# Extended Max-Occurrence with Normalized Non-Occurrence as MONO Term Weighting Modification to Improve Text Classification

Cristopher C. Abalorio[1], Ruji P. Medina[3]
Graduate Programs
Technological Institute of the Philippines, Quezon City, Philippines

Ariel M. Sison[2]
School of Computer Studies
Emilio Aguinaldo College
Manila, Philippines

Gleen A. Dalaorao[4]
College of Computing and Information Sciences
Caraga State University
Butuan City, Philippines

***Abstract*—The increased volume of data due to advancements in the internet and relevant technology makes text classification of text documents a popular demand. Providing better representations of the feature vector by setting appropriate term weight values using supervised term weighting schemes improves classification performance in classifying text documents. A state-of-the-art term weighting scheme MONO with variants TF-MONO and SRTF-MONO improves text classification considering the values of non-occurrences. However, the MONO strategy suffers setbacks in weighting terms with non-uniformity values in its term's interclass distinguishing power. In this study, extended max-occurrence with normalized non-occurrence (EMONO) with variants TF-EMONO and SRTF-EMONO are proposed where EMO value is determined as MO interclass extensions as improvements to address its problematic weighting behavior of MONO as it neglected the utilization of the occurrence of the classes with short-distance document frequency in non-uniformity values. The proposed schemes' classification performance is compared with the MONO variants on the Reuters-21578 dataset with the KNN classifier. Chi-square-max was used to conduct experiments in different feature sizes using micro-F1 and macro-F1. The results of the experiments explicitly showed that the proposed EMONO outperforms the variants of MONO strategy in all feature sizes with an EMO parameter value of 2 sets number of classes in MO extension. However, the SRTF-EMONO showed better performance with Micro-F1 scores of 94.85% and 95.19% for smallest to largest feature size, respectively. Moreover, this study also emphasized the significance of interclass document frequency values in improving text classification aside from non-occurrence values in term weighting schemes.**

***Keywords*—Extended MO; normalized NO; text classification; term weighting scheme**

## I. Introduction

As the volume of data has dramatically increased in the digital environment due to rapid advances in the internet and developing technology, many researchers focus on the data organization, data retrieval, and data mining. One widely used technology mentioned above is text classification [1]. Text classification (TC) labels text documents with categories from predefined classes according to their content. Spam SMS and email filtering [2][3][4], author recognition [5][6], classification in clinical text and medical documents [7][8][9], sentiment analysis [10][11], and short text classification[12] [13][14] are topics in research under the different subdomains category of TC in literature.

In TC, modeling in the feature vector space used Bag-Of-Words (BOW) [14], which extracts unique words representing the text document in the collection. The machine learning algorithm assigns documents to predefined classes. The VSM numerical feature vector represents each document with significant weights thru Term Weighting Scheme (TWS). TWS plays a vital role as it directly affects the classifying performance and simplifies the classifier's jobs. Three factors comprise the TWS for TC: term frequency factor (TFF), collection frequency factor (CFF), and length normalization factor (LNF). The TFF refers to the ratio of the number a term occurs [15][16][17], while the CFF [1] refers to the ratio of the information of the specific term to each document in a text collection. The LNF is used to normalize text collection containing different document lengths. TWS is categorized into two: supervised (STWS) and unsupervised (UTWS) Term Weighting Scheme [18][19]. STWS includes the class information in the weighting process, while UTWS ignores class information in setting weights to terms. Currently, STWS is preferred to use as it outperforms previous weighting strategies [20]. Its latest state-of-the-art is the SRTF-MONO, a variant of the MONO strategy [21] that added value to the non-occurrence of a term in stressing weights.

However, the MONO strategy suffers setbacks in weighting terms with non-uniformity of values in its class document frequency such as (1) a case of giving equal values to $MONO_{Local}$ leads to equal $MONO_{Global}$ weights containing different max-occurrence and non-occurrence; (2) the occurrences of interclass distinguishing power of a term proven in literature [22] [23] to give good term weighting ability leading successful classifying performance is neglected as it focused more class' document frequency for non-occurrence members.

The motivation of this study is to address the issue of MONO in its problematic weighting behavior to terms. In generating max-occurrence, the standard MONO formula neglects the utilization of the occurrence of the classes with short-distance document frequency in non-uniformity values.

These occurrences are the concentration of values representing the interclass distinguishing power essential in improving classification performance in STWS. The top classes with higher class occurrences could affect the belongingness of a document and should not be treated as non-occurrences. Thus, this study aims to enhance MONO by extending the max-occurrence group to cater to succeeding classes with higher document frequency values to set its real class distinguishing power of a term; and employing normalization to the non-occurrences to address the imbalanced distribution of the document frequency of classes.

The remaining parts of the paper are ordered as follows. The literature review that introduces several TWS and examines the state-of-the-art MONO strategy is presented in Section 2. The enhancement and required data, techniques, and evaluation are presented in Section 3. In Section 4, we present the results and discussion of the study. Finally, in Section 6, we conclude and recommend developing this study in text classification subdomains.

