A Dynamic Two-Layers MI and Clustering-based Ensemble Feature Selection for Multi-Labels Text Classification

Adil Yaseen Taha[1], Sabrina Tiun[2], Abdul Hadi Abd Rahman[3]*, Masri Ayob[4], Ali Sabah[5]
Center for Artificial Intelligence Technology (CAIT)
Faculty of Information Science and Technology
University Kebangsaan Malaysia
43600 Bangi, Selangor
Malaysia

***Abstract*—Multi-label text classification deals with the issue that arises from each sample being related to multiple labels. The text data suffers from high dimensionality. In order to resolve this issue, a feature selection (FS) method can be implemented for efficiently removing the noisy, irrelevant, and redundant features. Multi-label FS is a powerful tool for solving the high-dimension problem. With regards to handling correlation and high dimensionality problems in multi-label text classification, this paper investigates the various heterogeneous FS ensemble schemes. In addition, this paper proposes an enhanced FS method called dynamic multi-label two-layers MI and clustering-based ensemble feature selection algorithm (DMMC-EFS). The proposed method considers the: 1) dynamic global weight of feature, 2) heterogeneous ensemble, and 3) maximum dependency and relevancy and minimum redundancy of features. This method aims to overcome the high dimensionality of multi-label datasets and acquire improved multi-label text classification. We have conducted experiments based on three benchmark datasets: Reuters-21578, Bibtex, and Enron. The experimental results show that DMMC-EFS has significantly outperformed other state-of-the-art conventional and ensemble multi-label FS methods.**

Keywords*—Multi-label text classification; high dimensionality; filtering method; ensemble clustering; ensemble MI feature selection

I. INTRODUCTION

In multi-label text classification, each sample is related to one or more classes at the same time. The difference between main key to a multi-label learning and single label learning is that the labels in the multi-label learning are related and inclusive. Thus, the problems related to multi-label learning are more challenging to solve. In the field of machine learning and data mining, multi-label learning is an endeavor task that greatly suffers from high dimensionality [1] [2].

The limitation of this research in multi-label text learning process, there is a significant number of irrelevant, redundant, and disruptive information. The number of involved features is usually large. The high dimensionality of multi-label text data results in challenges such as poor performance, over-fitting, and anything from computational to classification complexity. Some existing multi-label feature selection (FS) methods can be considered in order to minimize the effect of the irrelevant and redundant features that disrupts the learning process [3]. A label or a class can be a non-convex region which is a union of several overlapping or disjointed sub-regions. As a result, they may suffer from large memory requirements or poor performance. FS is a method that aims to discover a minor subset of features that can define the original features of the dataset or something better [4] [5], and it can be regarded as an effective way to manage the problem of high dimensionality. FS can reduce the dimensionality of the original data by speeding up the learning process and building comprehensible learning models with quality generalization performance. In multi-label learning, there is a need to implement multi-label feature reduction techniques so that they remove any irrelevant features and transform high dimensional documents into low dimensional ones. Many algorithms exist that can simplify the multi-label FS sets, but they neglect the interrelations among multi-label FS sets. However, multi-label filter-based FS algorithms consider the label interactions and are able to promptly and effectively select features [1] [6] by evaluating the measures. Several researches [7] [8] [9] [10], have proposed the adaption of single-label FS techniques.

The multi-label FS algorithms are designed based on the decomposition of multi-label learning into a number of single-label classification, and thus, they ignore the correlation between the different labels. By reviewing the existing studies, it can be assumed that the single-label filter-based FS methods are not appropriate for multi-label datasets. Therefore, after taking several factors into consideration, it seems reasonable to propose:

The first priority of the FS method should be to maximize the feature-class dependency and minimize the feature-feature conditional redundancy. FS methods help reduce the redundant dimensions without suffering the loss of the total information. These redundant features [11] [12] [13] provide overlapped information about the selected feature.

Secondly, a good ensemble FS method should take into account the functional diversity of the data. In other words, the ensemble FS method should reduce the possibility of overvoting caused by the other existing FS methods [14].

*Corresponding Author

Thirdly, the FS method [15] should consider the dynamic changes of the selected features along with the class and dynamic global weight of the feature.

Therefore, the following are the expected key contributions of this paper:

1) Investigating the several state-of-the-art conventional multi-label FS methods that have been derived from different mathematical and statistical concepts for generating different FS solutions. The aim is to identify and select the best multi-label FS methods that can be used in the ensemble FS method. The expected outcome of this endeavor is to identify features that are effective in accomplishing the intended tasks.

2) Proposing two multi-label ensemble FS methods: multi-label Mean ensemble FS method and multi-label Plurality Vote ensemble FS method.

3) Designing a new dynamic multi-label MI and clustering-based ensemble FS method that considers the functional diversity and dynamic changes of the selected features along with the class and dynamic global weight of the feature.

Thus, this paper is presented in several sections where Section II briefly reviews the related work; Section III briefly describes the FS methods that have been used in this study; Section IV explains the framework of our proposed multi-label ensemble FS method that solves the problem of multi-label high-dimensionality in multi-label text classification; Section V presents the classifier models used in the experiments; Section VI presents the experimental work; Section VII presents the experiment results; The results discussion is presented in Section VIII; and lastly, Section IX concludes the paper.

II. Related Work

In multi-label text classification, the goal of an FS method is to reduce the feature space dimensions and improve the classification efficiency and performance by removing redundant and irrelevant (disruptive) features. In multi-label text classification, there is a need for a method that ensures multi-label feature reduction by subtracting the irrelevant features and transforming high dimensional documents into low dimensional ones. Many existing algorithms simplify the multi-label FS sets but neglect the interrelations among the features of multi-label data sets. Multi-label filter-based FS methods consider label interactions and promptly and effectively selects features based on evaluating measures [3].

