Hybrid Support Vector Machine based Feature Selection Method for Text Classification

Thabit Sabbah[1], Mosab Ayyash[1], and Mahmood Ashraf[2]
[1]Faculty of Technology and Applied Sciences, Al-Quds Open University, Palestine
[2]Department of Computer Science, Federal Urdu University, Pakistan

Abstract: *Automatic text classification is an effective solution used to sort out the increasing amount of online textual content. However, high dimensionality is a considerable impediment observed in the text classification field in spite of the fact that there have been many statistical methods available to address this issue. Still, none of these has proved to be effective enough in solving this problem. This paper proposes a machine learning based feature ranking and selection method named Support Vector Machine based Feature Ranking Method (SVM-FRM). The proposed method utilizes Support Vector Machine (SVM) learning algorithm for weighting and selecting the significant features in order to obtain better classification performance. Later on, hybridization techniques are applied to enhance the performance of SVM-FRM method in some experimental situations. The proposed SVM-FRM method and its enhancement are tested using three text classification public datasets. The achieved results are compared with other statistical feature selection methods currently used for the said purpose. Results evaluation shows higher and superior F-measure and accuracy performances of the proposed SVM-FRM on balanced datasets. Moreover, a noticeable performance enhancement is recorded due to the application of the proposed hybridization techniques on an unbalanced dataset.*

Keywords: *Feature ranking, text classification, feature selection, SVM-based weighting, hybridization, dimensionality reduction.*

Received February 12, 2018; accepted April 22, 2018

1. Introduction

Machine learning-based text categorization and classification techniques are effective and preferable solutions for the rapidly increasing amount of online textual contents [11]. The Text Classification (TC) refers to assigning a document to one of the predefined classes or categories [16]. Text representation and feature selection are among the fundamental steps in TC that enable the classification algorithms to deal with textual content [20] and to reduce the dimensionality of feature space.

The Vector Space Model (VSM) has proved to be an effective text representation method in several text processing related domains including summarization [5] and categorization [21]. In VSM model, documents are represented vectors of weighted features, where the features and the weighting methods can be of different types [26]. In spite of the available feature types and weighting methods, the huge dimensionality of feature space is a major problem that should be reduced to decrease the computational complexity, increase the classification algorithms performance, and reduce the required resources for data processing [28]. Therefore, many dimensionality reduction methods were proposed in last several decades.

The Feature Selection (FS) methods are part of the dimensionality reduction methods that aim at downsizing the dimensionality of feature space. When applying an FS method, the most informative features are selected while the less important and uninformative features are eliminated, based on the assumption that removing such features will not significantly affect the quality of the classification [12]. However, selecting the most informative features involves the process of weighting and ranking all features in the feature space. In general, this process is based on the statistical analysis of feature space that analyses the intrinsic characteristics of the document [8] or the corpus [26].

In TC domain, many methods for feature weighting and ranking have been used frequently such as Term Frequency-Inverse Document Frequency (TF-IDF), Term Frequency (TF), Term Frequency-Relevance Frequency (TFRF), Document Frequency (DF), Chi-square (CHI), Entropy, Inverse Document Frequency (IDF), Information Gain (IG), and Correlation [27]. Based on the literature available in this area, various less common methods have also been proposed for feature weighting and ranking [27]. These ranking methods depend on the statistical analysis of feature space.

Therefore, this paper proposes a feature ranking and selection method in which the weights are computed based on a learning algorithm. The proposed method is based on the assumption that the more informative and important the features are, the higher the weights are assigned by the learning

algorithm. As such, the feature selection based on these weights will eventually lead to higher classification performance.

This paper consists of the following sections in addition to this introductory section. Section 2 presents the concepts of some statistical based dimensionality reduction methods that are widely used in TC domain. Additionally, it explains the basics of the SVM learning algorithm which is utilized in the method proposed in this paper. Section 3provides a detailed description of the proposed Support Vector Machine based Feature Ranking Method (SVM-FRM) and its hybridization-based enhancement. Section 4 highlights the used datasets and the conducted experiments. Section 5 presents the obtained results along with a discussion on the major findings. Section 6 provides the conclusion on this paper and suggests headlines for future studies to be conducted by the research team.

2. Related Works

This section presents major concepts of the dimensionality reduction methods based on feature ranking by the statistical analysis of IG, Correlation, Chi-square, and the SVM learning algorithm. Although, there are many feature reduction (i.e., selection) method in TC classification presented through years (see [13]), none of these methods has been proofed to be superior, and several of these methods showed poor performance, therefore, based on our initial exploration and experiments the methods IG, Correlation, and Chi-square were selected as it showed some competitive results to the proposed method.

2.1. Information Gain (IG)

IG is among the most commonly applied feature selection methods [18]. It is a statistic that measures the goodness of an attribute (i.e., feature). As previously referred to, feature reduction methods aim at determining and applying the most useful attributes for distinguishing the different classes of a given feature space. Therefore, IG measure can indicate how important each of the attributes is, by calculating the weight (relevance) of an attribute in terms of the class attributes. The higher the weight of an attribute, the more distinguished it is considered to be.

The IG of a feature *f* is defined as the information gained by doing the split of the feature space based on that particular feature, which is mathematically expressed as follows [29]:

$$IG(F) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(f)\sum_{i=}^{m} P_r(c_i|f)\log P_r(c_i|f) + P_r(\bar{f})\sum_{i=}^{m} P_r(c_i|\bar{f})\log P_r(c_i|\bar{f}) \quad (1)$$

Where m is the number of categories, $P(c_i)$ is the probability of category c_i, $Pr(f)$ and $Pr(\bar{f})$ are the probabilities of presence and absence of feature f, $P(c_i|f)$ and $P(c_i|\bar{f})$ are the conditional probabilities of category c_i considering occurrence and nonappearance of feature f, respectively.