## II. Literature Review

### A. Schemes on Assigning Weights to Terms

First, Term Frequency – Inverse Document Frequency (TF-IDF), a TWS, topped the list in chronological order on research related to assigning weights to terms [24]. The IDF is one of the pioneering strategies in setting weights to terms adapted from the studies in the area of information retrieval used in text classification task proposed by Karen Spärck Jones, implied that the assigning of weights to terms should take place under the collection frequency factor (IDF) to utilize the terms effectively. This collection frequency factor is adapted to join with term frequency (TF). As an improvement in TWS for TC, Deisy et al. modified the IDF called MIDF and successfully outperformed Weighted IDF and TF-IDF [25]. The results revealed computation of MIDF is characteristically easier, and there was an improved TC performance compared with the other term weighting schemes. Sabbah et al. proposed several TWS as TF-IDF’s modification (mTFmIDF, mTF, TFmIDF, and mTFIDF) which are developed with improved results than standard TF, TF-IDF, and Entropy. They also substantiated the importance of employing ELM, NB, SVM, and KNN classifiers with popular text corpora. In another study, Debole and Sebastiani proposed several term weighting strategies, namely TF-GR, TF-IG, and TF-CHI, based on feature selection methods named Gain Ratio, Information Gain, and Chi-Square [26]. The study commenced the idea of adding the class information in weighting to terms in TC, which was later called supervised term weighting schemes (STWS). The strategy improved the classifying performance compared to the previously mentioned and traditional TWS.

A novel collection frequency factor under the STWS category, introduced by Lan et al., namely TF-RF based on relevance frequency [27]. TF-RF focused terms in its terms' class distribution considering the positive and negative ratio. This new TWS showed an improved performance than Binary, TF, and TF-IDF (unsupervised TWS) and TF-IG, TF-CHI2, and TF-LogOR (supervised TWS). Another term weighting method derived from TF-RF, namely LogTFRFmax strategy proposed by Xuan and Le Quang for TC [28]. They showed that TF-RF's classification accuracy could be increased by combining reduced TF values with RF. Liu et al. introduced a TWS intended for unbalanced text datasets, namely TF-PB [29]. TF-PB is derived using the term's interclass and intraclass distribution. They showed that classification accuracy could be increased by utilizing information on a term's inner-class distribution from unbalanced text datasets. Log-TFTRR introduced by Ko is derived from class distribution by using the negative and positive class probabilities of terms [30]. Log TF-TRR showed better performance than (TF-CHI, TF-RF) over the use collection of text data such as Korean UseNet, Reuters-21578, and 20-Newsgroups.

A study called Positive Impact Factor (PIF) introduced by Emmanuel et al. showed better TC performance concerning computing time and classification accuracy upon experiments on Classic3 text collection [31]. Altınçay and Erenel investigated previously proposed and most-used TWS for TC [32]. They discovered that the ratios and term occurrence probabilities are the reasons for relative performance differences in giving weights to terms. A new collection frequency factor introduced by Altınçay and Erenel in another study derived from the logarithm of term frequencies [33]. They showed that the lesser term frequency values bore better classification performance combined with their proposed term weighting scheme. They also indicated that the distribution of TF in the collection of the text suggests the usable form of the TF factor.

Another related study, a proposed method by Cai et al. in tagging systems wherein it modeled the resource along with the user's profile, indicating the utilization of normalized form of term frequency [34]. Badawi and Altınçay introduced another study implemented for binary text classification using a termset weighting strategy [35]. They wanted to prove using the bag-of-words approach as an effective method. They showed that their process produces document vectors successful for TC. Later, they introduced new cardinality statistic-based TWS [36]. Two collection frequency factors integrated with Bag-of-Words (BOW) were used in this new term weighting strategy. They emphasized the success of text classification performances in setting weights to terms generated by the standard BOW approach can be increased with n-term setting its values to n = 2, 3, 4.

In another study, a term weighting scheme adapted from information of the term’s document was introduced by Ren and Sohrab. [37]. They showed that TF-IDF-ICSDF outperforms five previous TWS (TF-CC, TF-IDF, TF-PB, TF-OR, and TF-RF) on 20 Newsgroups, RCV1-v2, and Reuters-21578 with Centroid, SVM, and NB classifiers. Escalante et al. used genetic programming to improve TWS [38]. They stated that genetic programming aims to acquire which combination of units makes greater discriminative TWS. They showed that genetic programming generated superior classification results than latest and popular strategy on assigning weight to terms. A new collection frequency factor for TWS introduced by Chen et al. named Inverse Gravity Moment (IGM) adapted from a statistical model [22]. IGM has two variants, namely TF-IGM and SRTF-IGM. They stated that the adequate distinguishing power is taken from the information of the interclass distribution of the documents. They showed that IGM performed better classification than the 5 TWS (RF, CHI2, Prob, ICSDF, IDF) using KNN and SVM classification algorithm on the popular text corpora (20-Newsgroups, TanCorp, and Reuters-21578). Dogan et al. improved the IGM by reorganizing the IGM formula to address similar values given to different term weights [23]. The improvement in the IGM provides better classification results than the standard IGM. Sabbah et al. implemented a hybridized term weighting strategy for terrorism activity detection in texts [39] by creating a set of features from the combination of small feature subsets taken from TF, IDF, TF-IDF, Entropy, and Glasgow TWS. They also indicated that the successful text classification could be improved by utilizing small sets of features representing the most significant terms inside the text.

With all the previous and successfully introduced supervised TWS, the concept is always the involvement of the available class information. In contrast, another collection frequency factor STWS presented by Dogan et al., called max-occurrence and non-occurrence (MONO) [21], added the value of non-occurrence or the absence of a term in the distribution of terms in the documents on each class upon stressing weights. MONO comes with two variants, TF and SRTF (the squared root of TF). SRTF-MONO variant outperforms the classification performance of the previous STW using the news dataset.

### *B. Assigning Weight to Terms using MONO Strategy*

The standard MONO term weighting scheme [21] according to Dogan et'al:

*1)* Assume that the document frequency df of a term $t_i$ or $df_{t_i}$ from the j classes of a text collection is shown in (1). The $df_{t_i}$ collection is sorted in descending order. The head of the left has the highest value, and the tail in the right has the lowest as shown in (2).

