur. Two novel term/feature weighting schemes, TF-MONO and SRTF-MONO, are proposed based on the weighting strategy called MONO.

The proposed term/feature weighting schemes are compared with 5 different popular state-of-the-art schemes named TF-IDF, TF-IDF-ICF, TF-RF, TF-IDF-ICSDF and TF-IGM by using SVM and KNN classifiers upon the Reuters-21578, 20-Newsgroups, and WebKB datasets. The experimental results indicate that the proposed SRTF-MONO term weighting method generally outperformed the performances of other 5 popular term weighting methods over all datasets and with both classifiers. Moreover, the TF-MONO term weighting scheme has promising results on the popular text datasets like Reuters-21578, 20-Newsgroups compared with other five term weighting schemes including TF-IGM.

The rest of the study is organized as follows: Studies about existing term weighting methods in the literature are summarized in Section 2. Existing term weighting schemes employed in the experiments and proposed term weighting schemes

are given in Sections 3 and 4, respectively. Datasets, experimental settings, classifiers, success measures, and experimental results are presented in the Section 5. Lastly, the conclusions are presented in Section 6.

2. Related works

The origin of proposed methods about term weighting extends to TF-IDF (Sparck Jones, 2004). Inverse Document Frequency (IDF), indeed, is a method proposed by Karen Spärck-Jones to select effective query selection in Information Retrieval domain (Sparck Jones, 2004). This selection method is adapted for term weighting by combining with term frequency (TF). Sabbah et al. proposed 4 term weighting schemes (mTFmIDF, TFmIDF, mTFIDF, and mTF) by editing standard TF and IDF factors (Sabbah et al., 2017). They stressed that main characteristic of these schemes is inclusion of quantity of deficient terms to the weighting process of terms, additionally. They demonstrated that proposed mTFIDF, mTF, and mTFmIDF schemes supply better results than popular term weighting schemes named Entropy, TF-IDF, and TF. They also verified that performance gains obtained with proposed schemes that employ SVM, KNN, NB, and ELM classifiers on the 3 well-known text datasets are statistically significant.

Debole and Sebastiani compared the performances of standard TF-IDF with three term weighting schemes (TF-CHI, TF-GR and TF-IG) which are adapted from three different feature selection methods named Chi-Square, Gain Ratio and, Information Gain (Debole & Sebastiani, 2004). They used the concept of "Supervised term weighting" first time in this study and they stated that use of class information on weighting process could improve the classification success of term weighting scheme.

Lan et al. introduced a novel term weighting scheme namely TF-RF based on 'relevance frequency' (Lan et al., 2009). Relevance frequency focuses on the ratio of the quantity of text documents belonging to positive and negative classes where term occurs. They show that TF-RF scheme has a better performance over three unsupervised schemes (TF, Binary and TF-IDF) and three supervised schemes (TF-IG, TF-CHI2, and TF-LogOR). In another research, Xuan and Le Quang introduced LogTF-RF_{max} scheme for text classification (Xuan & Le Quang, 2014). They showed that combining RF with reduced term frequency values increases the performance of TF-RF. Liu et al. proposed TF-PB term weighting approach for classification of unbalanced text datasets (Liu et al., 2009). TF-PB is based on calculation of two ratios showing inter-class distribution and intra-class distribution of terms. They demonstrated that using inner-class distribution information of a term on the weighting can increase the classification accuracies of documents belonging to minor categories on unbalanced text datasets. Ko introduced Log_TF-TRR term weighting scheme based on class information by utilizing positive and negative class distributions (Ko, 2015). Ko indicated that Log_TF-TRR demonstrated better performance than other supervised schemes (TF-CHI, TF-RF) in addition to TF-IDF on the Reuters-21578 text dataset, 20-Newsgroups dataset, and Korean UseNet dataset.

Altınçay et al. did an analysis on the performance differences of six commonly-used term weighting schemes which were introduced for text classification before (Altınçay & Erenel, 2010). They concluded that the ratios, which were used by each term weighting scheme, and term occurrence probabilities which were calculated for terms' weights, caused these relative performance differences for term weighting schemes. In another study, they (Erenel & Altınçay, 2012) proposed a new collection frequency factor which were derived from term frequencies. They indicated that their proposed scheme has better performance on classification tasks having small term frequencies. They also emphasized that the optimal form of the term frequency factor depends upon the distribution of term frequencies on used text collection. Badawi and Altınçay proposed a termset-based-representation method for binary text classification (Badawi & Altınçay, 2014). They tried to testify the effectiveness of their method using bag-of-words approach on three datasets. Then, they demonstrated that proposed method generates more effective documents vectors for text classification. Corresponding researchers introduced a new termset weighting scheme (Badawi & Altınçay, 2017) based on cardinality statistics. In this study, they employed two available collection frequency factors with BOW representing approach with weight termset. They indicated that using n-termsets with BOW approach can improve the classification performance of term weighting methods weighted with pure BOW approach.

Ren and Sohrab introduced TF-IDF-ICSDF and TF-IDF-ICF schemes based on document-class information of terms (Ren & Sohrab, 2013). They tested proposed schemes by comparing them with 5 different term weighting schemes (TF-PB, TF-IDF, TF-RF, TF-OR and TF-CC) on the Reuters-21578, RCV1-v2, and 20 Newsgroups datasets using Centroid, NB and SVM classifiers. They showed that especially TF-IDF-$ICS_{\&}$ F scheme generally has better classification results compared to other five term weighting schemes.

Escalante et al. introduced a new learning method to maximize the success of term weighting methods by genetic programming (Escalante et al., 2015). They revealed that main mission of genetic program is to learn which basic units are combined for creating more discriminative term weighting schemes. They demonstrated that the classification results of term weighting methods produced with genetic program was superior than traditional and popular term weighting schemes. Chen et al. introduced two novel term weighting schemes, TF-IGM and SRTF-IGM, derived from a model in statistics named Inverse Gravity Moment (IGM) (Chen et al., 2016). They expressed that IGM measured distinguishing power of a term more effectively than popular collection frequency factors (RF, CHI2, Prob, ICSDF, IDF) in literature. They showed the superiority of TF-IGM over other popular term weighting schemes using SVM and KNN classifiers on the benchmark text datasets (Reuters-21578, 20-Newsgroups, and TanCorp).

Sabbah et al. introduced a hybridized term weighting approach to accurately detect terrorism activities in textual contents (Sabbah, Selamat, Selamat, Ibrahim, & Fujita, 2016). The main function of the proposed hybridized approach is to create a feature set by combining small feature subsets obtained from term weighting schemes named TF, IDF, TF-IDF, Glasgow

and Entropy. With acquired experimental results, they emphasized that the overall success of text classification could be enhanced with use of little mini feature sets that combined most substantial terms in the text. Alsmadi and Hoon introduced SW scheme for weighting short texts obtained from Twitter collections (Alsmadi & Hoon, 2019). They tested SW weighting method using a decision tree, support vector machine, logistic regression, and k-nearest neighbor algorithms on the Sanders and self-collected tweet datasets. They stated that SW is more successful than traditional weighting methods (Binary, TF-IDF, TF-RF, TF, TF-IDF-ICSDF, Delta, CHI2 and TF-IG) to generate weighting solutions for short texts. Rao et al. proposed LGT scheme for first story detection which determines first story of any event-based problem about information analysis and security (Rao et al., 2017). They tested LGT collectively modelling the topical association, global and local elements of each story. They showed that LGT has better performance than fundamental schemes.

Feng et al. introduced lrp scheme based upon a probabilistic model (Feng et al., 2018). They stated the proposed scheme assigns very closer values to zero if term is not discriminative enough. They showed that proposed method can be effectively used on large and small text collections. Li et al. proposed CEW weighting scheme which is a combined form of BCD, log and, entropy weighting (EW) word weights (Li, Zhang, Li, Ouyang, & Cai, 2018). In this study, CEW was employed to award informative terms while other word weights were used to punish common terms which mostly occur. They did experiments to see the performance of CEW scheme on document clustering and topic quality. They indicated that CEW outperformed existing term weighting schemes on the standard text collections including short texts.