In [16], proposed an ensemble filter-based FS technique for multi-label data classification. As suggested in [16], ensemble FS provides relatively stable feature ranking and reduces the negative effects of the change in the training dataset. This technique combines the results of four FS methods in order to create an ensemble method. In [17], proposed an ensemble method that employs a prediction risk and forward search strategy for creating an ensemble FS method. They were used to evaluate the importance of selection of the features in order to generate a feature subset that can be employed to improve the classifier's performance. In [18], incorporated a mutual information measure in an ensemble method in order to create an optimal subset of features. The approach combines multiple algorithms, such as Info Gain, Gain-ratio, Relief, Chi-square and Symmetric Uncertainty. In [19], an ensemble method multi-label FS algorithm based on information entropy (EMFSIE) was proposed. The core idea of this method is to accomplish information gain for evaluating the correlation between the feature and the label set and more effectively filtering out the irrelevant features. In [20], proposed an improved global FS scheme (IGFSS) on an ensemble method that combines the superior functionality of a filter-based global FS method and a one-sided local FS method. The idea behind IGFSS is to allow the feature set almost equally represent each class in the dataset. In [11], presented a new FS method that is based on term frequency reordering of document-level (TRDL). The TRDL uses the document's frequency to measure the unbalanced factors in the data set and considers the effect of the term "frequency" on the ordering the importance of the features. Author in [21] also proposed a new text FS method based on the mutual information using sample variance (MIUSV). MIUSV is a typical variation in terms of distribution and also calculates the mutual information score of the term. In [10] proposed a fast-multi-label FS method, which is called MLFR that implements an information-theoretic feature ranking. The method in [10] speeds up the search process by scrapping the dispensable calculations and identifying the important label combinations for accomplishing a fast-multi-label FS. The method demonstrates the relationship between the labels and the features using a graphical scheme. The proposed method of [10] was used to solve a problem with datasets that contain discrete values, as it used a Symmetric Uncertainty criterion for evaluating the features. In addition, by reviewing the proposed methods in [22] [23], it can be stated that both methods used an adaptation entropy calculation in order to calculate information gain for each feature in the multi-label dataset. The features were then selected based on the resultant top scores. An FS method, which was proposed in [24]. [24], was based on information gain. The proposed methods identifies the relationship between the features and the labels in order to discover the importance of each feature in the multi-label dataset.

Based on the conclusions provided by [16] [17] [18], it was found that the FS ensemble methods provide promising results for solving the high dimensionality problem in multi-label text classification. It is crucial to further investigate the use of the ensemble filter-based method in different applications, such as using it for multi-label FS methods wherein the results are expected to be higher.

This work examines the various heterogeneous FS ensemble schemes and proposes a dynamic multi-label two layers MI and clustering-based ensemble feature selection algorithm (DMMC-EFS). The proposed method takes the following factors into account: 1) dynamic global weight of the feature; 2) heterogeneous ensemble; 3) maximum dependency and relevancy and minimum redundancy of the features. In the following section (Section 3), we will discuss the various FS ensemble methods before we venture into explaining in detail our proposed method on multi-label FS (Section 4).

III. Multi-Label FS Ensemble Methods

FS is an important for ensuring the attainment of an effective multi-label text classification system. Adapting an FS method improves the performance of text classification tasks in terms of their learning speed and effectiveness. An FS method also reduces the number of data dimensions and removes any irrelevant, redundant, and disruptive data. FS methods can be effective to solve the classification problem of multi-label datasets. They can improve the performance of the tasks and even speed up the process. They can remove the irrelevant, noisy, and redundant data. Several approaches, including filter methods and wrapper methods, have been considered in order to perform FS for multi-label learning. The filter-based FS methods do not take into account features redundancy and feature class dependency, and their results are inconsistent with the available classifiers. On the other hand, the wrapper methods usually produce better results, but their drawback includes the risk of overfitting and high computational complexity. The ensemble methods [25] are also popular methods for FS in case of high dimensional datasets. However, the redundancy of the features among themselves and all the class labels is not considered by the existing ensemble-based FS methods.

In order to design an effective multi-label FS method so that it can remove the irrelevant features and handle the high dimensionality problem, it should be able to ensure minimum redundancy among the selected features and have maximum dependency between features and all the class labels [18] [26] [27] [28] [19] [1] [29] [6]. In addition, the multi-label FS method should scalable, not computationally demanding, and be as fast as the filtering methods, and they should perform well like the wrapper methods.

In order to handle the correlation and high dimensionality problems in multi-label text classification, this work investigates the various heterogeneous FS ensemble schemes and proposes an FS method (DMMC-EFS). The baseline FS method and multi-label FS method is described in Section 3.1 and Section 3.2, respectively.

A. Baseline Feature Selection Method

Several FS methods have been analyzed in order to select features from each sample, including Information Gain, F-score, Relief, mutual information, and normalized mutual information. Based on the existing literature review [27] [28] [19] [6] [30], these methods and their extensions prove to be effective in case of multi-label FS, and in addition, they are able to cope with the feature-label correlation [1].

1) Information Gain (IG): IG is an FS algorithm [31] [32] that is used to measure the quality of the features in solving the machine learning problem. The appearance or absence of a feature is measured in order to what extent it contributes to the attainment of a correct classification result. IG is one of the most popular and commonly used FS in the multi-label text classification system. It is formally defined by using the following equation:

$$IG(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p\big(x(i),y(j)\big)\, log\big(x(i)\big) \qquad (1)$$

Here, $p\big(x(i)\big)$ indicates the likelihood of feature x, and $p\big(x(i),y(j)\big)$ is the joint likelihood when $\big(x(i),y(j)\big)$ is denoted simultaneously.

2) F-score: F-score is a multi-label FS method [3] that evaluates the discriminative ability of the features. F-score estimates the relevance of the features based on their ability to discriminate between the groups of the target variable and discrimination within each group. A higher F-score indicates an increased likelihood that this feature is discriminative. It is formally defined through the following equation:

$$F-score_i = \frac{\sum_{k=1}^{c}(\overline{f_i^k}-\overline{f_i})}{\sum_{k=1}^{c}\left[(\frac{1}{N_i^k}-1)\sum_{j=1}^{N_i^k}(\overline{f_{ij}^k}-\overline{f_i^k})\text{^}2\right]} \qquad (2)$$

Here, c is the number of labels, and n is the number of features; N_i^k is the number of samples of the feature i in label k, (k = 1, 2,…, c; i = 1, 2,…, n), x_{ij}^k is the j the training sample for the feature i in class k, (j = 1, 2,…, N_i^k), $\overline{f_i}$ is the mean value of feature i from all labels, and$f_i^k$is the mean of the ith feature of the samples in label k.

3) Relief: Relief is the most effective and commonly used FS [5] [30] [33], in multi-label text classification system. The Relief randomly selects instances from the training data, and then estimates the features' relevance to a class based on the closest data that can be found. It assigns a high weight to the features based on each instance's ability to differentiate between the classes [32] [30]. Relief algorithm is the only individual evaluation filter-based algorithm that is capable of detecting feature dependencies.