Although IG is a good measure for an attribute's relevance, it has lower performance when it is applied to attributes that can take a large number of distinct values. More details on IG can be found in [13].

2.2. Correlation

Correlation statistic is used to measure the linear association (correlation) between two attributes (i.e., features), where attributes of higher correlation weight are considered to be more relevant. A correlation is defined as a number ranging from -1 to +1 that represents the degree of association between two attributes (let these attributes be X and Y). A positive association between X and Y is represented by a positive value for the correlation while a negative correlation value implies an inverse or negative association [14]. The correlation of two attribute vectors X and Y is defined as follows:

$$Correlation(X,Y) = \frac{\sum_{i=1}^{n}(X(i)-\bar{X}).(Y(i)-\bar{Y})}{(n-1).\sigma(X).\sigma(Y)} \quad (2)$$

Where n is the number of samples (i.e., document). $\overline{X}$, $\sigma(X)$ and $\overline{Y}$, $\sigma(Y)$ are the means and standard deviations of X and Y, respectively.

However, using correlation for feature selection involves finding a subset of features in which the features are correlated as less as possible among each other. Besides, each of these the features, has to be correlated with classes vector as much as possible. Usually, correlation based feature selection is based on heuristic search strategies to find the appropriate feature subset in a reasonable period of time [23].Therefore, the metaheuristic (M) of the selected subset is usually approximated based on the Equation:

$$Ms = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (3)$$

Where k is the number of features contained in the subset of features S, while $\overline{r_{cf}}$ and $\overline{r_{ff}}$ are the mean of feature-class correlation and feature-feature inter-correlation, respectively [23].

2.3. Chi-Square

Similar to the IG, chi-square is a nonparametric statistical technique used to compute the lack of independence between the distributions of observed frequencies and the theoretically expected frequencies [30], where the higher the weight of an attribute, the more relevant it is. In general, chi-square statistics use nominal data, however, in TC domain it uses feature's frequencies instead of using means and variances. The value of the chi-square statistic is given by Equation (4):

$$\chi^2 = Sigma\left[\frac{(O-E)^2}{E}\right] \quad (4)$$

Where chi-square statistic is noted as x^2, O is the observed frequency, and E is the expected frequency. More details on chi-square in TC domain are provided in [10].

2.4. Support Vector Machine

SVM is one of the most effective and popular supervised learning algorithms, in which it depends on learning from a training set to find a hyper-plane that can separate the cases of binary classes [22]. The hyper-plane is located at the point in the hyper-space that maximizes the distance between the support vectors which are the closest positive and negative samples. Two components play a vital role in Linear SVM; one is a weight vector $\ddot{W}$ which is perpendicular to the hyper-plane, while the other is the bias $\dot{b}$ which is the offset of hyper-plane from the origin. The class of an unlabelled example $\dot{X}$ is determined by calculating the value $f(\dot{x})$, where $f(\dot{x}) = \ddot{W}\dot{X} + \dot{b}$. If the computed value of $f(\dot{x})$ is greater than or equal to zero, the example is classified as positive, otherwise, it is classified as negative.

SVM algorithm has many advantages that make it preferable among other classification tools. Among which is the ability to handle extremely large feature spaces besides the well-handling of high dimensional feature vectors and redundant features which are the features that can be predicted from others [17]. SVM has also been proved to be among the best performing machine learning approaches [19] in various domains including text classification. Although, SVM is an effective binary classifier that has been utilized by many existing projects as text classifier, it can be applied for multi-label classification problems. For example, [7] presents a comprehensive empirical comparison study in which many different SVM algorithms were tested on various publicly available text classification datasets.

In this paper, the research team utilizes the weight vector $\ddot{W}$ generated by SVM learning algorithm for feature selection based on the assumption that for each w_i represents the contribution and importance of feature f_i to the separation hyper-plane.

The use of SVM as feature ranking method is not a new method, as it has been discussed earlier in [9] and recently in [31]. However, none of the earlier works examine the method in TC domain where the problem of classification is known to be of a high dimensional feature space. For instance, in [9] the method was proposed for Causality Challenge where the training and testing sets might have different distributions, and it was tested on multiple datasets where the maximum number of features is less than 5000 binary features. Therefore, this work utilizes the SVM as feature ranking method in TC and examines its performance using three different Arabic TC datasets. Additionally, this work presents a hybridized method in which SVM method is hybridized with two feature ranking methods to achieve higher classification performance. Next sections discuss in detail the steps of applying SVM as feature ranking method and the proposed hybridized method.

3. Support Vector Machine based Feature Ranking Method

This paper presents a SVM-FRM. This method utilizes the SVM learning algorithm in order to assign weight values to the features in the feature space. Then, these weights are used as ranking criteria to select the features of highest weights for classification. The proposed method is the core part of the general text classification approach that usually consists of three steps. The major three steps of the applied approach are summarized as follows:

1. The Term Frequency-Inverse Document Frequency (TFIDF) weighting method is used in VSM text representation.
2. The SVM-FRM is applied for feature ranking.
3. The top K effective features are used in the classification process.

Figure 1 shows the steps of the applied approach.