We can summarize the rest of the works about term weighting as follows: Deisy et al. introduced MIDF scheme based on Modified Inverse Document Frequency factor for text categorization (Deisy et al., 2010). Emmanuel et al. proposed positive impact factor (PIF) scheme for text categorization (Emmanuel, Khatri, & Babu, 2013). Ropero introduced a novel term weighting scheme based upon Fuzzy-Logic (Ropero, Gómez, Carrasco, & León, 2012). Rashid et al. introduced a fuzzy topic modelling method for short texts (Rashid, Shah, & Irtaza, 2019). In another study, several term weighting approaches (Luo, Chen, & Xiong, 2011; Wei, Feng, He, & Fu, 2011) were presented for semantic analysis while some term weighting studies (Abdel Fattah, 2015; Deng, Luo, & Yu, 2014) proposed for sentiment analysis. Besides, there exist some studies (Ke, 2014; Santhanakumar, Columbus, & Jayapriya, 2018) that proposed for Information Retrieval domain. Lastly, some of the presented studies about term weighting are oriented to specific classification algorithms such as naïve Bayes (Jiang, Li, Wang, & Zhang, 2016; Kim & Kim, 2016; Zhang, Jiang, Li, & Kong, 2016), centroid-based-classifiers (Lertnattee & Theeramunkong, 2004; Nguyen, Chang, & Hui, 2013), and SVM (Haddoud, Mokhtari, Lecroq, & Abdeddaïm, 2016).

3. Term weighting schemes

Several term weighting schemes have been proposed for text classification so far. In this section, five state-of-the-art popular term weighting schemes which are executed in experiments are explained in details. TF-IDF, TF-IDF-ICF, TF-RF, TF-IDF-ICSDF, and TF-IGM term weighting schemes are utilized as baseline methods to compare performances of proposed two schemes.

3.1. TF-IDF

TF-IDF weighting is based on the calculation of Term Frequency (TF) and Inverse Document Frequency (IDF) values of a specific term (Sparck Jones, 2004). While TF is the occurrence of related term in a specific document, IDF shows the value of collection frequency factor of related term for whole data collection. The TF-IDF weighting formula of term t_i is shown in the following equation.

$$W_{TF.IDF}(t_i) = TF(t_i, d_k) * log\left(\frac{D}{d(t_i)}\right) \tag{1}$$

In Eq. (1), $TF(t_i\ d_k)$ shows the quantity of occurrence of term t_i in text document d_k, $d(t_i)$ is the quantity of text documents in which t_i occurs and D shows the total quantity of text documents in whole dataset. TF-IDF assigns high IDF scores to infrequent terms in data collection. Since it does not use class information of terms in weighting process, TF-IDF is known as an unsupervised term weighting scheme in text classification domain.

3.2. TF-RF

TF-RF assigns weighting scores to terms according to binary approach. In other words, it calculates the inter-class-distributions of terms for each class by using probabilities of positive and negative classes (Lan et al., 2009). Calculated inter-class-distribution values are used to obtain the relevance frequencies (RF) of terms. The authors state that high frequency terms, which occur more in positive class than in negative class, should have higher RF values in weighting process because they are more effective for discriminating positive samples from the negative ones. The TF-RF weighting formula of t_i term is given in Eq. (2).

$$W_{TF.RF}(t_i) = TF(t_i, d_k) * \max_{j=1}^{C}\left\{\log\left(2 + \frac{a_{ij}}{\max(1, c_{ij})}\right)\right\} \tag{2}$$

In this equation, a_{ij} and c_{ij} are the quantities of text documents in positive class (regarding to class C_j) and negative class (not regarding to class C_j) that contain term t_i, respectively. C is the entire number of classes in the whole dataset and *max* expression shows that the maximum of class-based weight values are assigned to term t_i. Because of using class information of term, TF-RF is mentioned as supervised term weighting scheme in text classification domain.

3.3. TF-IDF-ICF

The success of term weighting schemes (such as TF-RF) utilized class information of terms in their term weighting process has recently attracted the researchers' attention to developing new supervised schemes. TF-IDF-ICF is also one of the proposed schemes in the light of these thoughts. In weighting strategy of this scheme, Inverse Class Frequency information (ICF) of term is also used in addition to TF-IDF information (Ren & Sohrab, 2013). ICF focuses on calculating the ratio of the quantity of classes and total quantity of classes that contain term t_i. Unlike TF-RF, TF-IDF-ICF generates one weighting score for every term, hence it does not use any globalization policy in weighting. The TF-IDF-ICF weighting stage is demonstrated with following equation.

$$W_{TF.IDF.ICF}(t_i) = TF(t_i, d_k) * \left(1 + log\left(\frac{D}{d(t_i)}\right)\right) * \left(1 + log\left(\frac{C}{c(t_i)}\right)\right) \tag{3}$$

In this equation, $c(t_i)$ expresses the quantity of the classes where term t_i occurs, C also expresses the total quantity of classes in entire dataset. As in the TF-IDF term weighting scheme, TF-IDF-ICF also assigns high weighting scores to infrequent words, which occur less in all text documents or classes in whole dataset.

3.4. TF-IDF-ICSDF

This scheme is one of the other proposed schemes in Ren and Sohrab's study in which Inverse Class Space Density Frequency (ICSDF) is added as second collection frequency factor to TF-IDF information of terms (Ren & Sohrab, 2013). Unlike ICF, ICSDF considers also distributions of inter-class documents when calculating weighting values of each term. The weighting value of term t_i based upon TF-IDF-ICSDF term weighting scheme is calculated as in the following equation.

$$W_{TF.IDF.ICSDF}(t_i) = TF(t_i, d_k) * \left(1 + log\left(\frac{D}{d(t_i)}\right)\right) * \left(1 + log\left(\frac{C}{\sum_{j=1}^{C} \frac{df_{t_{ij}}}{D_j}}\right)\right) \tag{4}$$

In this equation, $df_{t_{ij}}$ is class-specific DF values of term t_i in class j and D_j is the quantity of text documents in class j. *TF-IDF-ICSDF* has also a supervised learning characteristic because of using class information of terms and it generates one weighting score for each term in the weighting stages of them.

3.5. TF-IGM

TF-IGM is one of the lately proposed supervised term weighting schemes and it combines TF and Inverse Gravity Moment (IGM) information of terms as collection frequency factor. IGM is a statistical model and adapted for term weighting (Chen et al., 2016). The IGM method calculates inter-class-distributions of terms by concentrating on their document frequencies when trying to find out their distinguishing powers. The TF-IGM weighting calculation is presented in Eq. (5).

$$W_{TF.IGM}(t_i) = TF(t_i, d_k) * \left(1 + \lambda * \frac{f_{i1}}{\sum_{r=1}^{C} f_{ir} * r}\right) \tag{5}$$

In this equation, f_{ir} expresses class-specific document frequencies of term t_i. In IGM weighting strategy, related document-frequencies are sorted in descending order and multiplied by rank values expressed with r (r = 1,2,3,. . .,C). Hence f_{ir} shows the document frequency of class in which term t_i occurs most. This equation has also λ balance parameter whose value range is defined as 5.0–9.0. The assumed value of λ is also identified as 7.0 in the referenced paper.

4. The proposed novel weighting schemes

In this section, we propose a novel term weighting strategy called MONO to effectively reflect terms' distinguishing abilities to vector space model. Two novel term weighting schemes named TF-MONO and SRTF_MONO are proposed by

combining TF and Squared Root of TF (SRTF) term frequency factors with proposed MONO term weighting strategy. Subsection 4.1 includes explanation about general term weighting problems that existing popular weighting approaches have difficulty to deal with and main motivation of the study while the proposed term weighting strategy and schemes are explained in subsection 4.2. Lastly, the superiorities of the proposed MONO term weighting approach over the existing popular term weighting approaches are given in subsection 4.3.