4) Mutual Information (MI): Relief is the most effective and commonly used FS in multi-label text classification system [27] [30].

$$\mathrm{MI}(x,y) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{p}\big(x(i),y(j)\big)\, log\, \frac{\mathrm{p}(x(i),y(j))}{\mathrm{p}(x(i))\times p(y(j))} \qquad (3)$$

Here, $p\big(x(i)\big)$, is the likelihood of incidence of a feature x, and $p\big(x(i),y(j)\big)$ is the joint likelihood when, $\big(x(i),y(j)\big)$ happens simultaneously. Depending on the MI definition, the filter process is described by the following steps:

a) Compute the MI of the features.

b) Use the MI values in order to calculate the mean and their standard deviation.

c) Remove any feature that has an MI value below the value acquired by subtracting the standard deviation from the mean.

5) Normalized Pointwise Mutual Information (NPMI): The measure of the mutual information FS provides a formal way to model the mutual information between the terms and the classes [6] [34]. The mutual information MI (t, c) between the term t and the class c can be defined on the basis of the level of co-occurrence between a feature f_j and a class c_i. In this work, the normalized pointwise mutual information FS method has been adopted in order to select the features for each class according to the co-occurrence measure between a feature f_j

and a class c_i. The normalized pointwise mutual information (NPMI) between the feature and its class [34] [35] can be calculated using the following equations:

$$PMI(class = c_i,) = lin\frac{p(c_i,f_j)}{p\,(c_i).p(f_j)} \quad (4)$$

$$NPMI\big(class = c_i, f_j\big) = \frac{PMI(c_i,f_j)}{\sum_{f_k} PMI\,(c_i,f_k)} \quad (5)$$

B. Multi-Label Mean Ensemble Feature Selection Method (ME-mean)

The multi-label mean ensemble FS method [29] calculates the mean feature scores across all the FS methods and then finds the overall mean value. This value is used to create the final feature list. Let us consider n data samples, $S^1, \ldots\ldots, S^n$, base feature methods, $FS^1, \ldots\ldots, FS^n$. Here, each FS method SF^i selects a list of m features $F^i = \{f^1, \ldots\ldots, f^m\}$ from the data sample S^i. The final score or ensemble score of a feature f^j from any list of features is calculated using the following equation:

$$E_{Score(f^j)} = \frac{\sum_{i=1}^{n} SF^i_{Score(f^j)}}{n} \quad (6)$$

After calculating the ensemble mean score for each feature from all lists of features, the final list of the mean ensemble FS method containing only m features that have the highest high ensemble scores is developed.

C. Multi-Label Plurality Vote Ensemble Feature Selection Method (ME-PV)

In the plurality vote ensemble [29], each FS selects its preferred list of features. These lists are used to select the candidate features. The selected features are based on the number of times they appear in the multiple lists. Once a feature is selected from a list, it is removed from the list. This process is repeated according to the number of required candidate features [29]. It should be noted that most of the votes are not required for the selection of the candidate features. Let us consider n number of data samples $S^1, \ldots\ldots, S^n$, base feature methods, $FS^1, \ldots\ldots, FS^n$, each FS method, SF^i,selects a list of m features $F^i = \{f^1, \ldots\ldots, f^m\}$ from the data sample S^i. The final score or ensemble score of a feature f^j from a list of features can be calculated using the following equations:

$$SF^i_{Score(f^j)} = \begin{cases} 1 & if \quad f^j \in F^j \\ 0 & otherwise \end{cases} \quad (7)$$

$$E_{Score(f^j)} = \frac{\sum_{i=1}^{n} SF^i_{Score(f^j)}}{n} \quad (8)$$

IV. Dynamic Multi-Label Two-Layers MI and Clustering-Based Ensemble Feature Selection Algorithm (DMMC-EFS)

This section illustrates the proposed multi-label dimension reduction technique that takes into account the 1) dynamic global weight of a feature; 2) heterogeneous ensemble; 3) maximum dependency and relevancy and minimum redundancy of the features. After the subsets of features are produced using baseline FS methods, the dynamic ensemble multi-label FS algorithm obtains a subset of useful features by combining the outputs of each method with each method in order to enhance the performance of the multi-label classification algorithm. The following figure provides the detailed steps of the proposed dynamic multi-label two layers MI and clustering-based ensemble FS methods (DMMC-EFS) (see Fig. 1).

A. Data Partitioning

Using random sampling, the dataset is partitioned into multiple samples (based on *j*). The process involves randomly shuffling the instances in order to ensure that the samples, $P_1, \ldots\ldots, P_j$, in each partition are properly balanced. Each data sample contains equal or almost equal number of instances from all the classes.

B. Baseline Feature Selection Methods Step

For each data sample or partition P_j , all the FS methods, $FM_1, \ldots\ldots, FM_k$, are applied to compute the FS values depending on the raw feature values of the sample to produce its selected feature list. The selected feature lists are sorted and passed on to the next stage. Each feature lists, FS_i^j, consists of all the features from the data sample or partition P_j, using the FS method FM_i. Specifically, each primary feature subset FS_i^jconsists of the top τ k features in P_j that are selected and sorted according to the filter-based measure values FM_i.

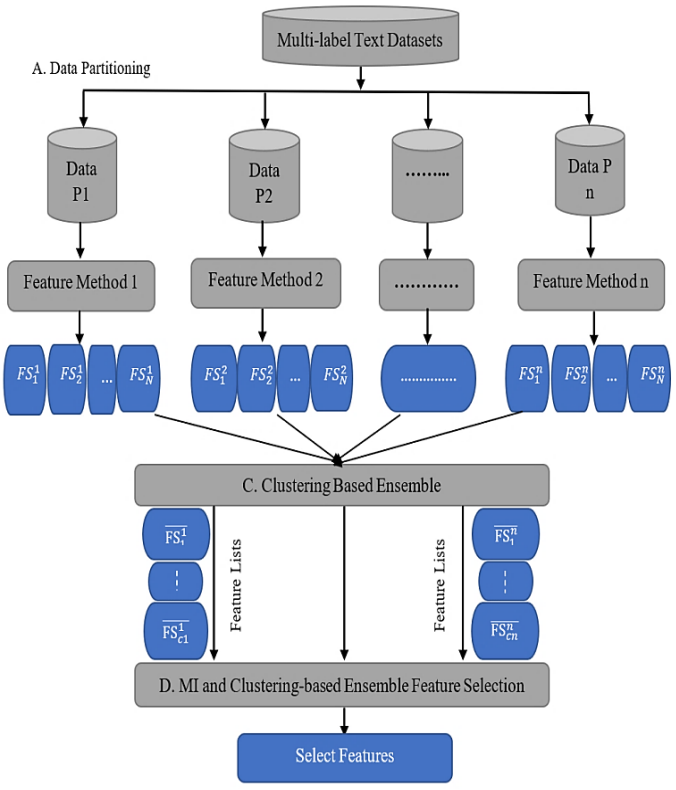


Fig. 1. The Diagram of the Proposed Dynamic Multi-Label Two Layers MI and Clustering-based Ensemble FS Methods (DMMC-EFS).