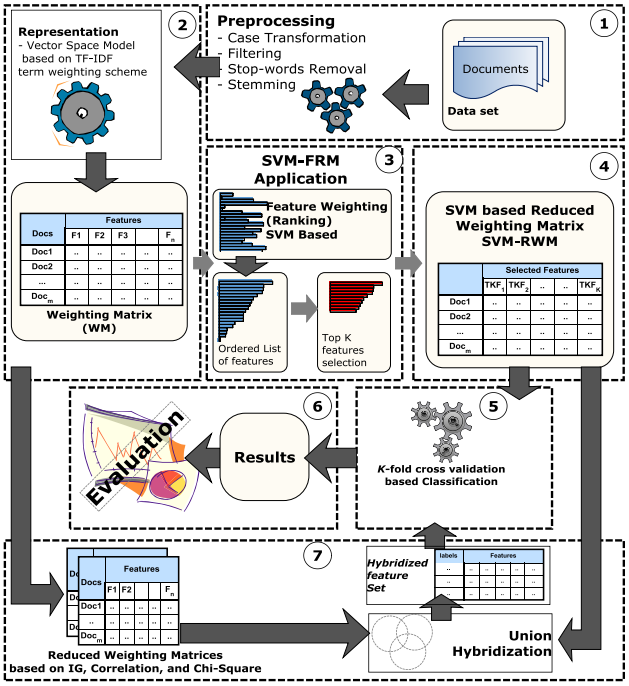

Figure 1. Steps of text classification approach.

The applied approach as seen in Figure 1 consists of six steps, sections 3.1 to 3.6 describe these steps in detail.

3.1. Pre-Processing

This step includes the application of case transformation, filtering, stop-words removal and stemming methods. Filtering includes eliminating the non-words tokens from the text such as numbers, Latin words, and Html tags. In this research, filtering also removes from the text the words of less than four or more than fifteen characters in length. Stop-words removal process usually removes the meaningless tokens from the text. The default stop-words list for the Arabic language included in Rapid Miner Studio v7.5 was the one referred to for the purposes of this research. Stemming is the process of reducing inflected words to their word stem. It should be said that the simple form of stemming is to treat related words as synonyms of the same stem when even this stem may not be a valid root. The Arabic Light stemmer is referred to in this step in this research.

3.2. Text Representation

The corpus, in this step, is represented based on the VSM in which the different term weighting formula can be considered for document vector creation [27]. This research considers the Term Frequency-Inverse Document Frequency (TF-IDF) weighting formula because of its popularity and efficiency in the domain of text classification. The result of this step is a matrix $M_{d \times f}$, where d is the number of documents in the corpus and f is size of feature space (i.e., the number of features), the entry w_{ij} of the matrix represents the weights of the feature j in the document i. In this research, features are considered as the unigram token basis (i.e., each single word is considered as one feature).

3.3. SVM-FRM Application

The main contribution of this research is focused in this step, where the SVM learning algorithm is utilized to rank the features of the feature space. The SVM algorithm assigns a weight value to each feature as a result of the training process. In normal usage of SVM as a classifier, these weights help the SVM to learn the hyper-plane that separates positive examples from negative examples of the dataset. However, in this research, these weights are employed as a ranking method of the features. The features ranked high are assumed to be more distinguishable, hence lead to better classification results. The output of applying SVM for feature ranking will produce a weight assigned to each feature in the unordered list of features. Thus, the list of features will be sorted in descending order according to the value of the weight. Then, Top K features will be selected to be used in next step.

3.4. Reduced Feature Space Construction

After selecting the Top K features, the reduced feature space based on these features should be established before the classification. The construction of the reduced feature space is performed using the algorithm shown in Algorithm 1.

Algorithm 1: Reduced feature space construction algorithm

Input:
Feature Space FS_{mxn} / m is the count of rows, and n is the count of columns.*
List of Top K features.
Output:
Reduced Feature Space RFS_{mxk}, / m is the count of rows, and k is the count of selected features (columns).*/*
Start
For each feature in FS
If feature is among the Top K features
Include feature in Reduced Feature Space
End If
Loop
End

3.5. Classification

In this step, the constructed reduced feature space is passed to the classification algorithm. Usually, the feature space can be treated into two ways. One is splitting it into two parts known as Training and Testing parts. The training part is used to train the classifier to construct the classification model, while

the testing part is used to measure the performance of the constructed model.

The other way is to split the matrix into *K* equal (or approximately equal) parts known as folds (usually 10 folds). Then, the classification training and testing processes are performed *K* rounds. In each round, one fold is considered as Testing, while the remaining *K-1* folds are used for Training. In this case, the performance is calculated by averaging of the performances obtained from all rounds. This research follows the second way and applies the classification based on the stratified 10-folds cross validation model [33] using the SVM classifier. In stratified splitting each fold will contain samples from all classes in proportions that are equal to the classes' proportions in the dataset.

3.6. Evaluation

Commonly, in TC domain, the metrics such as Precision, Recall, F-measure, and Accuracy are used to evaluate the performance of approaches. Thus, these are quite helpful in providing an overall performance evaluation of the presented classification approach. However, literature review shows that high precision and recall values are hard to be achieved simultaneously as low values of recall may be the price of obtaining high levels of precision and vice versa [20]. This research has considered the Accuracy metric in addition to averaged F-measure metric as the weighted harmonic mean of precision and recall for evaluation. Generally, text classification or categorization is a multiple class classification problem, in which the Precision, Recall, and F-measure metrics are calculated per class using the Equations 5, 6, 7 and 8.

$$Pci = |TPci| / [|TPci| + \dashv |FPci| \dashv] \quad (5)$$

$$Pci = |TPci| / [|TPci| + \dashv |FPci| \dashv] \quad (6)$$

$$F - measureci = 2 * [((Pci * Rci))/(Pci + Rci)] \quad (7)$$

$$Accuracyci = \frac{|TPci| + |TNci|}{|TPci| + |FPci| + |FNci| + |TNci|} \quad (8)$$