4.1. General observations and motivations

In text classification, one of the important problems to which researchers try to find various solutions is to represent document vectors effectively. For this aim, it is crucial to assign proper weights to terms when determining the relations between documents and their contents (terms). Therefore, an effective term weighting scheme should assign reasonable weights that can represent terms according to their class distinguishing abilities. There exist too many term weighting schemes in literature but it is difficult to say that these schemes have generated ideal weights reflecting terms' real distinguishing abilities for all cases. For example, unsupervised schemes such as TF-IDF do not use class information of terms; therefore, it does not fully reflect terms' class distinguishing power on classification space. Hence the classification performance of TF-IDF is generally limited in text classification tasks. Supervised weighting schemes in literature show that the classification successes of term weighting schemes could be improved with suitable use of class information of terms. The success of any term weighting schemes generating only one weight for each term also shows that the weighting strategies of terms should contain all-classes rather than positive/negative class in binary approaches.

However, the traditional or popular supervised weighting methods in literature mostly focus on assigning weights to terms according to their occurrence information in classes. For example, *TF-RF* uses the ratio of the quantity of documents belonging to positive class and negative class in which term t_i occurs while *TF-IGM* uses the quantity of documents of all classes in which term t_i occurs. Sorting class specific document frequencies (DF) of terms and focusing DFs of class (or classes) where terms occur most also provide more successful multi-class classification performance than other schemes for *TF-IGM*. Similarly, this class/document information utilized in other ways in *TF-IDF-ICF*, *TF-IDF*, *TF-IDF-ICSDF* and other several term weighting schemes. Utilizing class information is an effective path; however, this may not be sufficient to show the distinguishing power of terms. Indeed, the non-occurrence information of terms can also be used together with occurrence information of terms to reflect terms' real distinguishing abilities. It can be explained with the following simple example:

Assume that having two terms called as t_1 and t_2 occurring in documents belonging to two classes in a small text collection. Assume that term t_1 occurs in all documents regarding the first class and it occurs in few documents regarding the second class. Then, it can be said that the class distinguishing ability of t_1 is high and it must be assigned with high weight value in the term weighting period. In case it is assumed that the term t_2 occurs in all documents belonging to the first class and it does not occur in any documents belonging to the second class. Then, it seems that the class distinguishing ability of t_2 is higher than t_1. The weight value of t_2 must be higher than the weight value of t_1 because t_2 supplies better class discrimination for class-1.

In the light of literature review and above-mentioned information, it can be stated that an effective term weighting method should have the following characteristics:

- It should effectively use class information of terms.
- It should reflect the distinguishing power of terms as well as possible.
- It should especially focus on the occurrence information of terms in class they occurred most.
- It should effectively use the non-occurrence information as well as occurrence of terms on the supervised term weighting.
- It should generate weights for terms reflecting the distinctiveness information of them over all classes.

4.2. Proposed MONO weighting strategy and schemes

We developed a new term weighting strategy involving above-mentioned approaches for text classification. Assume that there exist j classes in a text collection and document frequencies of t_i are shown as follows:

$$df_{t_i} = \left\{ d_{i1}, d_{i2}, d_{i3}, \ldots, d_{ij-1}, d_{ij} \right\} \tag{6}$$

Proposed weighting approach consists of following stages on the weighting process of the term t_i respectively:

Firstly, class-specific document frequencies of term t_i are sorted in ascending order. Assume that the sorted list of document frequencies is as below.

$$sorted\_df_{t_i} \left\{ d_{i3}, d_{i1,} d_{i4}, \ldots, d_{ij}, d_{ij-1} \right\} \tag{7}$$

The class-specific *DF*s are divided into following groups as the class where t_i occurs most and rest classes.

$$sorted\_df_{t_i} = \left\{ \overbrace{d_{i3}}^{C_{t_i\_max}} \mid, \overbrace{\overline{d_{i1}}, \overline{d_{i4}}, \ldots, \overline{d_{ij}}, \overline{d_{ij-1}}}^{C_{others}} \right\} \tag{8}$$

The first group is represented with *max-occurrence* (*MO*) ratio while second group is represented with *non-occurrence* (*NO*) ratio as follows.

$$sorted\_df_{t_i} = \left\{ \overbrace{d_{i3}}^{MO} \mid, \overbrace{\overline{d_{i1}}, \overline{d_{i4}}, \ldots, \overline{d_{ij}}, \overline{d_{ij-1}}}^{NO} \right\} \tag{9}$$

MO_{t_i} value, representing the ratio between the quantity of text documents in class where t_i occurs most and total quantity of text documents in the corresponding class, is calculated.

$$MO_{t_i} = \frac{D_{t_i\_max}}{D_{total}\left(t_i\_max\right)} \tag{10}$$

After calculating MO_{t_i} value, NO_{t_i} value is calculated. NO_{t_i} value represents the ratio between the quantity of text documents in the rest of the classes where t_i does not occur and total quantity of text documents in the rest of the classes.

$$NO_{t_i} = \frac{D_{\overline{t_i}}}{D_{total}\left(\overline{t_i}\right)} \tag{11}$$

Then, the product of *MO* and *NO* ratios are calculated. The result of the product operation is assigned as the local *MONO* weight of term t_i.

$$MONO_{Local}\left(t_i\right) = \overbrace{\left[\frac{D_{t_i\_max}}{D_{total}\left(t_i\_max\right)}\right]}^{MO_{t_i}} * \overbrace{\left[\frac{D_{\overline{t_i}}}{D_{total}\left(\overline{t_i}\right)}\right]}^{NO_{t_i}} \tag{12}$$

Finally, global *MONO* weight of term t_i is calculated as following:

$$MONO_{Global}\left(t_i\right) = \left[1 + \alpha * MONO_{Local}\left(t_i\right)\right] \tag{13}$$

In this strategy, α is a balance parameter to set the ranges of global weights values in weighting period. Its value ranges and default value are defined as 5.0–9.0 and 7.0, respectively.

In this work, two novel term weighting schemes based upon $MONO_{Global}$ collection frequency factor are proposed. These two schemes are called as TF-MONO and SRTF-MONO. The weighting formulas of them are shown in the following equations.

$$TF - MONO = TF\left(t_i, d_k\right) * \left[MONO_{Global}\left(t_i\right)\right] \tag{14}$$

$$SRTF - MONO = SRTF\left(t_i, d_k\right) * \left[MONO_{Global}\left(t_i\right)\right] \tag{15}$$

In Eq. (15), $SRTF(t_i,d_k)$ is squared root forms of raw term frequency (*TF*) values of term t_i in document d_k.

4.3. Weighting specifications and advantages of proposed MONO strategy

In this subsection, the weighting strategy of the proposed MONO weighting is shown with several examples having different document distributions. Assume that document frequencies of these terms (t_1, t_2, and t_3) are [100,0,0] and [100,75,0] and [100,100,100] respectively in a dataset consisting of three classes where each class has 100 documents. The local weights of t_1, t_2 and t_3 are calculated by MONO strategy as follows.

$$MO\left(t_1\right) = \frac{100}{100} = 1 \quad NO\left(t_1\right) = \frac{200}{200} = 1 \quad MONO_{Local}\left(t_1\right) = \frac{100}{100} * \frac{200}{200} = 1$$

$$MO\left(t_2\right) = \frac{100}{100} = 1 \quad NO\left(t_2\right) = \frac{125}{200} = 0.63 \quad MONO_{Local}\left(t_2\right) = \frac{100}{100} * \frac{125}{200} = 0.63$$

$$MO\left(t_3\right) = \frac{100}{100} = 1 \quad NO\left(t_3\right) = \frac{0}{200} = 0 \quad MONO_{Local}\left(t_3\right) = \frac{100}{100} * \frac{0}{200} = 0$$

As seen in the above mentioned calculations, $MONO_{Local}$ generates weight scores ranging between 0 and 1. If term occurs in entire documents in a particular class and not occurs in the other classes like t_1, $MONO_{Local}$ assigns 1 to it. If term occurs in all documents in all classes like t_3, $MONO_{Local}$ assigns 0 to it. Assuming α = 7, global MONO weights for t_1, t_2 and t_3

Table 1
Local MONO weighting for different MO and NO levels.