C. First Ensemble Layer (Clustering Ensemble Step)

FS methods with similar statistical and mathematical concepts may generate an alike output. If an ensemble is created by combining such similar methods, this can lead to strongly biased results. In order to avoid such bias, the FS methods that are used for an ensemble should be carefully selected. However, identifying FS methods with similar backgrounds may not be obvious, and in order to make the proposed multi-label FS ensemble more general and function well regardless of the selected baseline FS methods, we propose a graph-based clustering of group or similar ensemble intermediate FS lists that are produced using similar FS methods. With the help of the data sample or partition P_j, k intermediate feature lists are produced using the base FS methods; k intermediate feature lists are aggregated to $c < k$ feature lists, as shown in Fig. 1. The clustering step is done in order to identify and categorize the similar FS methods based on their similar outputs. Doing so should reduce the chances of allowing the similar methods to overvote the other ones, which can lead to higher diversity. The step-by-step flow of the clustering step has been summarized below:

Step1: Similarity Graph Construction: Given the data sample or partition P_j and its k intermediate feature lists that are produced by k base FS methods, $FS^j = \{\mathrm{FS}_1^j, \dots, \mathrm{FS}_k^j\}$, each output of the FS method (a feature list) over the partition P_j is represented as a node. The edge between the two nodes, E_{xy}, is computed with each pair of intermediate feature lists $(\mathrm{FS}_x^j, \mathrm{FS}_y^j)$, where the output of two baseline FS methods x and y, with the following equation:

$$E_{xy} = \cos\left(\mathrm{FS}_x^j, \mathrm{FS}_y^j\right) = \frac{\left|\mathrm{FS}_x^j\right| * \left|\mathrm{FS}_y^j\right|}{\sqrt{\left(\mathrm{FS}_x^j\right)^2 * \left(\mathrm{FS}_y^j\right)^2}} \quad (9)$$

Step2: Node Clustering Based on Edges Weights Estimation: If their edge weight is higher than a threshold t > 0.70, two node pairs is clustered together. The threshold value is measured experimentally. The resulted feature list contains the features of both the nodes. The value of the feature f is calculated using the following equation:

$$V(f) = \begin{cases} average\left(V_x(f), V_y(f)\right) & if\ f \in \text{both } \mathrm{FS}_x^j, \mathrm{FS}_y^j \\ V_x(f) & if\ f \in \mathrm{FS}_x^j \\ V_y(f) & if\ f \in \mathrm{FS}_y^j \end{cases} \quad (10)$$

Step3: Graph Reconstruction and Node Clustering Repetition: Repeat step (1) to recalculate edge weights between the clustered node and remaining graph nodes and then step (2) in order to continue clustering the nodes as long as their edge weight is higher than the threshold value. After this has been done, the k intermediate feature lists are aggregated to $c < k$. So, the output of this phase is $c_j < k_j$ and the feature lists for each partition, $P_j\ FS^j = \{\overline{\mathrm{FS}_1^j}, \dots, \overline{\mathrm{FS}_{cJ}^j}\}$.

D. Second Ensemble Layer of the Dynamic MI-based Multi-Label Feature Selection Algorithm

In the second ensemble layer, the ensemble FS method [17] takes into account the dynamic change of the selected features along with the class and dynamic global weight of the feature. In addition, the ensemble FS method measures the importance of the feature based on a criterion that has been adapted in order to maximize the dependency between the candidate feature and all class labels and minimize the conditional redundancy between the candidate feature and the selected features [36] [37] [6]. The maximal relevancy, i.e., the correlation and the minimal redundancy condition, ensures that the selected feature subset contains the most class-discerning information.

The DMM-EFS is based on a few factors which have been described as follows:

1) The first factor is the dynamic sample weight, and it considers the weight $w(\overline{FS}, h, q)$.of feature h in the feature list q.

2) The second factor is the average weight (ASW_h) of feature h from all the FS lists (samples). Based on this factor, the DMM-EFS is able to evaluate the importance of each feature in all the partitions P and all FS lists Q using the following equation:

$$ASW_h = \sum_{p \in P} \sum_{q \in Q} w(\overline{FS}, h, p, q) \quad (11)$$

3) The third factor is the size (ai) of the selected features in the sample by the FS method (i.e., ai represents the size of the features that appear in the sample). This factor is used to dynamically reduce the feature weight.

4) The fourth and last factor is the maximum sample weight of the overall samples which is used to map the feature weight $fw(i,j)$ into a higher value if the assigned value of the weight at the level of all the samples are not high.

The proposed algorithm, as shown in algorithm 1, uses the dynamic sample weight (DSW) and defines the rest of the aspects as follows: a set of samples or FS methods $D^{t \times n}$, where t is a unique set of features, while n is the number of used base FS methods (number of samples). Dynamic feature weight (DFW) is the weight of each feature in each of the base FS method or in each sample. DSW is the number of base methods that selects feature using the following equations:

$$SFWij = \frac{(fw(i,j) * ASWj * ai)}{ASW_i} \quad (12)$$

$$DSWj = \frac{DSWj}{sumSFj} \quad (13)$$

As shown in the algorithm 1, the DMM-EFS consists of several steps. With Step (i), it calculates the size of the features in each sample. Then, with Step (ii), it measures the maximum weight of the features in the sample of each sample. Following this, Step (iii) finds out the average weight of the feature j in all the samples. If feature j is selected from several samples by using FS methods, then it has various weights (different weights in different samples); this step will calculate the average weight of feature j in all the samples. With Step (iv), it calculates the overall weight of the feature in all the samples, this step computes the weight of feature j in all the FS methods (samples). (Note: For each sample, a FS method is applied to the selected feature on the basis of their weight). After that, Step (v) calculates the dynamic global weight of each feature from the weights of the feature j in all the samples. Finally,

Step (vi) selects the features based on their calculated dynamic global weight. This means that features with dynamic sample weigh greater than the threshold is selected using the ensemble algorithm.