Where Pc_i, Rc_i are the Precision and Recall of class c_i, respectively. TPc_i is the count of documents correctly labelled to be in class c_i, and FPc_i is the number of documents incorrectly labelled by the classifier to be in class c_i. FNc_i is the number of documents incorrectly identified not to be in class c_i, and TNc_i is the number of documents correctly labelled not to be in class c_i. In spite of the fact that text classification is usually considered as a multi-class classification problem, the averaged F-measure is calculated in this research as Equation 9 shows, where *n* is the number of classes (i.e., categories) in the dataset.

$$Averaged\ F - measure = 2 * \left[\frac{\left(\frac{\sum_{i=1}^{n} P_{ci}}{n}\right) * \left(\frac{\sum_{i=1}^{n} R_{ci}}{n}\right)}{\left(\frac{\sum_{i=1}^{n} P_{ci}}{n}\right) + \left(\frac{\sum_{i=1}^{n} R_{ci}}{n}\right)}\right] \quad (9)$$

3.7. Hybridization

This step is an extension of the basic approach in which different feature selection methods are hybridized for further performance enhancement, the details of this step are discussed in section 6.

4. Datasets and Experiments

This section discusses the used datasets and the conducted experiments.

4.1. Datasets

In order to evaluate the performance of the proposed method, experiments were conducted on three common Arabic text classification collections: BBC, Watan, and Abuaiadah datasets. These datasets were selected to test the proposed method in different situations such as balance and dataset size in terms of the number of documents. A brief description of these corpora is provided next while Figure 2 shows the statistical distribution of documents in these datasets.

- *Watan dataset [1]*: This corpus contains more than 20000 documents that fall into six categories which are: culture, religion, economy, local news, international news and sports. Originally, the numbers of documents in these categories are not equal. Thus, the researchers selected 9900 documents that are equally distributed over the categories. The aim of considering this dataset is testing the performance of the proposed method under the big sized dataset condition. In Arabic TC domain, this corpus is popular and has been used widely in many works such as in [2, 6].
- *Abuaiadah dataset [3]*: This is a balanced dataset that consists of 2700 documents distributed equally in nine categories which are: economy, health, law, literature, politics, religion, sport, and technology. The documents of this corpus are of the same size (approximately 2 Kilobytes) and collected from various resources. Even though this dataset is new, it has been used in many Arabic TC works such as [4, 15].
- *BBC dataset [25]*: BBC is an unbalanced free dataset that consists of 4763 documents. The documents in this dataset are distributed in seven different classes which are: business and economy, Middle East news, Misc, newspapers highlights, science and technology, sports, and world news. This dataset is used widely in Arabic TC such as in [24, 32].

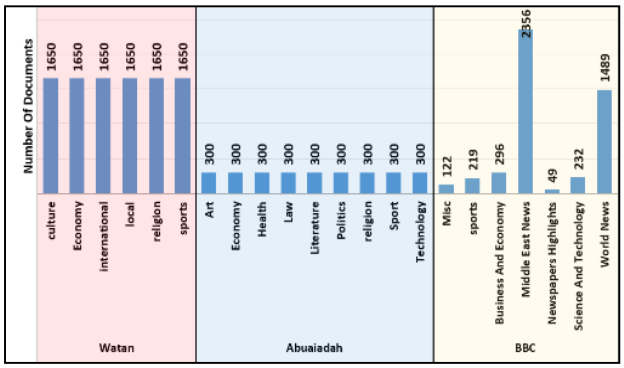

Figure 2. Documents distribution in considered datasets.

As seen in Figure 2, the Watan and Abuaiadah datasets are balanced datasets (i.e., the count of documents are equal in all categories) with a different number of categories. It is important to note that the Watan dataset is a big-sized dataset while Abuaiadah dataset is small-sized one. The BBC dataset, however, is an unbalanced dataset with an adequate number of documents.

4.2. Experiments

The proposed SVM-FRM was tested against two of the commonly applied traditional feature ranking (selection) techniques that are IG and Chi Square (Chi^2). Three datasets were used for the purpose of making the referred to comparison. The performance of SVM-FRM was tested against these methods based on different numbers of features. As explained in section 3.3, all features in the feature space are ranked based on SVM-FRM and other ranking methods as well, individually. Feature subsets of different sizes were selected and considered for classification. The sizes of these feature subsets are 100, 500, and 1000 to 5000 features (in intervals of 500 features). The Top K ranked features according to each of the ranking methods were selected each time and the experiment is carried out. The Rapid miner Studio V7.5 software was used to handle these experiments. Figure 3 shows the structure of the basic process in Rapid miner.

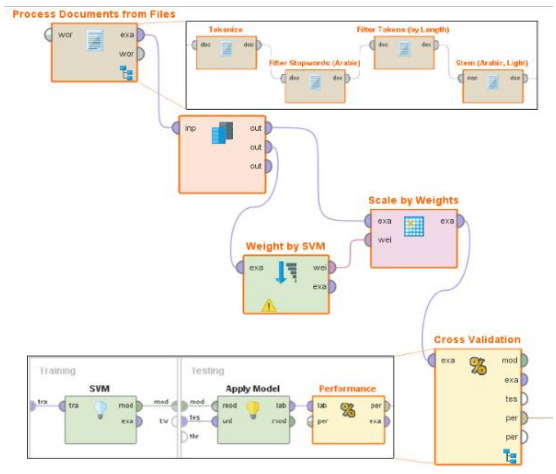

Figure 3. Structure of SVM-FRM in Rapid miner software.