*2)* The sorted $df_{t_i}$ is divided into two groups; one has the highest class $df_{t_i}$ values and the other for the rest of the classes. Groups are categorized into two the max-occurrence (MO) ratio group and the non-occurrence (NO) ratio group as shown in (3).

*3)* $MO_{t_i}$ value, corresponds the ratio between total document frequencies and the total number of documents available on the class with maximum occurring $t_i$. After calculating the $MO_{t_i}$ value, the $NO_{t_i}$ value is calculated as shown in (4). $NO_{t_i}$ value is calculated on classes excluded from $MO_{t_i}$ calculation. The value is computed from the ratio between the quantity of document frequencies and the total number of the documents of the class within the NO ratio group.

*4)* In order to obtain the $MONO_{Local}$ weights of a term the MO and NO is multiplied as shown in (5).

*5)* Finally, $MONO_{Global}$ weight of term $t_i$ is calculated as shown in (6). In the aforementioned equation in (6), α parameter is presented. The purpose of α is to set balance to weights to global values in the weighting stage where values ranges from 5.0 to 9.0 and 7.0 as its default.

*6)* The two TWS based upon $MONO_{Global}$ collection frequency factor are shown in (7). The $SRTF(t_i, d_k)$ is squared root of $TF$ values of the term $t_i$ in text document $d_k$.

$$df_{t1} = \left\{d_{i1}, d_{i2}, d_{i3}, d_{i2}, \dots, d_{ij-1}, d_{ij}\right\} \qquad (1)$$

$$sorted\_df_{t1} = \left\{d_{i3}, d_{i1}, d_{i4}, \dots, d_{ij}, d_{ij-1}\right\} \qquad (2)$$

$$sorted\_df_{t_i} = \left\{ \underbrace{\overbrace{d_{i3}}^{C_{t_i\_max}}}_{eq} \;|, \underbrace{\overbrace{\overline{d_{\imath 1}}, \overline{d_{\imath 4}}, \dots, \overline{d_{\imath J}}, \overline{d_{\imath J-}}}^{C_{others}}}_{eq} \right. \quad sorted\_df_{t_i} = \left\{ \underbrace{\overbrace{d_{i3}}^{MO}}_{eq} \;|, \underbrace{\overbrace{\overline{d_{\imath 1}}, \overline{d_{\imath 4}}, \dots, \overline{d_{\imath J}}, \overline{d_{\imath J-}}}^{NO}}_{eq} \right. \qquad (3)$$

$$MO_{t_i} = \frac{D_{t_{i1}\_max}}{D_{total(t_i\_max)}} \qquad NO_{t_i} = \frac{D_{\bar{t_\imath}}}{D_{total(\bar{t_\imath})}} \qquad (4)$$

$$MONO_{Local}(t_i) = \underbrace{\overbrace{\left[\frac{D_{t_{i1}\_max}}{D_{total(t_i\_max)}}\right]}^{MO_{t_i}}}_{eq} * \underbrace{\overbrace{\left[\frac{D_{\bar{t_\imath}}}{D_{total(\bar{t_\imath})}}\right]}^{NO_{t_i}}}_{eq} \qquad (5)$$

$$MONO_{Global}(t_i) = [1 + \alpha * MONO_{Local}(t_i)] \qquad (6)$$

$$TF - MONO = TF(t_i, d_k) * \ [MONO_{Global}(t_i)]$$
$$SRTF - MONO = SRTF(t_i, d_k) * \ [MONO_{Global}(t_i)]$$
(7)

### *C. Empirical Observations of Issues in MONO Local and Global Weights*

This subsection illustrates empirical observations on weighting terms using standard MONO strategy.

Assume that there exist three terms ($t_4$, $t_5$, $t_6$) where the frequencies of its documents in four classes are {100, 60, 0, 0}, {200, 40, 10, 10}, and {100, 55, 5, 0} respectively. Then, assume that the entire documents of each class are uneven, which are 200, 300, 400, and 500, respectively. The standard $MONO_{Local}$ and $MONO_{Global}$ produced similar values in this scenario.

Supervised TWS considers intraclass and interclass distinguishing essential factors in specifying weights to terms [40]. Intraclass within a specific class, while interclass in multiple classes. MO represents intra-class and NO as the interclass. However, the existing interclass is not utilized because NO neglects actual interclass existence values and obtains non-existence. Class information improves classification [27], such as TF-CHI [26] [41] considers the term's intraclass distribution. Moreover, the used of interclass [22][23] outperforms the previous TWS in classification performance. Terms ($t_4$-$t_6$) using MONO ignore interclass occurrence as it selects one member of a MO group and the rest to NO group even to other classes containing higher class-document frequency. It failed to fully represent the distinguishing power of a term as it assigns equal scores, as shown in Table I in its weights.

TABLE I. RESULTS OF THE CASE SCENARIO

| Terms | Document Frequencies | No. of documents in every class | MO | NO | MONO Local | MONO Global |
|---|---|---|---|---|---|---|
| $t_4$ | (100, 60, 0, 0) | (200, 300, 400, 500) | 0.5 | 0.95 | 0.475 | 4.325 |
| $t_5$ | (100, 40, 10, 10) | (200, 300, 400, 500) | 0.5 | 0.95 | 0.475 | 4.325 |
| $t_6$ | (100, 55, 5, 0) | (200, 300, 400, 500) | 0.5 | 0.95 | 0.475 | 4.325 |

## III. METHODOLOGY

The enhancement of MONO strategy is categorized into two major processes. The extended max-occurrence and non-occurrence normalization is shown in Fig. 1.