	Terms	# of classes /# of doc.	Document Frequencies	Sorted according to Distinctiveness (Intuitively)	MO values	NO values	Local MONO values
	t_1		[100, 0, 0, 0, 0]		1	1	1
	t_2		[78, 0, 0, 0, 0]		0.78	1	0.78
Scenario 1	t_3	5 / 100	[57, 0, 0, 0, 0]	$t_1 > t_2 > t_3 > t_4 > t_5$	0.57	1	0.57
	t_4		[31, 0, 0, 0, 0]		0.31	1	0.31
	t_5		[13, 0, 0, 0, 0]		0.13	1	0.13
	t_6		[1000, 1000, 0, 0]		1	0.667	0.667
	t_7		[1000, 760, 0, 0]		1	0.747	0.747
Scenario 2	t_8	4 / 1000	[1000, 470, 0, 0]	$t_{10} > t_9 > t_8 > t_7 > t_6$	1	0.843	0.843
	t_9		[1000, 235, 0, 0]		1	0.922	0.922
	t_{10}		[1000, 2, 0, 0]		1	0.9993	0.9993

are calculated as 8, 5.41, and 1 in the proposed weighting strategy. Although the local MONO weight value of one term is assigned as 0, the corresponding term can be represented with its TF value in vector space model in global MONO weighting strategy.

Table 1 shows that how the proposed MONO strategy performs weighting for terms having different MO and NO levels in two different scenarios. Assume that t_1, t_2, t_3, t_4, and t_5 terms are extracted from a dataset which has 5 classes and each class has 100 documents (Scenario 1). However, terms t_6, t_7, t_8, t_9, and term t_{10} belong to another dataset having 4 classes and 4000 documents (Scenario 2).

While changes in the MO levels are illustrated with scenario 1 when NO values are fixed, changes in the NO levels are shown with scenario 2 when MO values are fixed. It should be noted that proposed MONO weighting approach takes into account not only the information that term occurs in how many class (or classes), but also occurrence of corresponding term in number of documents (document-scaling) in relevant class (or classes). We can explain the advantage of document-scaling by comparing with one of the recent methods named TF-IGM term weighting. IGM method assigns same local values to terms t_1- t_5 (Eq. (5)) since all of them occur only in one class, but it is implicitly seen that the intuitive class-distinguishing abilities of these terms are not equal ($t_1 > t_2 > t_3 > t_4 > t_5$). MONO assigns 1, 0.78, 0.57, 0.31, 0.13 scores to these terms and these scores supply better representation in terms of their class distinguishing potentials. Also, another advantage of MONO weighting strategy is the usage of terms' non-occurrence information on the weighting calculation process. This scaling provides more reasonable representations for most-distinctive and distinctive terms. From this point of view, most-distinctive term can be expressed as the term occurring frequently in one class and not occurring in other classes. Distinctive term can be expressed as the term occurring in one class and occurring rarely in other classes. If we explain it by giving an example from Scenario 2 in Table 1, t_{10} is more distinctive than t_6 because t_6 occurs in all documents of first two classes and t_{10} occurs only in two documents of the second class.

Let's explain other advantages of MONO weighting strategy by comparing it with term weighting schemes explained in Section 3 with another scenario. In scenario 3, assume that document frequencies of these terms (t_{11}, t_{12} and t_{13}) are [23, 15, 48, 19, 39] and [48, 0, 96, 0, 0] and [0, 1, 0, 2, 0] respectively in a dataset including five classes where each class has 100 documents. Since total number of text documents in which t_{11} and t_{12} occur are equal, IDF, ICSDF and IDF-ICSDF assign same weighting scores to them. But the class-distinguishing ability of t_{12} is higher than t_{11} because t_{11} occurs in all classes while t_{12} occurs only in two classes. By MONO weighting strategy, local weights of t_{11} and t_{12} are 0.36 and 0.84, respectively. Moreover, by setting α as 7, they will be represented with 3.52 and 6.88 global MONO weights in VSM, respectively. It can be stated that MONO provides better representations in comparison to TF-IDF, TF-ICSDF and TF-IDF-ICSDF weighting. For another example from scenario 3, t_{12} and t_{13} has equal RF_{max} and ICF scores, but it is obvious that they do not have the same class distinguishing abilities. For ICF weighting, the reason of this equality is that both terms occur in two classes. RF_{max} also represents t_{12} and t_{13} with same weights values because the ratios of a_{ij} / c_{ij} in Eq. (2) for each one are equal for both t_{12} and t_{13}. It can be stated that t_{12} has higher class distinguishing potential than t_{13} since it occurs in more documents in same number of classes. If corresponding terms are weighted by MONO, the local weights of them will be 0.84 and 0.02. So, they are represented with 6.88 and 1.14 scores by $MONO_{Global}$ in VSM. In conclusion, MONO provides more reasonable representation in VSM while weighting these terms. On the other hand, it is possible to give more examples for problematic weighting scenarios as above-mentioned ones.

5. Experimental study

In this section, experimental information about used datasets, pre-processing, feature selection, term weighting, evaluation methods, and classifiers are presented. However, multiclass-classification results obtained from seven term weighting schemes and discussions are also given in this section.

Table 2
Reuters-21578 text dataset.

No	Class Label	Training Samples	Testing Samples
1	earn	2840	1083
2	acq	1596	696
3	money-fx	206	87
4	grain	41	10
5	crude	253	121
6	trade	251	117
7	interest	190	75
8	ship	108	36

Table 3
20 Newsgroups dataset.

No	Class Label	Training Samples	Testing Samples
1	alt.atheism	500	500
2	comp.graphics	500	500
3	comp.os.ms-windows.misc	500	500
4	comp.sys.ibm.pc.hardware	500	500
5	comp.sys.mac.hardware	500	500
6	comp.windows.x	500	500
7	misc.forsale	500	500
8	rec.autos	500	500
9	rec.motorcycles	500	500
10	rec.sport.baseball	500	500
11	rec.sport.hockey	500	500
12	sci.crypt	500	500
13	sci.electronics	500	500
14	sci.med	500	500
15	sci.space	500	500
16	soc.religion.christian	500	497
17	talk.politics.guns	500	500
18	talk.politics.mideast	500	500
19	talk.politics.misc	500	500
20	talk.religion.misc	500	500

5.1. Datasets and preprocessing

In the experimental side, three different benchmark text datasets named Reuters-21578, 20-Newsgroups, and WebKB are used to evaluate the classification performances of all term weighting schemes. The detailed information about processed datasets are presented in the following subsections.

5.1.1. Reuters-21578

It is one of the most popular unbalanced text dataset which is commonly preferred for text classification. We used Reuters-ModApte split (Asuncion & Newman, 1994) which has top-10 classes belonging to Reuters-21578 and consists of single and multi-labelled text documents. Since multi-labelled documents are removed on the reading texts, 'wheat' and 'corn' classes become empty then these classes are deleted. Class and document distributions about Reuters-21578 dataset are given in the following Table 2. We have not executed any train and test segmentation on the experiment period since Reuters-ModApte split has already own train and test document splits.

5.1.2. 20-Newsgroups

It is also one of the widely used balanced text dataset which contains total 19,997 documents from twenty classes (Asuncion & Newman, 1994). Except one class, the all classes have 1000 documents in the 20-Newsgroups dataset. In the experimental period, almost fifty-fifty separation is done for train and text processes. Table 3 consists of information about 20-Newsgroups dataset.

5.1.3. WebKB

This dataset is another unbalanced benchmark dataset used for text classification tasks. It contains web pages from 7 different classes which are collected from Computer Science departments of four different universities (Craven, McCallum, PiPasquo, Mitchell, & Freitag, 1998). In the experimental side, four-class subset, which has total 4199 documents, is used by separating training (67 %) and test splits (33 %). Detailed information about WebKB is shown in Table 4.

In the pre-processing step, lowercase conversion, alphabetic tokenization, removing the stop-words and stemming pre-processing methods are applied to document contents obtained from above mentioned three datasets. Moreover, the terms

Table 4
WebKB dataset.