Algorithm 1: DMM-EFS

Input: *sample feature matrix // contains the weight of each feature in each sample*

Output: *A new subset of features*

Begin

Step 1: Find the size of features in each sample

ai *=Calculate_Sample_Size (si)*

Step 2: Find the max weight in each sample samples

MSW_i *= Calculate_max_**weight** (si)*

Step 3: Find the average weight of feature j in all samples

ASW_j*=CalculateAverageWeight (feature j).*

Step 4: Calculate weight of feature in all samples

$for\ j\ = 1\ to\ t\ \ do$ *// Number of Features*
$for\ i = 1 to\ n\ do$ *// Number of Samples*
$SFWij = (fw(i,j) * ASWj\ * ai)\ /\ ASW_i)$
//SFWij is the weight of feature i in sample j
Endfor
End for

Step 5: Calculate Dynamic global weight of each feature

$for\ j\ = 1\ to\ t\ \ do\ //\ number\ of\ features$
$for\ i = 1\ to\ n\ do\ //\ number\ of\ samples$

$if\ \ feature\ \ j\ appears\ in\ the\ sample\ i\ then$
$Updating\ \ DSWj = DSWj\ +\ SFWij$

$DSWj\ =\ DSWj\ /\ sumSFj$
$Endfor$
$endfor$

Step 6: select features based on their calculated dynamic global weights

$F_{setD} = \{\}$ *// final selected features set*
$for\ j\ = 1 to\ t\ do$
$if\ DSWj >= threshold\ then$
FR_j
$\leftarrow Compute\ feature_{redundancy}(f_j, fs) with\ all\ the\ selected$
$feature\ fs\ \in\ F_{setD}$
if Calculated information is less than α for all selected features in F then
F←F∪{f}
end
$newsubsetD = newsubsetD \cup feature\ j$
End if
end for
$Return\ newsubsetD$

V. Classifcation Models

For the evaluation, two multi-label classification learning models: chain of classifier (CC), which is based on binary relevance method, and AdaBoost.MH are adopted. CC model [38] [39] can be trained independently using different datasets. This work utilizes three proven binary classifiers, namely, support vector machines (SVM) classifier, K-nearest neighbor (KNN) classifier, and Naive Bayes (NB) [40] [41]. These classifiers can be selected to construct the classifier chain. Based on different sets of domains, the training of each classifier was done independently using a data set from each domain. On the other hand, AdaBoost.MH model can adaptively adjust the weight distribution of the training samples and choose the best weak classifier out of the sample weight distribution by consistently combining all the weak classifiers, and vote by a given weight in order to build a strong classifier. AdaBoost.MH is a multi-label version of AdaBoost algorithm [42] [15]. However, these models were selected, as they have been considered as two of the high-performance state-of-the-art classification models [15] [42] [39], and they are often used to solve problems related to high dimensionality of datasets.

VI. Experimental Work

This section describes the datasets and measurements that have been used to evaluate the proposed method. The experiments were evaluated using a 5-fold cross-validation technique.

A. Multi-Label Text Dataset

Table I presents the three datasets that have been used in this work: Reuters-21578, Bibtex, and Enron and are publicly available for the multi-label text domain. In Table I, the number of features, instances, labels, cardinality, and average imbalance ratio per label (avgIR) are displayed. Cardinality measures the average number of classes for each instance, whereas density denotes the cardinality divided by the total number of labels. The datasets are available at the Mulan website (http://mulan.sourceforge.net/datasets-mlc.html).

TABLE I. Summary Description of the Multi-label Text Classification Datasets

Dataset	Instances	Features	Labels	Cardinality	avgIR
Reuters-21578	6000	500	103	1.462	54.081
Bibtex	7395	1836	159	2.402	12.498
Enron	1702	1001	53	3.378	73.953

B. Evaluation Metric

The results of the experiment on multi-label classification were measured using the following three evaluation metrics: Precision, Recall, and F measure [39] [11] [43] [44], using equations 14, 15, and16, respectively. These evaluation metrics are well-known in this domain for making comparisons.

$$\text{M\_PRECISION} = \sum_{i=1}^{d} \frac{\text{TP}_\text{i}}{\text{PT}_\text{i}+\text{FP}_\text{i}} \quad (14)$$

$$\text{M\_RECALL} = \sum_{i=1}^{d} \frac{\text{TP}_\text{i}}{\text{PT}_\text{i}+\text{FP}_\text{i}} \quad (15)$$

$$\text{M}_{\text{F}\beta} = \sum_{i=1}^{d} \frac{(\beta^2+1)\text{Pr}\times\text{Re}}{\beta^2\text{Pr}+\text{Re}} \quad (16)$$

VII. Experiment Results

This section evaluates and compares the five individual FS methods: Information Gain (IG), F-score (F), Normalized Mutual Information (NMI), Relief (R), and Mutual Information (MI) and the multi-label ensemble FS methods: multi-label mean-based ensemble feature (ME-mean) selection method and multi-label plurality vote ensemble FS method (ME-PV) with our proposed method, dynamic multi-label two layers MI and clustering-based ensemble FS method (DMMC-EFS).

Three experiments are conducted: The first experiment (Experiment I) is conducted using conventional and ensemble FS methods on Reuters-21578 corpus; the second, experiment (Experiment II) is conducted using conventional and ensemble FS methods on Bibtex corpus; and the third experiment (Experiment III) is conducted using conventional and ensemble FS methods on Enron corpus.

A. Experiment I: Evaluation of the Proposed Ensemble FS Method and the Conventional FS Methods on Reuters-21578 Corpus

This subsection evaluates five state-of-the-art conventional multi-label FS methods (FSMs): IG, F, NMI, R, and MI and three multi-label ensemble FS methods: ME-mean, ME-PV and DMMC-EFS. The effect of these methods is studied using two classification models: CC model, which combines three classifiers (SVM, KNN and NB), and AdaBoost.MH.

All experiments in this subsection are conducted on Reuters-21578 corpus benchmark dataset. The macro-averaging *F*-measure of the CC and AdaBoost.MH with the eight FS methods (FSMs) are displayed in Table II.

With a focus only on the conventional multi-label FS methods, both NMI and MI multi-label FS methods achieve the best performance with all the classifiers. The performance of the two conventional FS methods: IG and F-score is below average. The main reason is that both NMI and MI multi-label FS methods use feature-class mutual information to select relevant features.

With a focus on both multi-label ensemble and conventional FS methods, the results from all the multi-label ensemble methods are better than the results that are obtained using the conventional FS methods. DMMC-EFS multi-label ensemble method achieves the highest performance in terms of macro-averaging F-measure outperforming both the multi-label ensemble and conventional methods. As mentioned in Section 4, DMMC-EFS considers the feature-class and feature-feature correlation in order to select relevant and non-redundant features and also the dynamic change of the selected features along with the class and dynamic global weight of the feature.