MO computation selects only a single class with the highest document frequency values, previously shown in (3). In the modification, extended max-occurrence (EMO) is proposed to cater to interclass occurrence values essential for supervised term weighting schemes. EMO refers to the number of class members of the MO group comprising the ratio of the number of documents in a class where the term mostly occurs and its total number of documents in that class. In the j classes in (1), assume that $EMO = 2$ where $1 < EMO < d_{ij-1}$then MO group covers $d_{i3}$ and $d_{i1}$ representing ratios as shown in (8), $EMO_{t_i}$ calculates the new value of the weight of a term as shown in (9). In non-occurrence normalization, the original NO formula in (4) is modified. As the EMO is extended, NO members are reduced. With $EMO = 2$ then each NO member is individually calculated in percentage and normalized as shown in (10). The rest of the formulas for assigning weight to a term using the EMONO is shown in (11) for local weights and (12) for global weights on different variants.

$$sorted\_df_{t_i} = \left\{ \overbrace{d_{i3}, d_{i1}}^{MO} |, \overbrace{\overline{d_{\imath 4}}, \overline{d_{\imath 5}}, \dots, \overline{d_{\imath J}}, \overline{d_{\imath J-1}}}^{NO} \right\} \tag{8}$$

$$EMO_{t_i} = \frac{MO_{t_{i1}} + MO_{t_{i2}}}{D_{total}\left(MO_{t_{i1}} + MO_{t_{i2}}\right)} \tag{9}$$

$$NO_{t_{i1}} = \frac{NO_{\overline{t_{\imath 1}}}}{D_{total}(\overline{t_{\imath 1}})} * 100 \qquad NO_{t_{i2}} = \frac{NO_{\overline{t_{\imath 2}}}}{D_{total}(\overline{t_{\imath 2}})} * 100 \qquad Normalized\_NO_{t_i} = \frac{NO_{t_{i1}} + NO_{t_{i2}}}{200} \tag{10}$$

$$EMONO_{Local}(t_i) = EMO_{t_i} * \ Normalized\_NO_{t_i} \qquad EMONO_{Global}(t_i) = [1 + \alpha * EMONO_{Local}(t_i)] \tag{11}$$

$$TF - EMONO = TF(t_i, d_k) * \ [EMONO_{Global}(t_i)] \qquad SRTF - EMONO = SRTF(t_i, d_k) * \ [EMONO_{Global}(t_i)] \tag{12}$$

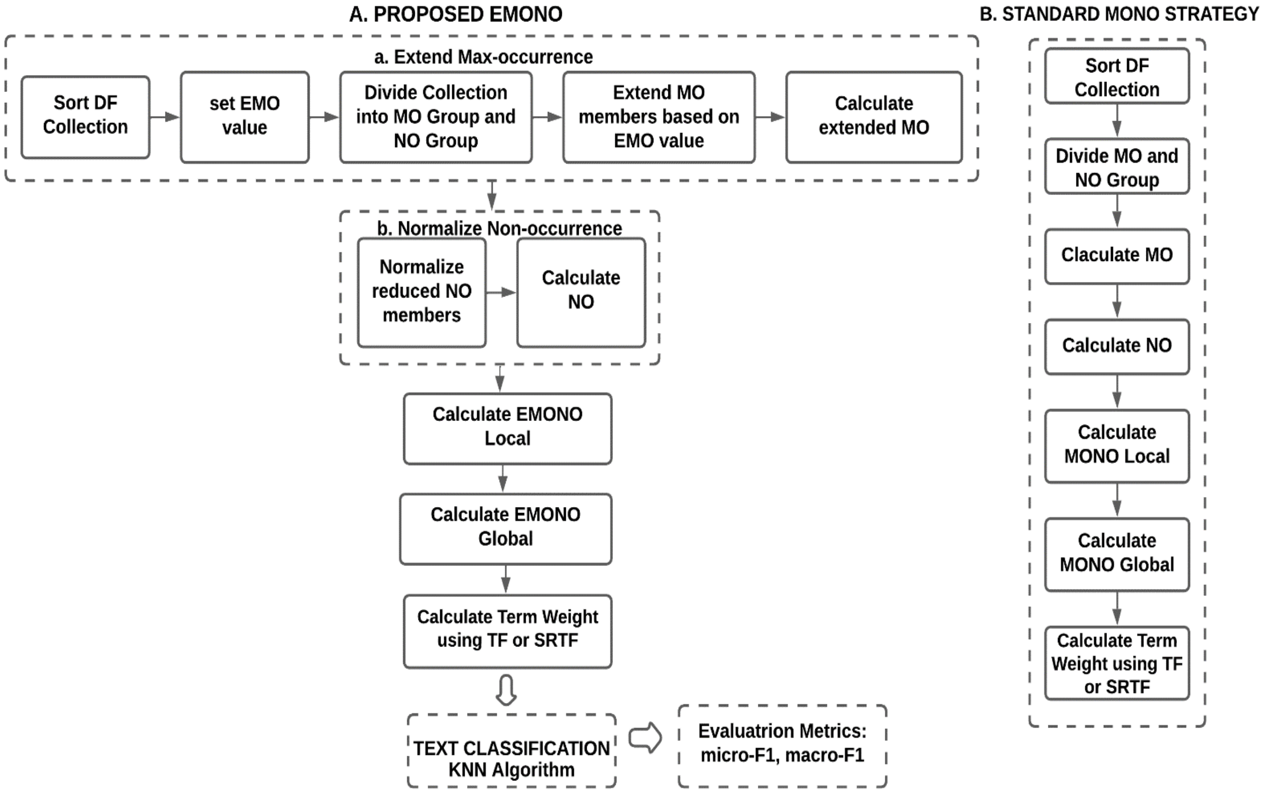
A. PROPOSED EMONO
a. Extend Max-occurrence
Sort DF Collection
set EMO value
Divide Collection into MO Group and NO Group
Extend MO members based on EMO value
Calculate extended MO
b. Normalize Non-occurrence
Normalize reduced NO members
Calculate NO
Calculate EMONO Local
Calculate EMONO Global
Calculate Term Weight using TF or SRTF
TEXT CLASSIFICATION KNN Algorithm
Evaluatrion Metrics: micro-F1, macro-F1
B. STANDARD MONO STRATEGY
Sort DF Collection
Divide MO and NO Group
Claculate MO
Calculate NO
Calculate MONO Local
Calculate MONO Global
Calculate Term Weight using TF or SRTF