5. Results and Discussion

This section presents the accuracy and averaged F-measure results on the considered corpora.

Figures 4, 5, and 6 show the accuracy results of experiments completed on the datasets Watan, Abuaiadah, and BBC, where the Full FS is the full feature space of each dataset which counts 86389, 43462, and 38630 features, respectively.Table1 shows the averaged F-measure benchmarking results.

Figure 4. Accuracy results on Watan dataset.

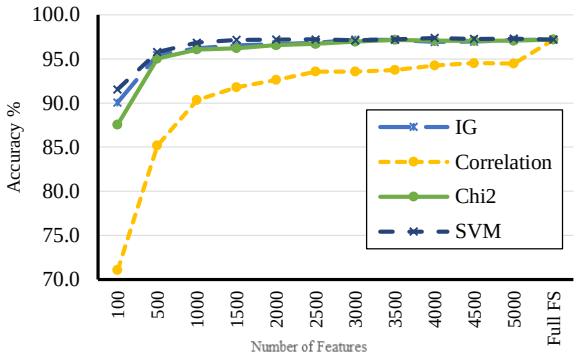

Figure 5. Accuracy results on Abuaiadah dataset.

As seen in Figures 4 and 5, the proposed SVM-FRM outperformed other traditional feature ranking methods on Watan and Abuaiadah datasets, with a maximum accuracy of 93.94% and 97.37% respectively. Documents in each of these datasets are distributed equally over dataset's categories (i.e., balanced datasets).

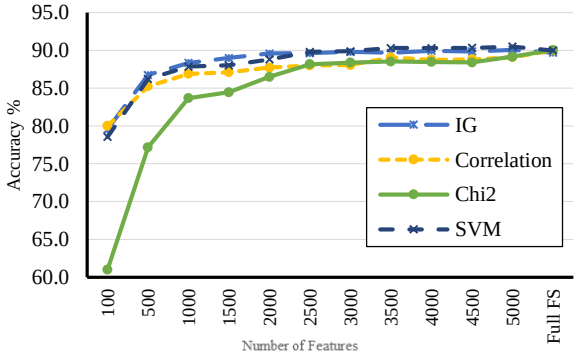

Figure 6. Accuracy results on BBC dataset.

However, the Watan dataset can be considered as big sized dataset as it consists of 9900 documents with a full feature space of 86389 features. The other

dataset i.e., Abuaiadah dataset, however, is a small dataset that contains less than 40000 features. The results in Figures 4 and 5 indicate the ability of the proposed method to perform well in the condition of balanced datasets in spite of the dataset's size.

Results in Figure 6 (i.e., accuracy results on BBC dataset) show that the proposed method outperforms the Correlation and Chi square feature ranking methods only. Similar to Abuaiadah dataset, BBC dataset is a small dataset with less than 40000 features. Still, BBC is an unbalanced dataset where the counts of documents in dataset's categories are not equal. In this case, our experimental results show that the IG feature ranking method outperformed other methods for the subsets that consisted of less than 2500 features, while the proposed SVM-FRM outperformed other methods for the subsets that contained 3500 features and more with a maximum accuracy value of 90.49%.

Besides, the accuracy results based on the Full FS (i.e., full feature space) are equal in spite of the feature ranking method per dataset. This case indicates that all features in the feature space are included in the classification process, where a large number of noisy and less important features are considered, leading to a very long learning and classification time.

The reported averaged F-measure benchmarking results in Table 1 show that the proposed SVM-FRM not only outperformed other methods on the balanced datasets (i.e., Watan and Abuaiadah datasets) but also obtained superior performance, with a maximum average F-measure values of 93.94% and 97.38%, respectively. On the contrary, none of the benchmarked methods showed superior average F-measure performance on the unbalanced BBC dataset. The maximum average F-measure value is obtained by the Chi^2 method based on the feature set of size 3000 features, as shown in Table 1.

Table 1. Averaged F-measure benchmarking results.

No. of Features	Watan Dataset				Abuaiadah Dataset				BBCDataset			
	IG	*Cor*[a]	*Chi*^{2}	*SVM*	*IG*	*Cor*[a]	*Chi*^{2}	*SVM*	*IG*	*Cor*[a]	*Chi*^{2}	*SVM*
100	81.27	73.32	83.85	86.76	90.46	72.66	88.65	91.79	76.01	52.87	66.87	49.52
500	89.43	87.23	90.84	91.91	95.49	86.09	95.09	95.83	84.59	65.12	79.81	66.31
1000	91.41	89.55	92.11	92.70	96.26	90.95	96.11	96.87	85.63	67.11	84.12	68.64
1500	91.73	90.64	92.45	93.02	96.53	92.08	96.25	97.16	86.10	67.51	85.06	68.98
2000	92.09	91.17	92.79	93.57	96.68	92.83	96.58	97.21	87.31	75.30	85.73	78.98
2500	92.77	91.56	92.93	93.57	96.83	93.73	96.73	97.24	87.04	76.89	87.12	84.25
3000	92.74	91.64	93.19	93.83	97.19	93.73	96.98	97.12	86.92	77.45	**87.50**	84.15
3500	92.83	91.97	93.40	93.90	97.16	93.91	97.16	97.23	87.56	80.78	87.07	84.73
4000	92.95	92.33	93.35	93.90	96.93	94.43	97.08	**97.38**	87.68	80.78	86.84	86.10
4500	92.98	92.47	93.53	93.75	96.97	94.66	97.05	97.27	87.17	81.83	86.14	86.04
5000	93.25	92.54	93.47	**93.94**	97.16	94.62	97.08	97.30	87.05	81.62	86.65	86.57
Full FS	93.61	93.61	93.61	93.61	97.20	97.20	97.20	97.20	86.52	86.92	66.87	86.92
											[a] Correlation	

However, the performance of SVM-FRM shows close results against the IG and Chi^2 methods based on big-sized feature sets (i.e., feature sets contained of 5000 features and more), as shown in Figure 7.