A range of 71–90% was achieved by all the classifiers. AdaBoost.MH displays higher classification performance than the CC on the Reuters-21578 corpus in terms of all the multi-label ensemble and conventional FS methods. This may be due to the fact that AdaBoost.MH model produces alternating decision trees that can handle multi-label data.

In general, all the classification models with all the multi-label ensemble FS methods (ME-mean, ME-PV, and DMMC-EFS) achieve good results in terms of prediction performance on the Reuters-21578 (a high dimensional dataset) corpus. This is expected as the ensemble FS methods exploits the several FS methods by combining their strengths.

B. Experiment II: Evaluation of the Proposed Ensemble FS and the Conventional FS Methods on Bibtex Corpus

This subsection evaluates five state-of-the-art conventional multi-label FS methods: IG, F, NMI, R, and MI and three multi-label ensemble FS methods: ME-mean, ME-PV, and DMMC-EFS. The effect of these methods is studied using two classification models: CC model, which combines three classifiers (SVM, KNN and NB) and AdaBoost.MH. All the experiments in this subsection have been conducted on Bibtex corpus benchmark dataset. The macro-averaging F-measure of the CC and AdaBoost.MH along with the eight FSM selection methods are shown in Table III.

With a focus only on the conventional multi-label FS methods, unlike experiments on Reuters-21578 corpus, experiments on Bibtex corpus show that R multi-label FS method achieves the best performance among all the conventional FS methods irrespective of the classifier used. R, as a feature evaluation measure, more often selects smaller number of features than the other FS methods, without degrading the performance of the classifiers. This could be due to the fact that R considers interactions among the features [27].

With a focus only on the multi-label ensemble FS methods, DMMC-EFS multi-label ensemble method achieves the best performance with all the classifiers. As mentioned in Section 4, DMMC-EFS takes into account the feature-class and feature-feature interaction in order to select the relevant and non-redundant features and the dynamic change of selected features along with the class and dynamic global weight of the feature. Its results on Bibtex corpus is slightly higher than its results on Reuters-21578 corpus. This is due to the fact that Reuters-21578 dataset has higher dimensionality than Bibtex corpus.

TABLE II. PERFORMANCE (F-MEASURE) OF CC AND ADABOOST.MH WITH ALL MULTI-LABEL ENSEMBLE AND CONVENTIONAL FS METHODS ON REUTERS-21578

Feature Selection method	AdaBoost.MH	CC
IG	74.13	71.67
R	81.1	79.54
NMI	86.01	81.6
MI	85.23	80.9
F	77.17	76.36
ME-mean	86.62	82.62
ME-PV	86.58	82.99
DMMC-EFS	89.96	87.41

TABLE III. PERFORMANCE (F-MEASURE) OF CC AND ADABOOST.MH WITH ALL MULTI-LABEL ENSEMBLE AND CONVENTIONAL FS METHODS ON BIBTEX CORPUS

Feature Selection Method	AdaBoost.MH	CC
IG	75.25	70.85
R	86.03	83.28
NMI	84.41	80.2
MI	84.32	80.82
F	73.84	70.13
ME-mean	84.42	80.52
ME-PV	83.7	80.39
DMMC-EFS	91.31	88.81

With a focus on both the multi-label ensemble and conventional FS methods, DMMC-EFS multi-label ensemble methods achieve better results than those obtained using all the conventional FS methods. R multi-label FS method attain better performance than ME-mean and ME-PV multi-label ensemble FS methods. The ME-mean and ME-PV ensemble FS methods, which use NMI and MI, dominate other methods when they are combined, and the final feature list are strongly biased toward their choice. So, the ME-mean and ME-PV ensemble FS methods act as a single MI FS method. However, in terms of macro-averaging F-measure, the DMMC-EFS multi-label ensemble method performs the best among the other multi-label ensemble and conventional methods. As it has been mentioned above, DMMC-EFS takes into account feature-class and feature-feature interaction in order to select relevant and non-redundant features and the dynamic change of selected features along with the class and dynamic global weight of the feature.

A range of 70– 91% of performance is achieved by all the classifiers. AdaBoost.MH gave higher classification performance than the CC on Bibtex corpus with all the multi-label ensemble and conventional FS methods. This may be due to the fact that AdaBoost.MH model produces alternating decision trees that can handle multi-label datasets. In general, all the classification models (CC and AdaBoost.MH) with the multi-label ensemble FS methods (ME-mean, ME-PV and DMMC-EFS) achieve good results (between 80% and 91%) in prediction performance on Bibtex corpus' high dimensional datasets.

C. Experiment III: Evaluation of the Proposed Ensemble FS and Conventional FS Methods on Enron Corpus

This subsection examines five state-of-the-art conventional multi-label FS methods: IG, F, NMI, R, and MI and three multi-label ensemble FS methods: ME-mean, ME-PV, and the proposed method (DMMC-EFS). The effect of these methods is studied using two classification models: CC model, which combines three classifiers (SVM, KNN, and NB), and AdaBoost.MH. All the experiments in this subsection are conducted on Enron corpus benchmark dataset. The macro-averaging F-measure of the CC and AdaBoost.MH with the eight FSM selection methods are presented in Table IV.

Considering only the conventional multi-label FS methods, similar to the results on the Reuters-21578 corpus dataset (as shown in Table II), both NMI and MI multi-label conventional FS methods achieve the best performance on Enron corpus among all the conventional FS methods regardless of which classifier has been used. Considering only the multi-label ensemble FS methods, DMMC-EFS multi-label ensemble method achieves the best performance with all the classifiers. On both the multi-label ensemble and conventional FS methods, both DMMC-EFS and E-mean multi-label ensemble methods achieve performances higher than that of all the conventional FS methods. Both NMI and MI multi-label conventional FS methods attain better performance than ME-PV multi-label ensemble FS methods. However, DMMC-EFS multi-label ensemble method achieves the highest performance in terms of macro-averaging F-measure among both the multi-label ensemble and conventional methods.