Fig. 1. Framework of the Study.

### A. Dataset

Reuters-21578 is the benchmark dataset utilized in this study. It is one of the most popular unbalanced datasets preferred for text classification. Training and testing of data are taken from the Reuters-ModApte split, which is its segmentation for the experiments' training and testing. Due to removing the multi-labeled from the reading text in this dataset, the texts 'wheat' and 'corn' classes are emptied. There are a total of 7,215 documents with eight categories: earn (3735), acq (2124), money-fx (354), grain (45), crude (259), trade (332), interest (211), and ship (155).

### B. Pre-Processing

In preparing the dataset, the following are the preprocessing methods [21] applied to the documents: conversion to lowercases, removing stop-words, alphabetic tokenization, Porter Stemming [42], and eliminating of seldom occurring terms (retain words occur more than one time).

### C. Feature Selection

A feature selection method is preferred to manifest the proposed improvements' performance in a text classification dealing with high dimensionality. A standard statistical metric named chi-square max is used to obtain a selected number of features for this experiment. The total number for observations is 7,215 and the highest feature extracted is 9,237. The term weighting schemes are tested with top terms sorted in descending order and scored using the employed feature selection method {1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, and 9237} terms scored and sorted.

### D. Learning Algorithm

The well-known and most used classification algorithm K-Nearest Neighbor (KNN) is employed in this study to classify documents from the benchmark Reuters-21578 dataset. Aside from its simplicity, it is generally utilized for performing text classification in literature [43][44].

### E. Performance Evaluation

The text classification performance of this study was evaluated using micro-F1 and macro-F1 scores. Precision and Recall in (13) used to derived F1 scores preferred to use for uneven distribution of documents in the classes of multiclass text classification as shown in (14).

$$Precision_{c_k} = \frac{TP_{c_k}}{TP_{c_k}+FP_{c_k}} \qquad Recall_{c_k} = \frac{TP_{c_k}}{TP_{c_k}+FN_{c_k}}$$

$$F1_{c_k} = \frac{2*Precision_{c_k}*Recall_{c_k}}{Precision_{c_k}+Recall_{c_k}} \tag{13}$$

$$Micro-F1 = \frac{2*\sum_k^c TP_{c_k}}{2*\sum_k^c TP_{c_k}+\sum_k^c FP_{c_k}+\sum_k^c FN_{c_k}} \qquad Macro-F1 = \frac{1}{c}\sum_{k=1}^{c} F1_{c_k} \tag{14}$$

## IV. Result and Discussion

The results of the performance of the MONO and the proposed modification were analyzed in this section, along with the use of the classification algorithm KNN on Reuters-21578 corpus with all alpha values set to 6.0. The boldface values correspond to the highest F1 values in each micro and macro score. Table II shows the new local, and global weights generated using the proposed EMONO on case scenario with problematic weights mentioned in Table II.

Table III shows the classification performance of the proposed EMONO combined with term frequency on 1,000 and 9237 feature sizes. As EMO value implies the number of classes in MO extension, in 1000 features, EMO values 2, 3, and 4 have 0.927815207 as the highest scores for micro-F1, and only in the EMO value of 2 has 0.880240879 as the highest value for macro-F1. The proposed EMONO combined with the squared root of TF's EMO value of 2 has the highest obtained scores for micro-F1 and macro F1 with 0.948508181 and 0.893171284, respectively. In TF-EMONO 9237 features, the EMO value of 3 got the maximum micro-F1 score of 0.918671800 as the highest score, and the EMO value of 2 has 0.860379651 as the highest value for macro-F1. On the other hand, the classification performance of the proposed EMONO combined with the squared root of term frequency EMO value of 4 has the highest obtained scores of 0.957170356 and 0.909860837 micro-F1 to macro-F1, respectively.

The comparative results of the classification performance of the original MONO and proposed EMONO combined with TF and squared root of TF with 2 as the successful EMO value is shown in Table IV. The table shows that EMONO combined with both TF and SRTF have greater values than the original MONO strategy in micro-F1 and macro-F1. It is explicit that the proposed EMONO generally outperformed the original MONO strategy in all indicated feature sizes shown in the table.

Fig. 2 shows the classification performance of micro-F1 and macro-F1 in plot graph. The overall performance of the proposed EMONO is superior when combined with TF and squared root of TF explicitly shown in the plot graph.