No	Class Label	Training Samples	Testing Samples
1	course	620	310
2	faculty	750	374
3	project	336	168
4	student	1097	544

that occurred only once in each dataset are removed from extracted features. The final numbers of unique features extracted from Reuters-21578, 20-Newsgroups and WebKB are 9237, 39,050 and 7181, respectively.

5.2. Feature selection and weighting specifications

We used Chi-Square (CHI2) feature selection method (Chen et al., 2016) to test the performances of proposed schemes and other schemes in different feature sizes. The features are sorted in the descending order with respect to their $CHI2_{max}$ scores and these sorted feature subsets are used during feature selection stage. Top-(500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000 and 9237) and top-(500, 1000, 2000, 4000, 6000, 8000, 10000, 12000, 14,000 and 16,000) features are selected for Reuters-21578 and 20-Newsgroups datasets respectively while top-(100, 500, 1000, 2000, 3000, 4000, 5000, 6000 and 7000) are preferred for WebKB dataset.

In term weighting process, the experiments are carried out by using TF-IDF, TF-RF, TF-IDF-ICF, TF-IDF-ICSDF, and TF-IGM term weighting schemes as well as proposed TF-MONO and SRTF-MONO schemes. For TF-RF, the maximum globalization function named $TF\text{-}RF_{max}$ is used for globalizing class-based RF weights of each term. For TF-IGM and TF-MONO, α and λ coefficients are set to 6.0 in experiments executed on Reuters-21578 and WebKB datasets, while related coefficients are set to 7.0 in experiments executed on 20 Newsgroups dataset.

5.3. Classifiers and success measures

Well-known Support Vector Machines (SVM) and K-Nearest Neighbor (KNN) classification algorithms are utilized for classification process of text documents obtained from three datasets. The detailed information about these classifiers and their parameters are given in the following subsections.

5.3.1. SVM

It is the one of the most preferred classification algorithm used in pattern recognition and text classification research areas (Sabbah et al., 2017). One of the reasons of high popularity of this classifier is that it can work effectively on classifying high dimensional feature vectors, too. The learning process of SVM classifier is based upon constituting a linear or non-linear hyper-plane to distinguish positive samples from negative samples. This hyper-plane located at the position maximizing distance between positive and negative samples by means of some samples on the training set. This concept named margin-maximization is very important for reflecting the classifier's characteristics of SVM.

On the experimental side, LibSVM package (Chang & Lin, 2011) which supports multi-class classification is used with default parameters. So, SVM classifier with linear kernel is applied in the experiments.

5.3.2. KNN

KNN has a simple learning algorithm and it is commonly used for classification tasks in literature (Sabbah et al., 2016). The learning algorithm of KNN is based on estimation of class label of any test document according to K-nearest neighbor of training documents. The learning algorithm aims to find similar training documents to related text document by measuring closeness. KNN assigns related test document to the class, which is the predominant one among the classes of these K-nearest documents. Distance between test document and training documents can measured by using several distance measures in KNN learning (Prasath, Alfeilat, Lasassmeh, & Hassanat, 2017).

On the experimental side, KNN classifier is performed by setting k value as 15, 15, and 11 for Reuters-21578, 20-Newsgroups and WebKB datasets, respectively. However, cosine similarity metric is used for finding similarity between text documents in the experiments.

For extensive evaluation of the classification successes of performed term weighting schemes, micro-averaged ($micro\text{-}F_1$) and macro-averaged ($macro\text{-}F_1$) success measures are preferred in the evaluation stage. These metrics are compatible for measuring multi-class classification performances and consist of *Precision* and *Recall* success metrics which are commonly used in binary classification. Assuming with FP, TP, and FN shows false positive, true positive, and false negative values in confusion matrix, $micro\text{-}F_1$ and $macro\text{-}F_1$ values are calculated according to the following equations.

$$Precision_{c_k} = \frac{TP_{c_k}}{TP_{c_k} + FP_{c_k}} \quad Recall_{c_k} = \frac{TP_{c_k}}{TP_{c_k} + FN_{c_k}} \quad F1_{c_k} = \frac{2 * Precision_{c_k} * Recall_{c_k}}{Precision_{c_k} + Recall_{c_k}} \tag{16}$$

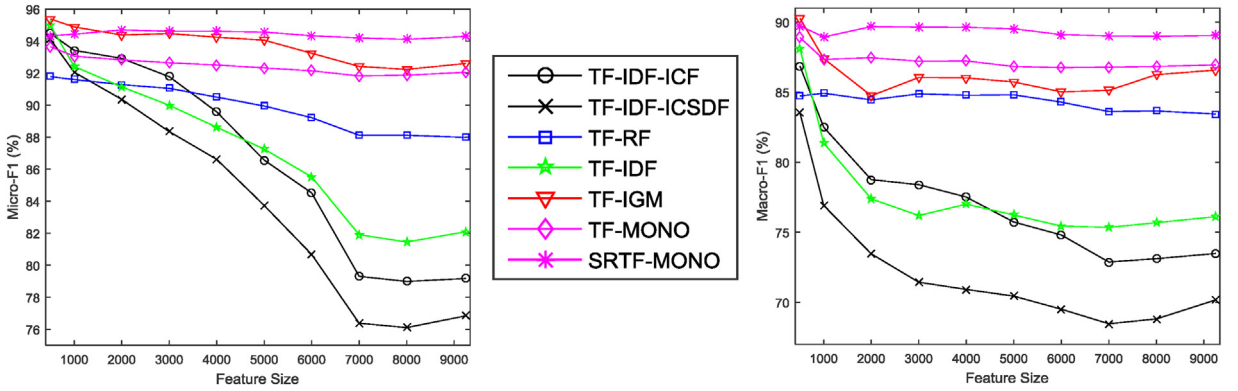

Fig. 1. Performance results acquired by 7 different term weighting schemes upon KNN (k = 15) classifier, and Reuters-21578 dataset.

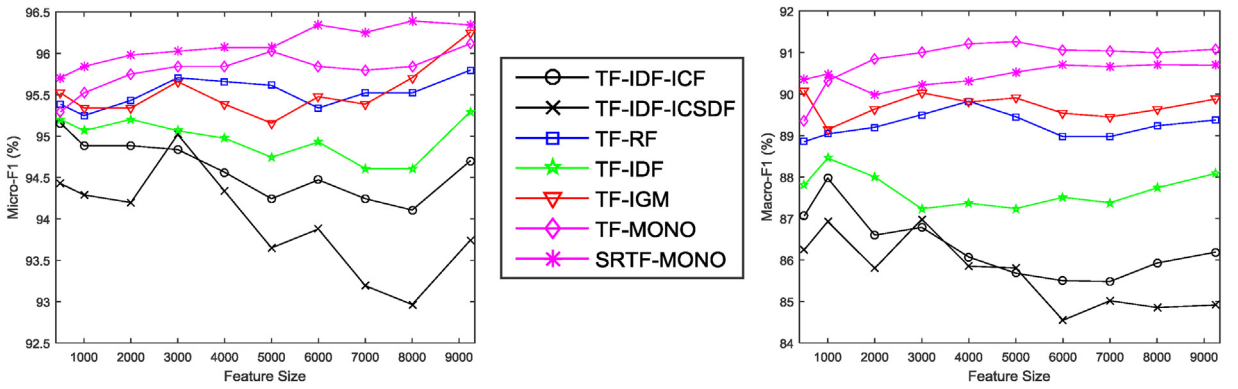

Fig. 2. Performance results acquired by 7 different term weighting schemes upon linear SVM classifier, and Reuters-21578 dataset.

$$Micro-F1 = \frac{2 * \sum_{k=1}^{C} TP_{c_k}}{2 * \sum_{k=1}^{C} TP_{c_k} + \sum_{k=1}^{C} FP_{c_k} + \sum_{k=1}^{C} FN_{c_k}} \quad Macro-F1 = \frac{1}{C}\sum_{k=1}^{C} F1_{c_k} \tag{17}$$

5.4. Multi-class classification results

The performance analysis of the proposed TF-MONO and SRTF-MONO and other executed popular five term weighting schemes are presented by using KNN and SVM classifiers upon the dataset named Reuters-21578, 20-Newsgroups and WebKB. The experiments are performed with several different feature sizes for more comprehensive success analysis.