TABLE IV. PERFORMANCE (F-MEASURE) OF CC AND ADABOOST.MH WITH ALL MULTI-LABEL ENSEMBLE AND CONVENTIONAL FS METHODS ON ENRON CORPUS

Feature Selection method	AdaBoost.MH	CC
IG	78.83	74.82
R	85.19	82.67
NMI	86.41	83.34
MI	86.7	84.2
F	77.7	73.88
ME-mean	87.27	83.37
ME-PV	86.11	84.57
DMMC-EFS	91.79	89.54

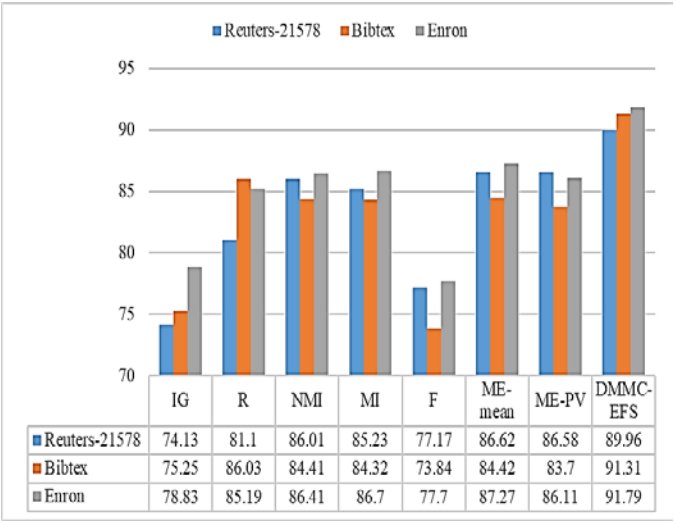
A range of 73–92% of performance is achieved by all the classifiers. In general, all the classification models with multi-label ensemble FS methods achieve good results between 84% and 91% of F-measure in the prediction performance on Enron corpus' high dimensional datasets.

VIII. RESULTS DISCUSSION

In order to evaluate and compare the performance of the conventional multi-label FS methods, ensemble FS methods and our proposed DMMC-EFS on all the data sets, the obtained results are presented in Fig. 2.

It can be observed in Fig. 2 that the F-measure is within the range of 86–91.7%; this demonstrates that the multi-label text classification models can be improved if the inherited high dimensionality problem is reduced. Fig. 2 also validates the stability of the proposed DMMC-EFS and the conventional methods. Stability is defined as the ability to behave the same way regardless of what dataset is being used. Since the complexity of the datasets varies, and they are derived from different sources, the stability could be found to be different for all the conventional FS methods. The proposed DMMC-EFS multi-label ensemble method is the most stable method, as it always achieves the top rank, and their results on all datasets is mostly consistent, which means that. It outperforms all the other methods on all the datasets, and the difference in the values of its performance from one dataset to another is minute.

By comparing the behavior of both the ME-mean and ME-PV multi-label ensemble FS methods with the behavior of both the NMI and MI multi-label conventional FS methods using the data presented in Tables II–IV, we notice that the ME-mean and ME-PV behave like the conventional FS methods and their results are effected by NMI and MI. As stated in [29], if the similar FS methods are combined, the result will be strongly biased towards their aggregated outputs. In fact, NMI and MI have similar underlying concepts, which means that they are derived from the same mathematical and statistical concepts. Therefore, they tend to produce similar outputs. NMI and MI [29] dominate other methods when they are combined, and the final feature list are strongly biased toward their output. This supports our hypothesis that similar methods outputs should be clustered together, so as to ensure that they have less chance to overvote the other methods which eventually widen the output diversity.



	IG	R	NMI	MI	F	ME-mean	ME-PV	DMMC-EFS
Reuters-21578	74.13	81.1	86.01	85.23	77.17	86.62	86.58	89.96
Bibtex	75.25	86.03	84.41	84.32	73.84	84.42	83.7	91.31
Enron	78.83	85.19	86.41	86.7	77.7	87.27	86.11	91.79

Fig. 2. Performance of All Multi-Label Ensemble and Conventional FS Methods on All Data Sets.

IX. Conclusion

This paper presents a scalable multi-label classification method that can handle the high dimensionality problem of the multi-label datasets. Firstly, this paper investigates the several state-of-the-art conventional multi-label FS methods. In addition, this paper proposes two multi-label ensemble FS methods: multi-label mean ensemble (ME-mean) FS method and multi-label plurality vote ensemble (ME-PV) FS method. Finally, this paper proposes a new dynamic multi-label two layers MI and clustering-based ensemble FS (DMMC-EFS) method that takes into account the 1) dynamic global weight of the feature; 2) heterogeneous ensemble 3) maximum dependency and relevancy and minimum redundancy of the features. The results show that the proposed multi-label FS methods significantly outperformed the other state-of-the-art conventional and ensemble multi-label FS methods. To conclude, it can be stated that an enhanced ensemble FS method, which takes into account the dynamic global weight of the feature, heterogeneous ensemble, and max dependency and relevancy and minimum redundancy of the features, can overcome the high dimensionality of the multi-label datasets and improve the performance of the multi-label text classification system. In future, it is recommended to extend the proposed method by adding more sophisticated feature selection methods to it. Additionally, it is also recommended to examine the performance of DMM-EFS method using different languages and datasets.

Acknowledgment

The authors would like to thank Universiti Kebangsaan Malaysia for supporting this project, grant codes: GGPM-2017-040 and PP-FTSM-2020.

References

[1] Xie, Y., Li, D., Zhang, D., & Shuang, H. (2017, June). An Improved Multi-label Relief Feature Selection Algorithm for Unbalanced Datasets. In International Conference on Intelligent and Interactive Systems and Applications (pp. 141-151). Springer, Cham.

[2] Alshalabi, H., Tiun, S., Omar, N., & Albared, M. (2013). Experiments on the use of feature selection and machine learning methods in automatic malay text categorization. Procedia Technology, 11(1), 748-754.

[3] Kashef S, Nezamabadi-pour H, An effective method of multi-label feature selection employing evolutionary algorithms. Swarm Intelligence and Evolutionary Compu-tation (CSIEC), 2017 2nd Conference on; 2017: IEEE.

[4] Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In 2014 Science and Information Conference (pp. 372-378). IEEE.

[5] Spolaôr, N., Cherman, E. A., Monard, M. C., & Lee, H. D. (2012, October). Filter approach feature selection methods to support multi-label learning based on relieff and information gain. In Brazilian Symposium on Artificial Intelligence (pp. 72-81). Springer, Berlin, Heidelberg.

[6] Li, F., D. Miao & W. Pedrycz 2017. Granular multi-label feature selection based on mutual information. Pattern Recognition 67: 410-423

[7] Zhang, L., & Duan, Q. (2019). A Feature Selection Method for Multi-Label Text Based on Feature Importance. Applied Sciences, 9(4), 665.

[8] Gharroudi, O., Elghazel, H., & Aussem, A. (2014, May). A comparison of multi-label feature selection methods using the random forest paradigm. In Canadian conference on artificial intelligence (pp. 95-106). Springer, Cham.

[9] Alhutaish, R., Omar, N., & Abdullah, S. (2015, November). A comparison of multi-label feature selection methods using the algorithm adaptation approach. In International Visual Informatics Conference (pp. 199-212). Springer, Cham.