TABLE II. EMONO Weights on Issues in Local and Global Weights

| Terms | Document Frequencies | No. of documents in every class | Extended MO | Normalized NO | EMONO Local | EMONO Global |
|---|---|---|---|---|---|---|
| $t_4$ | (100, 60, 0, 0) | (200, 300, 400, 500) | 0.32 | 1 | 0.32 | 3.24 |
| $t_5$ | (100, 40, 10, 10) | (200, 300, 400, 500) | 0.28 | 0.9775 | 0.2737 | 2.9159 |
| $t_6$ | (100, 55, 5, 0) | (200, 300, 400, 500) | 0.31 | 0.99375 | 0.3080625 | 3.1564375 |

TABLE III. MODIFIED MONO ON MINIMUM AND MAXIMUM FEATURE SIZES

| EMO | TF-EMONO 1,000 Features | | SRTF-EMONO 1,000 Features | | TF-EMONO 9,237 Features | | SRTF-EMONO 9,237 Features | |
|---|---|---|---|---|---|---|---|---|
| | micro-F1 | macro-F1 | micro-F1 | macro-F1 | micro-F1 | macro-F1 | micro-F1 | macro-F1 |
| 2 | **0.927815207** | **0.880240879** | **0.948508181** | **0.893171284** | 0.910490857 | **0.860379651** | 0.951876805 | 0.898198983 |
| 3 | **0.927815207** | 0.861238997 | 0.941289702 | 0.882265425 | **0.918671800** | 0.864875929 | 0.954764196 | 0.907354723 |
| 4 | **0.927815207** | 0.861238997 | 0.942252166 | 0.888250757 | 0.913859480 | 0.856517844 | **0.957170356** | **0.909860837** |
| 5 | 0.918190568 | 0.845061943 | 0.936477382 | 0.873495059 | 0.913859480 | 0.850101527 | 0.948989413 | 0.897765549 |
| 6 | 0.920596728 | 0.845835989 | 0.935514918 | 0.869062724 | 0.911934552 | 0.851904410 | 0.948026949 | 0.894754037 |
| 7 | 0.914821944 | 0.824183935 | 0.934552454 | 0.848188784 | 0.900384986 | 0.819861594 | 0.950914341 | 0.889930181 |

TABLE IV. COMPARATIVE RESULTS OF MONO AND MODIFIED

| No. of Features | micro-F1 | | | | macro-F1 | | | |
|---|---|---|---|---|---|---|---|---|
| | TF-MONO | TF-EMONO | SRTF-MONO | SRTF-EMONO | TF-MONO | TF-EMONO | SRTF-MONO | SRTF-EMONO |
| 1000 | 0.9095 | 0.9278 | 0.9365 | 0.9485 | 0.7952 | 0.8802 | 0.8749 | 0.8932 |
| 2000 | 0.9100 | 0.9211 | 0.9355 | 0.9398 | 0.7964 | 0.8650 | 0.8594 | 0.8696 |
| 3000 | 0.9076 | 0.9196 | 0.9346 | 0.9413 | 0.8038 | 0.8794 | 0.8702 | 0.8727 |
| 4000 | 0.9038 | 0.9153 | 0.9418 | 0.9442 | 0.7971 | 0.8703 | 0.8808 | 0.8856 |
| 5000 | 0.9033 | 0.9110 | 0.9403 | 0.9485 | 0.8031 | 0.8627 | 0.8808 | 0.8932 |
| 6000 | 0.8985 | 0.9086 | 0.9398 | 0.9514 | 0.8028 | 0.8624 | 0.8816 | 0.8985 |
| 7000 | 0.9023 | 0.9090 | 0.9437 | 0.9514 | 0.8104 | 0.8615 | 0.8891 | 0.9002 |
| 8000 | 0.9038 | 0.9100 | 0.9480 | 0.9519 | 0.8109 | 0.8619 | 0.8942 | 0.8966 |
| 9000 | 0.9076 | 0.9095 | 0.9471 | 0.9524 | 0.8121 | 0.8600 | 0.8934 | 0.8984 |
| 9237 | 0.9081 | 0.9105 | 0.9480 | 0.9519 | 0.8123 | 0.8604 | 0.8954 | 0.8982 |

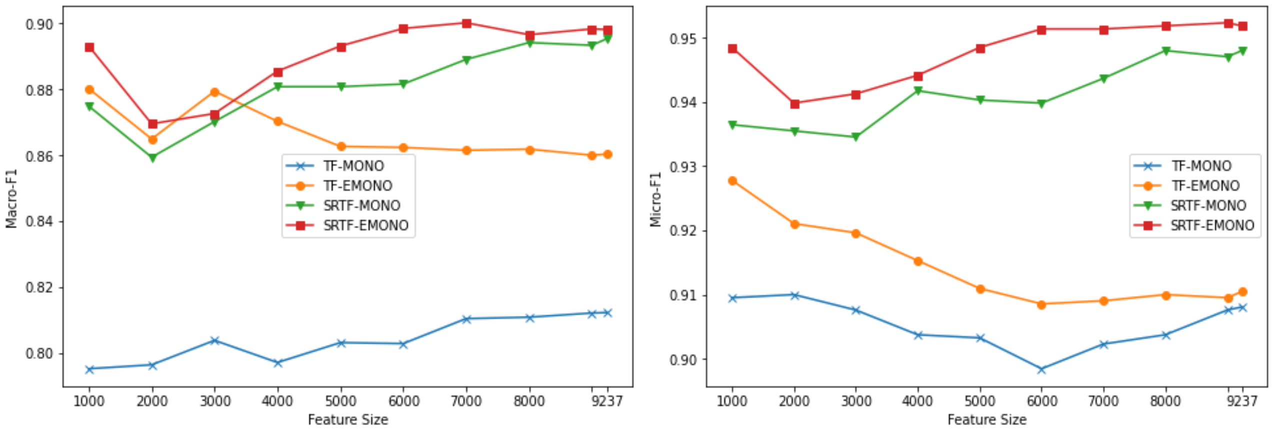

Fig. 2. Micro-F1 Acquired from 4 TWS on Reuters-21578 Dataset using KNN (k=5) with EMO=2.