Figs. 1 and 2 show the performances of classification for seven term weighting schemes obtained from KNN and SVM classifiers upon the Reuters-21578 dataset.

If the classification results which are obtained from KNN classifier upon the Reuters-21578 dataset are investigated, it can be seen that proposed SRTF-MONO scheme outperformed other all term weighting schemes according to both micro-F_1 and macro-F_1 measures. However, proposed TF-MONO has also better macro-F_1 results than popular TF-IGM and other four term weighting schemes for corresponding classifier and dataset. Moreover, the micro-F_1 and macro-F_1 performances of schemes based on IDF factor are generally decreasing while the feature sizes are increasing up to 7000 features in the experiments on KNN classifier. Besides, IDF based schemes have lower success values than other schemes while the lowest classification performance is obtained with TF-IDF-ICSDF scheme.

For SVM classifier, it is obviously seen that all proposed schemes have better both micro-F_1 and macro-F_1 values than other five term weighting schemes upon the Reuters-21578 dataset. The best micro-F_1 and macro-F_1 performance is obtained with SRTF-MONO and TF-MONO with SVM classifier upon Reuters-21578 dataset, respectively. Even TF-IGM and TF-RF showed closer classification performances, the macro- F_1 scores of TF-IGM is higher than TF-RF. IDF based schemes also showed lower performance than other schemes for SVM classifier while TF-IDF has better micro-F_1 and macro-F_1 scores among three IDF-based schemes.

If we perform a classifier-based comparison on the Reuters-21578 dataset, SVM has better micro-F_1 and macro-F_1 results than KNN classifier. However, it can be said that SVM is more successful for classifying documents represented with feature vectors of higher dimension. Lastly, TF-IDF-ICF scheme has better classification performance than TF-IDF up to 4000 features on the KNN classifier and TF-IDF is more successful than TF-IDF-ICF scheme in all feature sizes for SVM classifier.

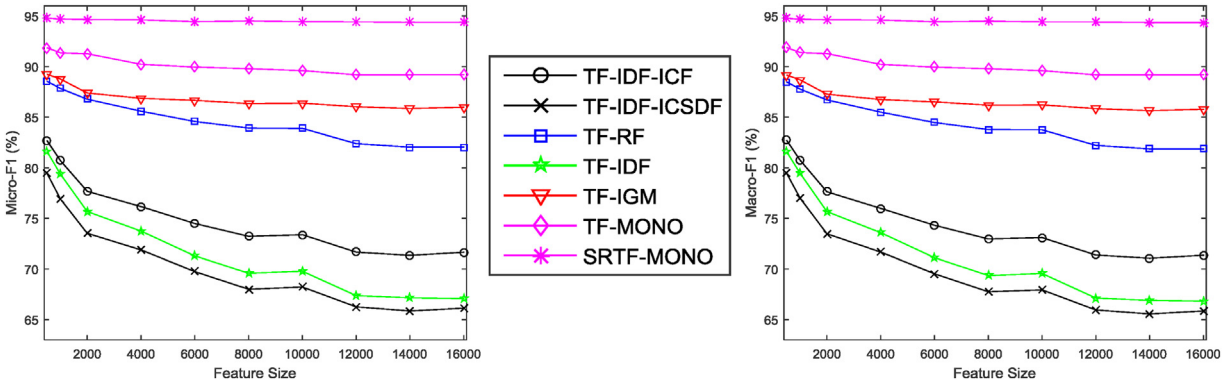

Fig. 3. Performance results acquired by 7 different term weighting schemes upon KNN (k = 15) classifier, and 20-Newsgroups dataset.

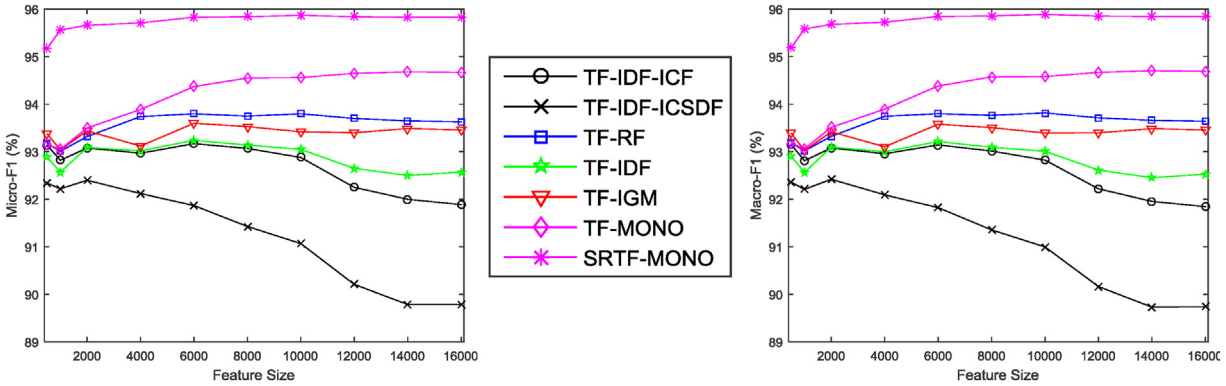

Fig. 4. Performance results acquired by 7 different term weighting schemes upon SVM classifier, and 20-Newsgroups dataset.

Figs. 3 and 4 show the performances of classification for seven term weighting schemes obtained from KNN and SVM classifiers on the 20-Newsgroups dataset.

For KNN classifier, the proposed SRTF-MONO and TF-MONO outperformed other schemes on the 20-Newgroups dataset with respect to both micro-F_1 and macro-F_1 results. The best classification performance is obtained with SRTF-MONO while the performance difference of SRTF-MONO and closest follower scheme named TF-IGM is at about 5 % with KNN classifier on the 20-Newsgroups dataset. General classification performance of schemes for KNN classifier on the 20-Newsgroups can be shown as SRTF-MONO > TF-MONO > TF-IGM > TF-RF > IDF-based schemes. However, lowest performance results are obtained from TF-IDF-ICSDF among seven term weighting schemes in experiments carried out with KNN classifier upon 20-Newsgroups dataset.

Proposed SRTF-MONO and TF-MONO have superior classification performances over other term weighting schemes with SVM classifier too. SRTF-MONO is also the most successful term weighting scheme among seven term weighting schemes compared along with KNN classifier on the 20-Newsgroups dataset. If we sort the general classification successes of seven term weighting schemes on the 20-Newsgroups dataset once again, the order is SRTF-MONO > TF-MONO > TF-RF > TF-IGM > IDF-based schemes, for SVM classifier. The general performance of TF-IDF-ICSDF is lower than other six term weighting schemes and decreases dramatically while the dimensions of feature vectors increase.

If we analyze the performance results on the 20-Newsgroups dataset according to classifiers, the classification performance of seven term weighting schemes obtained from SVM classifier are higher than those obtained from KNN classifier. For instance, while highest micro-F_1 and macro-F_1 values of SRTF-MONO are at about 94–95 % with KNN classifier, related values are at about 95–96 % with SVM classifier. Moreover, the successes of IDF-based schemes on KNN classifier decreases hardly than SVM when dimensions of feature vectors increase.

Figs. 5 and 6 show the performances of classification for seven term weighting schemes obtained from KNN and SVM classifiers on the WebKB dataset.

Concerning KNN classifier results on the WebKB dataset, it seems that proposed SRTF-MONO outperformed other schemes. The performance sorting is demonstrated as SRTF-MONO > TF-RF > TF-IGM > TF-MONO > IDF-based schemes in these experiments. IDF-based schemes have more stable classification performances on WebKB dataset when feature sizes increase. The reason of that may be related to the number of class and document distribution of WebKB. In other words, WebKB has less number of classes and documents than the other datasets.