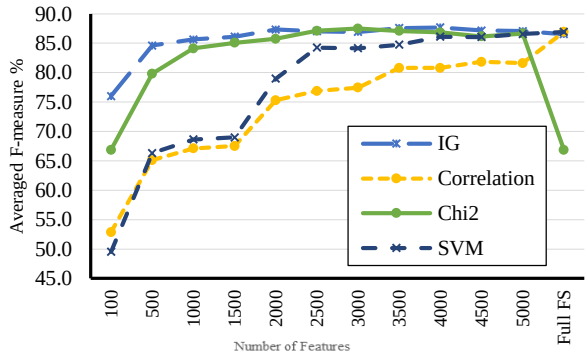

Figure 7. Averaged F-measure benchmarking results on BBC dataset.

The presented accuracy and average F-measure results can lead to a conclusion that the proposed SVM-FRM shows outstanding performance in the case of balanced datasets (such as Watan and Abuaiadah datasets), while it shows comparatively less performance when applied on the unbalanced dataset. Therefore, the research team focused on improving the proposed method to perform better on the unbalanced dataset as well, as shown in next section.

6. Improving SVM-FRM by Hybridization with IG and Chi^2

To further improve the performance of the proposed SVM-FRM method, we applied feature set hybridization in which the feature sets obtained by various ranking methods in addition to SVM-FRM were merged based on union operation from the Set Theory. It is assumed that this sort of hybridization will lead to higher performance of the proposed method as it will aggregate informative features obtained by different ranking methods into one feature set. Union based hybridization has been applied in some existing studies in the domain of feature selection methods such as [8].However, it has not been used with any machine learning based feature ranking method like the proposed SVM-FRM.

Mathematically, the union of two sets *A* and *B* is defined as the set of elements that are members of

either of the two sets. The union operation is commonly denoted by ∪, and is expressed as in equation (10):

$$A \cup B = \{x | x \in A \text{ or } x \in B\} \quad (10)$$

6.1. Implementation of Union Operation

Although, we handled our experiments using the Rapid miner software, still we noticed that the union operator in this software works in such a way that requires higher running time. To explain more about this issue, we suppose that FS1 and FS2 are the feature sets obtained by two different feature ranking methods and WM1 and WM2 are the weight matrices related to FS1 and FS2, respectively. The Rapid miner based union operator will duplicate the samples in the generated Hybridized Weighting Matrix (HWM) corresponding to the Hybridized Feature Set (HFS) while considering some entries of HFS as missing values. Figure 8 illustrates this issue.

As seen in Figure 8, the resulted Hybridized Weighting Matrix (HWM) which will be used for classification will be large in terms of number of instances, as the instances of the corpus will be repeated based on the number of feature sets being combined.

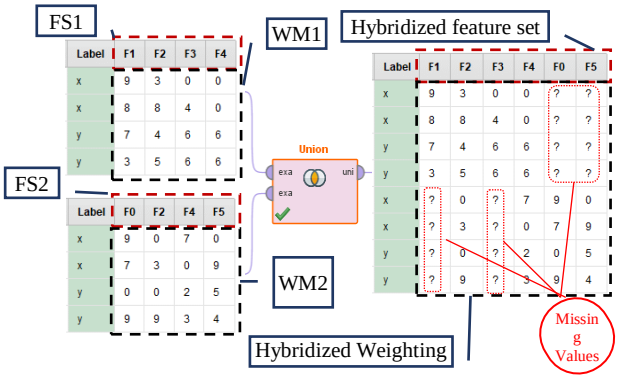

Figure 8. Illustration of union operation in rapid miner software.

For example, in our experiment, the union of feature sets based on IG, Chi², and SVM resulted an HWM that contained 14289 rows (i.e., the number of documents in the original dataset 4763 multiplied by 3). This huge number of instances will consume the extended classification time in addition to the less accuracy problem caused by the missing values occurred in the HWM.

Therefore, we implemented the union operation in Matlab, so that the resulted HWM will be generated as shown in Figure 9.In this implementation, the entries in HWM for the common features between the combined feature sets (i.e., F2 and F4 in the illustration example in Figures 8 and 9) are calculated by averaging the entries of these features.

Hybridized feature set

Label	F1	F2	F3	F4	F0	F5
x	9	1.5	0	3.5	9	0
x	8	5.5	4	0.0	7	9
y	7	2.0	6	4.0	0	5
y	3	7.0	6	4.5	9	4

HWM

Figure 9. HWM as generated by the implemented union operation.

From Figure 9, it is seen that the resulted HWM contains no missing values and the number of instances (i.e., rows) in the matrix is the same as the number of documents in the dataset.

6.2. Experiments based on Hybridization

As the current aim is to enhance the performance of SVM-FRM method on the unbalanced dataset, we conducted the hybridization-based experiments based on the combination of IG, Chi² and SVM feature ranking methods. Results based on the IG-SVM and Chi²-SVM combinations have not enhanced the performance significantly, however, the results based on IG-Chi²-SVM methods combination showed a noticeable accuracy enhancement. Table 2 shows combinations tried by the researchers and the recorded level of accuracy enhancement, while Figure 10 shows the comparison of accuracy results based on hybridization against the baseline results (i.e., results of individual methods IG, Chi², and SVM-FRM), and the IG-Chi² hybridization.