## V. CONCLUSION AND RECOMMENDATION

In this study, behavior on how the original MONO strategy assigned a weight to terms was comprehensively analyzed. Two new variants of the EMONO, namely TF-EMONO and SRTF-EMONO, are proposed. The MONO modifications employed in the study are the EMO parameter's utilization and the normalization of the non-occurrences. EMO parameter sets classes to cover in generating max-occurrence extension, and the normalization is utilized for class imbalance. The results of the experiments explicitly showed that the proposed improvement outperforms the variants of the original MONO strategy in all feature sizes with an EMO parameter value of 2 sets number of classes in MO extension. Even though there are feature sizes in which EMONO has a slight increase of micro-F1 and macro-F1, the increase is consistent and gradual in all features with a selected EMO value. However, the SRTF-EMONO showed better performance with Micro-F1 scores of 94.85% and 95.19% for smallest to largest feature size. It can be stated that the proposed EMONO term weighting scheme is superior in classification performance to the original MONO strategy. It is recommended that the proposed EMONO be implemented in text classification. It is also suggested to utilize it in other text classification subdomains.

## References

[1] M. Marcin Michał and J. Protasiewicz, "A Recent Overview of the State-of-the-Art Elements of Text Classification," 2018, doi: 10.1016/j.eswa.2018.03.058.

[2] O. Abayomi-alli, S. Misra, A. Abayomi-alli, and M. Odusami, "A review of soft techniques for SMS spam classification: Methods, approaches and applications," Eng. Appl. Artif. Intell., vol. 86, no. August, pp. 197–212, 2019, doi: 10.1016/j.engappai.2019.08.024.

[3] R. Pav, D. Ruano-ord, G. Silvana, and R. M, "Enhancing representation in the context of multiple-channel spam filtering," vol. 59, no. May 2021, 2022, doi: 10.1016/j.ipm.2021.102812.

[4] A. K. Uysal and S. Gunal, "A novel probabilistic feature selection method for text classification," Knowledge-Based Syst., vol. 36, pp. 226–235, 2012, doi: 10.1016/j.knosys.2012.06.005.

[5] O. Fourkioti, S. Symeonidis, and A. Arampatzis, "Language models and fusion for authorship attribution," Inf. Process. Manag., vol. 56, no. 6, p. 102061, 2019, doi: 10.1016/j.ipm.2019.102061.

[6] C. Akimushkin and D. R. Amancio, "On the role of words in the network structure of texts: application to authorship attribution," Phys. A Stat. Mech. its Appl., pp. 49–58, 2018.

[7] X. Li, M. Cui, J. Li, R. Bai, Z. Lu, and U. Aickelin, "A hybrid medical text classification framework: Integrating attentive rule construction and neural network," Neurocomputing, vol. 443, pp. 345–355, 2021, doi: 10.1016/j.neucom.2021.02.069.

[8] R. Matsuo and T. B. Ho, "Semantic Term Weighting for Clinical Texts," pp. 543–551, 2018.

[9] G. Mujtaba et al., "Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues," Expert Syst. Appl., 2018, doi: 10.1016/j.eswa.2018.09.034.

[10] A. Mee, E. Homapour, F. Chiclana, and O. Engel, "Sentiment analysis using TF – IDF weighting of UK MPs ' tweets on Brexit," Knowledge-Based Syst., vol. 228, p. 107238, 2021, doi: 10.1016/j.knosys.2021.107238.

[11] G. N. H, R. Siautama, A. C. I. A, and D. Suhartono, "Extractive Hotel Review Summarization based on TF / IDF and Adjective-Noun Pairing by Considering Annual Sentiment Trends," Procedia Comput. Sci., vol. 179, no. 2020, pp. 558–565, 2021, doi: 10.1016/j.procs.2021.01.040.

[12] I. Alsmadi, "Term weighting scheme for short-text classification: Twitter corpuses," Neural Comput. Appl., vol. 8, 2018, doi: 10.1007/s00521-017-3298-8.

[13] S. S. Samant, N. L. B. Murthy, and A. Malapati, "Improving Term Weighting Schemes for Short Text Classification in Vector Space Model," IEEE Access, vol. 7, pp. 166578–166592, 2019, doi: 10.1109/ACCESS.2019.2953918.

[14] V. N. Gudivada, D. L. Rao, and A. R. Gudivada, Information Retrieval : Concepts , Models , and Systems, 1st ed. Elsevier B.V., 2018.

[15] T. Dogan and A. K. Uysal, "On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification," Arab. J. Sci. Eng., 2019, doi: 10.1007/s13369-019-03920-9.

[16] L. Chen, L. Jiang, and C. Li, "Using modified term frequency to improve term weighting for text classification," Eng. Appl. Artif. Intell., vol. 101, no. November 2020, p. 104215, 2021, doi: 10.1016/j.engappai.2021.104215.

[17] Y. He, T. Li, Y. Huang, and S. Li, "Term Weight Algorithm Oriented Terms: Low Frequency Rather Than Little Occurrences," Procedia Comput. Sci., vol. 176, pp. 838–847, 2020, doi: 10.1016/j.procs.2020.09.079.

[18] F. Sebastiani, "Machine Learning in Automated Text Categorization," vol. 34, no. 1, pp. 1–47, 2002.