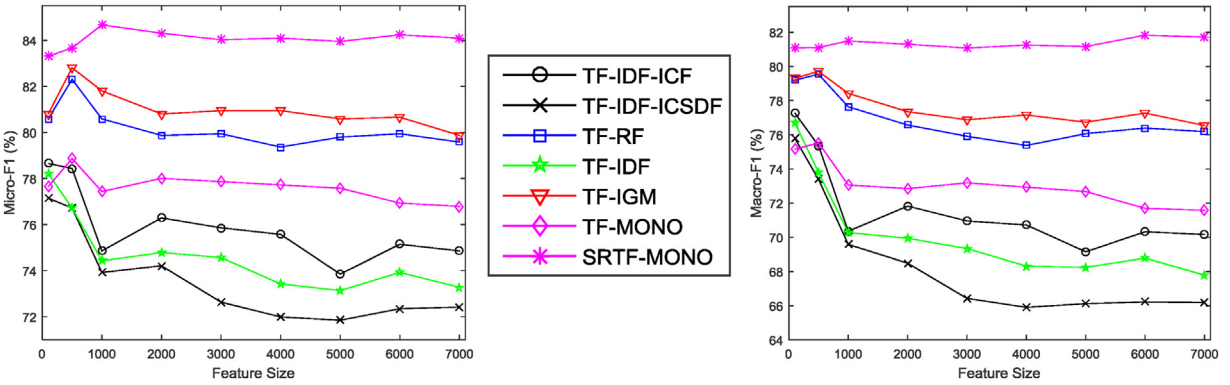

Fig. 5. Performance results acquired by 7 different term weighting schemes upon KNN (k = 11) classifier, and WebKB dataset.

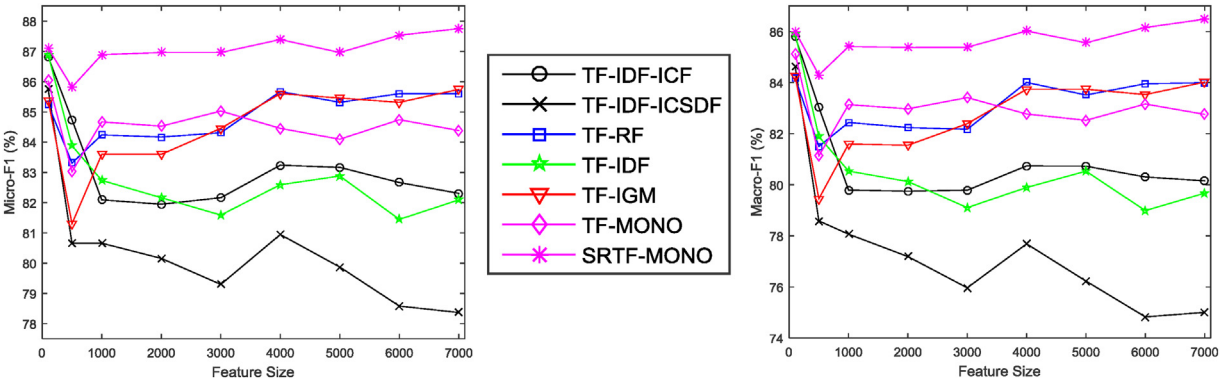

Fig. 6. Performance results acquired by 7 different term weighting schemes upon linear SVM classifier, and WebKB dataset.

The proposed SRTF-MONO has better classification results than other all term weighting schemes in case SVM classifier is used on the WebKB dataset. While the other proposed scheme TF-MONO has higher micro-F_1 and macro-F_1 successes than other five schemes up to 3000 features, the performance of TF-MONO is lower than TF-RF and TF-IGM when feature sizes are more than 3000. With SVM classifier, TF-IGM and TF-RF also showed closest classification successes when dimension of feature vector is higher than 3000 on the WebKB dataset. Also, the lowest classification results are obtained from TF-IDF-ICSDF.

The value range of classification performances obtained from seven term weighting scheme by using KNN classifier is larger than the ones obtained from same schemes by using SVM classifier on WebKB dataset. However, SVM provided more successful micro-F_1 and macro-F_1 values than KNN. For example, the best performing SRTF-MONO scheme has more classification performance on SVM than KNN.

It can be stated that the proposed SRTF-MONO scheme outperformed all other term weighting schemes without discriminating any datasets or classifiers. General classification performance of SRTF-MONO proved that it provides effective representations for text documents, especially the ones expressed with higher feature vectors on both unbalanced and balanced datasets. The other proposed TF-MONO scheme also has better classification performances on the benchmark datasets such as Reuters-21578 and 20-Newsgroups.

5.5. Discussions

Most of term weighting schemes which are proposed for text classification generate weighting solutions for terms depending on only information of class/document distribution in which related terms occurs when trying to discover their class distinguishing abilities. Although it is an effective way to focus on distribution of documents consisting of term in order to calculate their class distinguishing abilities, this information alone may not be adequate for reflecting real distinguishing power of related term in vector space model. In other words, the non-occurrence information should be comprehensibly added to weighting process of terms to achieve better document-term representations (the reason of that is explained with a simple example in subsection 4.1). Weighting terms that utilize both information of occurrence and non-occurrence could provide better representation that show its real class distinguishing power and improve classification performance of term weighting scheme. The proposed MONO weighting strategy combines occurrence and non-occurrence information of a term by using two important ratios named maximum occurrence (MO) and non-occurrence (NO) ratios, respectively. In addition

to this, the proposed MONO strategy has a sorting policy which sorts class specific document frequencies like TF-IGM since it finds discriminate terms and assigs higher weight values to them. The successes of the proposed TF-MONO and especially SRTF-MONO demonstrates that the representations of feature vectors significantly progressed with the proposed MONO strategy which uses non-occurrence information of terms.

6. Conclusions

In this study, we proposed a new term weighting strategy called as MONO and two term weighting schemes named TF-MONO and SRTF-MONO based on MONO. The proposed schemes and MONO strategy provided better discrimination for representing document vectors belonging to different classes by properly adding non-occurrence distributions of terms to their weighting process. The weighting process of terms in MONO is carried out with two important ratios named MO and NO that show the distributions of occurrence in the class where term occurs most and non-occurrence in the rest of the class (or classes). MO ratio tries to reach intra-class document-scaling level of term while NO aims to find their inter-class distinguishing level. The proposed MONO method comprehensibly combines these two important ratios in weighting calculation stages.

The proposed TF-MONO and SRTF-MONO term weighting schemes are compared with 5 popular state-of-the-art term weighting schemes by using SVM and KNN classifiers on the benchmark text datasets namely Reuters-21578, 20-Newsgroups, and WebKB. Experimental results show that SRTF-MONO outperformed all term weighting schemes for both SVM and KNN classifiers on three text datasets. SRTF-MONO is also rather robust and consistent for experiments which are carried out with higher feature sizes. Since the micro-F_1 and macro-F_1 performances of TF-MONO is also more successful than those of other five schemes on the universal text datasets such as 20-Newsgroups and Reuters-21578, it can be stated that TF-MONO is also a promising term weighting scheme for multi-class text classification area. The performance progresses proves that the proposed schemes represent terms better in VSM because they assign more proper weights reflecting their distinguishing powers.

It is obviously seen that the proposed schemes have better performance on the 20-Newsgroups and Reuters-21578 datasets in comparison to WebKB. While the first two datasets mostly consist of news documents, WebKB dataset has documents belonging to different specific topics. For this reason, it may be stated that the proposed schemes can be more effective on classifying or categorizing the collections including documents related to news.

Author contributions

Turgut Dogan: Conceived and designed the analysis; Collected the data; Performed the analysis; Wrote the paper

Alper Kursat Uysal: Conceived and designed the analysis; Contributed data or analysis tools; Performed the analysis; Wrote the paper

References

Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Odusami, M. (2019). A review of soft techniques for SMS spam classification: Methods, approaches and applications. *Engineering Applications of Artificial Intelligence*, *86*, 197–212.

Abdel Fattah, M. (2015). New term weighting schemes with combination of multiple classifiers for sentiment analysis. *Neurocomputing*, *167*, 434–442.

Alsmadi, I., & Hoon, G. K. (2019). Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing & Applications*, *31*(8), 3819–3831.

Altınçay, H., & Erenel, Z. (2010). Analytical evaluation of term weighting schemes for text categorization. *Pattern Recognition Letters*, *31*(11), 1310–1323.