Table 2. Accuracy improvement of different combinations of methods.

Method 1	Method 2	Method 3	Accuracy improved
IG	Chi²	--	Yes
IG	--	SVM	No
--	Chi²	SVM	No
IG	Chi²	SVM	Yes

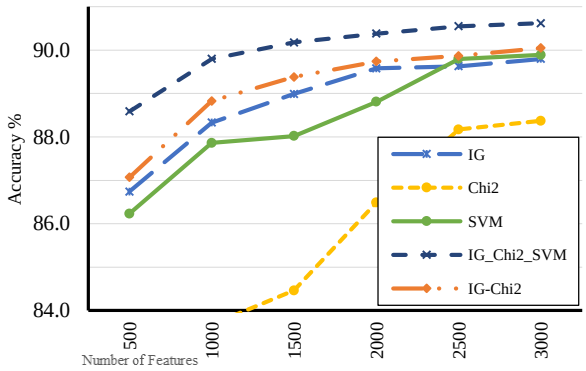

Figure 10. Comparison of accuracy results based on hybridization against the baseline results.

As seen in Figure 10, the accuracy results based on the hybridization on IG, Chi², and SVM-FRM methods outperformed the results of all individual methods with a maximum accuracy of 90.62% based on the union of feature sets that consist of 3000 features. Moreover, the IG-Chi²-SVM outperformed IG-Chi² combination which indicates that the features

selected by the SVM-FRM method are highly distinguishing and informative.

The results discussed in this section in addition to those discussed in section 5 supports the basic assumption of this research that the features ranked higher by the SVM algorithm can lead to the higher classification performances than the features ranked based on statistical methods.

7. Conclusions and Future Work

This research presented the SVM-FRM in which the weighting and ranking of features are based on the SVM learning algorithm. The benchmarking accuracy and average F-measure results with many different statistical feature selection methods on various datasets have concluded that the proposed SVM-FRM has an outstanding performance in the case of balanced datasets (such as Watan and Abuaiadah datasets).

However, it showed less performance on unbalanced datasets. In order to address this shortcoming, a hybridization approach of the proposed SVM-FRM method was presented based on union operation. The hybridized SVM-FRM method showed higher results than baseline methods and other combined methods on the unbalanced dataset. The future work of the research team will focus on enhancing the proposed method by applying other feature types and examining its performance using different classification algorithms and applying it on more public text classification datasets. Additionally, authors think about comparing the results of proposed method to wide range of feature selection methods exist in Text Classification domain.

References

[1] Abbas M., Smaili K., and Berkani D., "Comparing TR-Classifier and KNN by using Reduced Sizes of Vocabularies," *in Proceedings of the 3rd International Conference on Arabic Language Processing*, Rabat, 2009.

[2] Abbas M., Smaïli K., and Berkani D., "Evaluation of Topic Identification Methods on Arabic Corpora," Journal of Digital Information Management, vol. 9, no. 5, pp. 185-192, 2011.

[3] Abuaiadah D., El Sana J., and Abusalah W., "On The Impact of Dataset Characteristics on Arabic Document Classification," *International Journal of Computer Applications*, vol. 101, no. 7, pp. 31-38, 2014.

[4] Abuaiadah D., "Using Bisect K-Means Clustering Technique in the Analysis of Arabic Documents," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 15, no. 3, pp. 1-13, 2016.

[5] Alguliyev R., Aliguliyev R., and Isazade N., "An Unsupervised Approach to Generating Generic Summaries of Documents," *Applied Soft Computing*, vol. 34, pp. 236-250, 2015.

[6] Aliwy A., "Tokenization as Preprocessing for Arabic Tagging System," *International Journal of Information and Education Technology*, vol. 2, no. 4, pp. 348-353, 2012.

[7] Aphinyanaphongs Y., Fu L., Li Z., Peskin E., Efstathiadis., Aliferis C., and Statnikov A., "A Comprehensive Empirical Comparison of Modern Supervised Classification and Feature Selection Methods for Text Categorization," *Journal of the Association for Information Science and Technology*, vol. 65, no. 10, pp. 1964-1987, 2014.

[8] Bharti K. and Singh P., "Hybrid Dimension Reduction by Integrating Feature Selection with Feature Extraction Method for Text Clustering," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3105-3114, 2015.

[9] Chang Y. and Lin C., "Feature Ranking Using Linear SVM," *in Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI*, Hong Kong,pp. 53-64, 2008.

[10] Chen Y. and Chen M., "Using Chi-Square Statistics to Measure Similarities for Text Categorization," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3085-3090, 2011.

[11] Dasari D. and Rao K., "Text Categorization and Machine Learning Methods: Current State Of The Art," *Global Journal of Computer Science and Technology Software and Data Engineering*, vol. 12, no. 11, pp. 37-46, 2012.

[12] Efron M., Zhang, J. and Marchionini G., "Comparing Feature Selection Criteria for Term Clustering Applications," *in Proceedings of ACM SIGIR Conference on Research and Development in Informaion Retrieval*, Toronto, pp. 28-31, 2003.

[13] Forman G., "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289-1305, 2003.

[14] Guyon I. and Elisseeff A., "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[15] HmeidiI., Al-Ayyoub M., Abdulla N.,Almodawar A., Abooraig R., and Mahuoob N., "Automatic Arabic Text Categorization: A Comprehensive Comparative Study," *Journal of Information Science*, vol. 4, no. 1, pp. 114-124, 2014.

[16] Jiang S., Pang G., Wu M., and Kuang L., "An Improved K-Nearest-Neighbor Algorithm for Text Categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503-1509, 2012.