[19] H. Wu, X. Gu, and Y. Gu, "Balancing between over-weighting and under-weighting in supervised term weighting," Inf. Process. Manag., vol. 0, pp. 1–11, 2016, doi: 10.1016/j.ipm.2016.10.003.

[20] A. Alsaeedi, "A survey of term weighting schemes for text classification Abdullah Alsaeedi," vol. 12, no. 2, pp. 237–254, 2020.

[21] T. Dogan and A. K. Uysal, "A novel term weighting scheme for text classification: TF-MONO," J. Informetr., vol. 14, no. 4, p. 101076, 2020, doi: 10.1016/j.knosys.2012.06.005.

[22] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," Expert Syst. Appl., vol. 66, pp. 245–260, 2016, doi: 10.1016/j.eswa.2016.09.009.

[23] T. Dogan and A. K. Uysal, "Improved inverse gravity moment term weighting for text classification," Expert Syst. Appl., vol. 130, pp. 45–59, 2019, doi: 10.1016/j.eswa.2019.04.015.

[24] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," vol. 60, no. 5, 2004.

[25] C. Deisy, M. Gowri, S. Baskar, and N. Ramraj, "A NOVEL TERM WEIGHTING SCHEME MIDF FOR TEXT CATEGORIZATION," vol. 5, no. 1, pp. 94–107, 2010.

[26] F. Debole and F. Sebastiani, "Supervised Term Weighting for Automated Text Categorization *," pp. 784–788, 2004, doi: https://doi.org/10.1007/978-3-540-45219-5_7.

[27] M. Lan, C. L. Tan, S. Member, J. Su, and Y. Lu, "Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," vol. 31, no. 4, pp. 721–735, 2009.

[28] N. P. Xuan and H. Le Quang, "A New Improved Term Weighting Scheme for Text Categorization," vol. 1, pp. 261–270, 2014, doi: 10.1007/978-3-319-02741-8.

[29] Y. Liu, H. Tong, and A. Sun, "Imbalanced text classification: A term weighting approach," Expert Syst. Appl., vol. 36, no. 1, pp. 690–701, 2009, doi: 10.1016/j.eswa.2007.10.042.

[30] Y. Ko, "A New Term-Weighting Scheme for Text Classification Using the Odds of Positive and Negative Class Probabilities," vol. 66, no. 12, pp. 2553–2565, 2015, doi: 10.1002/asi.

[31] M. Emmanuel, S. M. Khatri, and D. R. R. Babu, "A Novel scheme for Term weighting in Text Categorization: Positive Impact factor," IEEE Int. Conf. Syst. Man, Cybern., 2013, doi: 10.1109/SMC.2013.392.

[32] H. Altınçay and Z. Erenel, "Analytical evaluation of term weighting schemes for text categorization," Pattern Recognit. Lett., vol. 31, no. 11, pp. 1310–1323, 2010, doi: 10.1016/j.patrec.2010.03.012.

[33] Z. Erenel and H. Altıncay, "Nonlinear transformation of term frequencies for term weighting in text categorization," vol. 25, pp. 1505–1514, 2012, doi: 10.1016/j.engappai.2012.06.013.

[34] Y. Cai, Q. Li, H. Xie, and H. Min, "Exploring Personalized Searches using Tag-based User Profiles and Resource Profiles in Folksonomy," Neural Networks, 2014, doi: 10.1016/j.neunet.2014.05.017.

[35] D. Badawi and H. Altincay, "A novel framework for termset selection and weighting in binary text classification," Eng. Appl. Artif. Intell., vol. 35, pp. 38–53, 2014, doi: 10.1016/j.engappai.2014.06.012.

[36] D. Badawi and H. Altıncay, "Termset weighting by adapting term weighting schemes to utilize cardinality statistics for binary text categorization," 2017, doi: 10.1007/s10489-017-0911-6.

[37] F. Ren and M. G. Sohrab, "Class-indexing-based term weighting for automatic text classification," Inf. Sci. (Ny)., vol. 236, pp. 109–125, 2013, doi: 10.1016/j.ins.2013.02.029.

[38] H. J. Escalante et al., "Term-weighting learning via genetic programming for text classification," KNOWLEDGE-BASED Syst., 2015, doi: 10.1016/j.knosys.2015.03.025.

[39] T. Sabbah, A. Selamat, R. Ibrahim, and H. Fujita, "Hybridized term-weighting method for Dark Web classification," Neurocomputing, 2015, doi: 10.1016/j.neucom.2015.09.063.

[40] U. I. Akpana and A. Starkey, "Review of classification algorithms with changing inter-class distances," vol. 4, no. November 2020, 2021, doi: 10.1016/j.mlwa.2021.100031.

[41] H. Peng, Y. Ma, Y. Li, and E. Cambria, "Learning multi-grained aspect target sequence for Chinese sentiment analysis," Knowledge-Based Syst., vol. 148, pp. 167–176, 2018, doi: 10.1016/j.knosys.2018.02.034.

[42] M. .Porter, "An algorithm for suffix stripping," vol. 40, pp. 211–218, 2006, doi: 10.1108/00330330610681286.

[43] T. Sabbah et al., "Modified Frequency-Based Term Weighting Schemes for Text Classification," Appl. Soft Comput. J., 2017, doi: 10.1016/j.asoc.2017.04.069.

[44] V. B. S. Prasath, H. Arafat, A. Alfeilat, A. B. A. Hassanat, O. Lasassmeh, and S. Ahmad, "Effects of Distance Measure Choice on KNN Classifier Performance - A Review," pp. 1–39, 2019.