[10] Lee, J., & Kim, D.-W. (2015a). Fast multi-label feature selection based on information-theoretic feature ranking.Pattern Recognition, 48(9), 2761–2771.

[11] Zhou, H., Zhang, Y., Liu, H., & Zhang, Y. (2018). Feature Selection Based on Term Frequency Reordering of Document Level. IEEE Access, 6, 51655-51668.

[12] Asaithambi, S. (2018). Why, How and When to apply Feature Selection.

[13] Liang, Y., Niu, D., & Hong, W. C. (2019). Short term load forecasting based on feature extraction and improved general regression neural network model. Energy, 166, 653-663.

[14] Seijo-Pardo, B., Bolón-Canedo, V., Porto-Díaz, I., & Alonso-Betanzos, A. (2015, June). Ensemble feature selection for rankings of features. In International Work-Conference on Artificial Neural Networks (pp. 29-42). Springer, Cham.

[15] Al-Salemi, B., Ayob, M., & Noah, S. A. M. (2018). Feature ranking for enhancing boosting-based multi-label text categorization. Expert Systems with Applications, 113, 531-543.

[16] Ndirangu, D., Mwangi, W., & Nderu, L. (2019). An Ensemble Model for Multi-label Classification and Outlier Detection Method in Data Mining.

[17] Gao, W., Hu, L., & Zhang, P. (2018). Class-specific mutual information variation for feature selection. Pattern Recognition, 79, 328-339.

[18] Hoque, N., Singh, M., & Bhattacharyya, D. K. (2018). EFS-MI: an ensemble feature selection method for classification. Complex & Intelligent Systems, 4(2), 105-118.

[19] Li, S., Zhang, Z., & Duan, J. (2014). An ensemble multi-label feature selection algorithm based on information entropy. Int. Arab J. Inf. Technol., 11(4), 379-386.

[20] Uysal, A. K. (2016). An improved global feature selection scheme for text classification. Expert systems with Applications, 43, 82-92.

[21] Agnihotri, D., Verma, K., & Tripathi, P. (2017, November). Mutual information using sample variance for text feature selection. In Proceedings of the 3rd International Conference on Communication and Information Processing (pp. 39-44).

[22] Clare, A., & King, R. D. (2001, September). Knowledge discovery in multi-label phenotype data. In european conference on principles of data mining and knowledge discovery (pp. 42-53). Springer, Berlin, Heidelberg.

[23] Pereira, R. B., Carvalho, A. P. D., Zadrozny, B., & Merschmann, L. H. D. C. (2015). Information gain feature selection for multi-label classification.

[24] Li, L., Liu, H., Ma, Z., Mo, Y., Duan, Z., Zhou, J., & Zhao, J. (2014, December). Multi-label feature selection via information gain. In International Conference on Advanced Data Mining and Applications (pp. 345-355). Springer, Cham.

[25] Pes, B. (2019). Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Computing and Applications, 1-23.

[26] Wang, J., Xu, J., Zhao, C., Peng, Y., & Wang, H. (2019). An ensemble feature selection method for high-dimensional data based on sort aggregation. Systems Science & Control Engineering, 7(2), 32-39.

[27] Spolaôr, N., Lee, H. D., Takaki, W. S. R., & Wu, F. C. (2015). Feature selection for multi-label learning: A systematic literature review and some experimental evaluations. International Journal of Computational Intelligence Systems, 8(sup2), 3-15.

[28] Jungjit, S. (2016). New Multi-Label Correlation-Based Feature Selection Methods for Multi-Label Classification and Application in Bioinformatics (Doctoral dissertation, University of Kent,).

[29] Drotár, P., Gazda, M., & Vokorokos, L. (2019). Ensemble feature selection using election methods and ranker clustering. Information Sciences, 480, 365-380.

[30] Reyes, O., Morell, C., & Ventura, S. (2015). Scalable extensions of the ReliefF algorithm for weighting and selecting features on the multi-label learning context.Neurocomputing, 161, 168–182.

[31] Ullah, I., & Mahmoud, Q. H. (2017, December). A filter-based feature selection model for anomaly-based intrusion detection systems. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 2151-2159). IEEE.

[32] Al-Salemi, B., Ayob, M., Noah, S. A. M., & Ab Aziz, M. J. (2017, November). Feature selection based on supervised topic modeling for boosting-based multi-label text categorization. In 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI) (pp. 1-6). IEEE.

[33] Xu, J. & Q. Ma 2018. Multi-label regularized quadratic programming feature selection algorithm with Frank–Wolfe method. Expert Systems with Applications 95: 14-31.

[34] Lim, H., Lee, J., & Kim, D. W. (2017). Optimization approach for feature selection in multi-label classification. Pattern Recognition Letters, 89, 25-30.

[35] Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In Mining text data (pp. 163-222). Springer, Boston, MA.

[36] Lee, J., Yu, I., Park, J., & Kim, D. W. (2019). Memetic feature selection for multi-label text categorization using label frequency difference. Information Sciences, 485, 263-280.

[37] González-López, J., Ventura, S., & Cano, A. (2019). Distributed Selection of Continuous Features in Multi-label Classification Using Mutual Information. IEEE transactions on neural networks and learning systems.

[38] Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009, September). Classifier chains for multi-label classification. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 254-269). Springer, Berlin, Heidelberg.

[39] Taha, A. Y., & Tiun, S. (2016). Binary Relevance (BR) Method Classifier of Multi-Label Classification for Arabic Text. Journal of Theoretical & Applied Information Technology, 84(3).

[40] Abdulameer, A. S., Tiun, S., Sani, N. S., Ayob, M., & Taha, A. Y. (2020). Enhanced clustering models with wiki-based k-nearest neighbors-based representation for web search result clustering. Journal of King Saud University-Computer and Information Sciences.

[41] Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36-54.

[42] Pant, P., Sabitha, A. S., Choudhury, T., & Dhingra, P. (2019). Multi-label classification trending challenges and approaches. In Emerging Trends in Expert Applications and Security (pp. 433-444). Springer, Singapore.

[43] Omar, N., Albared, M., Al-Moslmi, T., & Al-Shabi, A. (2014, December). A comparative study of feature selection and machine learning algorithms for Arabic sentiment classification. In Asia information retrieval symposium (pp. 429-443). Springer, Cham.

[44] Abdulameer, A. S., Saad, S., & Zakaria, L. Q. (2015). Trend detection in the arabic social media using voting combination. Journal of Theoretical and Applied Information Technology, 81(3), 432-443.