[17] Joachims T., "Text Categorization with Suport

Vector Machines: Learning with Many Relevant Features," *in Proceeding of Machine Learning: ECML*, Berlin, pp. 137-142, 1998.

[18] Lee C. and Lee G., "Information Gain and Divergence-Based Feature Selection for Machine Learning-Based Text Categorization," *Information Processing and Management*, vol. 42, no. 1, pp. 155-165, 2006.

[19] Lee L., Wan C., Rajkumar R., and Isa D., "An Enhanced Support Vector Machine Classification Framework by using Euclidean Distance Function for Text Document Categorization," *Applied Intelligence*, vol. 37, no. 1, pp. 80-99, 2012.

[20] Man, L., Tan C., Su J., and Lu Y.,"Supervised and Traditional Term Weighting Methods for Automatic Text Categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721-735, 2009.

[21] Liu L. and Özsu M.,Encyclopedia of Database Systems, Springer 2009.

[22] Meyer D. and WienT., "Support Vector Machines-the Interface to Libsvm in Package e1071," Technical Report, 2001.

[23] Onan A., "Classifier and Feature Set Ensembles for Web Page Classification," *Journal of Information Science*, vol. 42, no. 2, pp. 150-165, 2016.

[24] Raho G., Kanaan G., Al-Shalabi R., and Nassar A., "Different Classification Algorithms Based on Arabic Text Classification: Feature Selection Comparative Study," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 2, pp. 23-28, 2015.

[25] Saad M. and Ashour W., "OSAC: Open Source Arabic Corpora," *in Proceedings of the 6th International Conference on Electrical and Computer Systems*, Lefke, pp. 118-123,2010.

[26] Sabbah T., Selamat A., Selamat M., Ibrahim R., and Fujita H., "Hybridized Term-Weighting Method for Dark Web Classification," *Neurocomputing*, vol. 173, no. P3, pp. 1908-1926, 2016.

[27] Sabbah T., Selamat A., Selamat M.,Fujita H., Al-Anzi F., Viedma E., and Krejcar O., "Modified Frequency-Based Term Weighting Schemes for Text Classification," *Applied Soft Computing*, vol. 58, pp. 193-206, 2017.

[28] Sulic V., Pers J., Kristan M., and Kovacic S.," Efficient Dimensionality Reduction using Random Projection," *in Proceedings of the Computer Vision Winter Workshop*, Prague, pp. 29-36, 2010.

[29] Uysal A., "An Improved Global Feature Selection Scheme for Text Classification," *Expert Systems with Applications*, vol. 43, pp. 82-92, 2016.

[30] Yang Y. and Pedersen J., "A Comparative Study on Feature Selection in Text Categorization," *in Proceedings of the 14th International Conference on Machine Learning*, Nashville, pp. 412-420, 1997.

[31] Yang Z., He J., and Shao Y., "Feature Selection Based On Linear Twin Support Vector Machines," *Procedia Computer Science*, vol. 17, pp. 1039-1046, 2013.

[32] Yousif S., Elkabani I., Samawi V., and Zantout R., "Enhancement of Arabic Text Classification Using Semantic Relations With Part of Speech Tagger," *in Proceedings of 14th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*,Cambrigde,pp. 195-201, 2015.

[33] Zhang W., Yoshida T., and Tang X., "A Comparative Study of TF*IDF, LSI and Multi-Words for Text Classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758-2765, 2011.

Thabit Sabbah received his Bachelor of Computer Science BSc (CS), Master of Computer Science MSc (CS) from Al Quds University, Jerusalem / Palestine, and Doctor of Philosophy PhD in Computer Science from Universiti Teknologi Malaysia UTM, Malaysia in 1998, 2009 and 2015 respectively. His research interests are mainly focused on Data Mining, Text Mining and Classification, Information Retrieval, Machine Learning, and Artificial Intelligence. He has broad experience in administrative work, teaching and research. During the past 20 years he worked in many administrative and Academic positions. Currently, he is a Faculty Member in the Collage of Technology and Applied Sciences at Al Quds Open University / Palestine. Dr. Sabbah has received many academic and research awards. He has published a number of articles in high ranked International Journals, and many other research papers in International Conferences, Book Chapters, and he has been a reviewer of various International Journals and Conferences.

Mosab Ayyash received his Bachelor of Computer Science BSc (CS) from Al Quds University, Jerusalem / Palestine in 2003, and Master Degree (MSc) in Scientific Computing from Berzeit University in 2007. Currently, he is a Lecturer and Faculty Member of Computer Information Systems department / Collage of Technology and Applied Sciences at AL Quds Open University (QOU). His research interests are focused on the fields of Database System, Data mining, Project Management, and Data Analysis.

Mahmood Ashraf received his Bachelor of Computer Science BSc(CS), Master of Computer Science MSc(CS), second Master of Computer Science MS(CS) from Islamabad, Pakistan and Doctor of Philosophy PhD in Computer Science from Universiti Teknologi Malaysia UTM, Johar Bahru, Malaysia in 1999, 2002, 2008 and 2014 respectively. His areas of interests are: Human-Computer Interaction, Physintuitive Systems, Smart Environment, Text Classification, Machine Learning, Artificial Intelligence, and Intelligent User Interfaces. He has been administrative, academic and research Head of Islamabad Campus (as In charge Campus) of Federal Urdu University of Arts, Science and Technology (FUUAST) from 2017 to 2018. Dr. Mahmood Ashraf has published a number of research papers in National, International Conferences, Book Chapters and International Journals. He is Higher Education Commission (HEC)'s recognized MS/PhD supervisor. He has been a reviewer of various International Conferences and an International Journal.