Asuncion, A., & Newman, D. (1994). *UCI machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science (2007). In.

Badawi, D., & Altınçay, H. (2014). A novel framework for termset selection and weighting in binary text classification. *Engineering Applications of Artificial Intelligence*, *35*, 38–53.

Badawi, D., & Altınçay, H. (2017). Termset weighting by adapting term weighting schemes to utilize cardinality statistics for binary text categorization. *Applied Intelligence*, *47*(2), 456–472.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, *2*(3), 27.

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

Craven, M., McCallum, A., PiPasquo, D., Mitchell, T., & Freitag, D. (1998). *Learning to extract symbolic knowledge from the World Wide Web*. Carnegie-mellon univ pittsburgh pa school of computer Science.

Debole, F., & Sebastiani, F. (2004). Supervised term weighting for automated text categorization. In *Text mining and its applications*. pp. 81–97. Springer.

Deisy, C., Gowri, M., Baskar, S., Kalaiarasi, S., & Ramraj, N. (2010). A novel term weighting scheme MIDF for text categorization. *Journal of Engineering Science and Technology*, *5*(1), 94–107.

Deng, Z.-H., Luo, K.-H., & Yu, H.-L. (2014). A study of supervised term weighting scheme for sentiment analysis. *Expert Systems with Applications*, *41*(7), 3506–3513.

Dogan, T., & Uysal, A. K. (2018). The effects of globalization functions on feature weighting for text classification. *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 1–6.

Dogan, T., & Uysal, A. K. (2019). Improved inverse gravity moment term weighting for text classification. *Expert Systems with Applications*, *130*, 45–59.

Emmanuel, M., Khatri, S. M., & Babu, D. R. R. (2013). A novel Scheme for term weighting in text categorization: Positive impact factor. *2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2292–2297.

Erenel, Z., & Altınçay, H. (2012). Nonlinear transformation of term frequencies for term weighting in text categorization. *Engineering Applications of Artificial Intelligence*, *25*(7), 1505–1514.

Escalante, H. J., García-Limón, M. A., Morales-Reyes, A., Graff, M., Montes-y-Gómez, M., Morales, E. F., et al. (2015). Term-weighting learning via genetic programming for text classification. *Knowledge-based Systems*, *83*, 176–189.

Feng, G., Guo, J., Jing, B.-Y., & Sun, T. (2015). Feature subset selection using naive Bayes for text classification. *Pattern Recognition Letters*, *65*, 109–115.
Feng, G., Li, S., Sun, T., & Zhang, B. (2018). A probabilistic model derived term weighting scheme for text classification. *Pattern Recognition Letters*, *110*, 23–29.
Fourkioti, O., Symeonidis, S., & Arampatzis, A. (2019). Language models and fusion for authorship attribution. *Information Processing & Management*, *56*(6), Article 102061.
Haddoud, M., Mokhtari, A., Lecroq, T., & Abdeddaïm, S. (2016). Combining supervised term-weighting metrics for SVM text classification with extended term representation. *Knowledge and Information Systems*, *49*(3), 909–931.
Huang, C., Zhu, J., Liang, Y., Yang, M., Fung, G. P. C., & Luo, J. (2019). An efficient automatic multiple objectives optimization feature selection strategy for internet text classification. *International Journal of Machine Learning and Cybernetics*, *10*(5), 1151–1163.
Jiang, L., Li, C., Wang, S., & Zhang, L. (2016). Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence*, *52*, 26–39.
Joachims, T. (1996). *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization*. Carnegie-mellon univ pittsburgh pa dept of computer science.
Ke, W. (2014). Information-theoretic term weighting schemes for document clustering and classification. *International Journal on Digital Libraries*, *16*(2), 145–159.
Khreisat, L. (2009). A machine learning approach for Arabic text classification using N-gram frequency statistics. *Journal of Informetrics*, *3*(1), 72–77.
Kim, H. K., & Kim, M. (2016). Model-induced term-weighting schemes for text classification. *Applied Intelligence*, *45*(1), 30–43.
Ko, Y. (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*, *66*(12), 2553–2565.
Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 721–735.
Lertnattee, V., & Theeramunkong, T. (2004). Analysis of inverse class frequency in centroid-based text classification. *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on* (Vol. 2) (pp. 1171–1176).
Li, X., Zhang, A., Li, C., Ouyang, J., & Cai, Y. (2018). Exploring coherent topics by topic modeling with term weighting. *Information Processing & Management*, *54*(6), 1345–1358.
Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems with Applications*, *36*(1), 690–701.
Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, *38*(10), 12708–12716.
Nguyen, T. T., Chang, K., & Hui, S. C. (2013). Supervised term weighting centroid-based classifiers for text categorization. *Knowledge and Information Systems*, *35*(1), 61–85.
Parlak, B., & Uysal, A. K. (2019). On classification of abstracts obtained from medical journals. *Journal of Information Science*, 1–16.
Peng, H., Ma, Y., Li, Y., & Cambria, E. (2018). Learning multi-grained aspect target sequence for Chinese sentiment analysis. *Knowledge-based Systems*, *148*, 167–176.
Prasath, V., Alfeilat, H. A. A., Lasassmeh, O., & Hassanat, A. (2017). Distance and similarity measures effect on the performance of K-Nearest neighbor classifier-A review. *arXiv preprint arXiv*, 1708.04321.
Rao, Y., Li, Q., Wu, Q., Xie, H., Wang, F. L., & Wang, T. (2017). A multi-relational term scheme for first story detection. *Neurocomputing*, *254*, 42–52.
Rashid, J., Shah, S. M. A., & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management*, *56*(6), Article 102060.
Ren, F., & Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification. *Information Sciences*, *236*, 109–125.
Ropero, J., Gómez, A., Carrasco, A., & León, C. (2012). A Fuzzy Logic intelligent agent for information extraction: Introducing a new Fuzzy Logic-based term weighting scheme. *Expert Systems with Applications*, *39*(4), 4567–4581.
Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., et al. (2017). Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing*, *58*, 193–206.
Sabbah, T., Selamat, A., Selamat, M. H., Ibrahim, R., & Fujita, H. (2016). Hybridized term-weighting method for dark web classification. *Neurocomputing*, *173*, 1908–1926.
Santhanakumar, M., Columbus, C. C., & Jayapriya, K. (2018). Multi term based co-term frequency method for term weighting in information retrieval. *International Journal of Business Information Systems*, *28*(1), 79–94.
Sparck Jones, K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.
Uysal, A. K. (2016). An improved global feature selection scheme for text classification. *Expert Systems with Applications*, *43*, 82–92.
Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112.
Wang, H., & Hong, M. (2019). Supervised Hebb rule based feature selection for text classification. *Information Processing & Management*, *56*(1), 167–191.
Wei, B., Feng, B., He, F., & Fu, X. (2011). An extended supervised term weighting method for text categorization. *Proceedings of the International Conference on Human-Centric Computing 2011 and Embedded and Multimedia Computing 2011*, 87–99.
Wu, H., Gu, X., & Gu, Y. (2017). Balancing between over-weighting and under-weighting in supervised term weighting. *Information Processing & Management*, *53*(2), 547–557.
Xuan, N. P., & Le Quang, H. (2014). A New improved term weighting Scheme for text categorization. In *Knowledge and systems engineering*. pp. 261–270.
Yang, C., Zhang, H., Jiang, B., & Li, K. (2019). Aspect-based sentiment analysis with alternating coattention networks. *Information Processing & Management*, *56*(3), 463–478.
Yoon, H. G., Kim, H., Kim, C. O., & Song, M. (2016). Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling. *Journal of Informetrics*, *10*(2), 634–644.
Zhang, L., Jiang, L., Li, C., & Kong, G. (2016). Two feature weighting approaches for naive Bayes text classifiers. *Knowledge-based Systems*, *100*, 137–144.
Zhang, Y., Wang, M., Gottwalt, F., Saberi, M., & Chang, E. (2019). Ranking scientific articles based on bibliometric networks with a weighting scheme. *Journal of Informetrics*, *13*(2), 616–634.