ISSN: 0067-2904

The Evaluation of Accuracy Performance in an Enhanced Embedded Feature Selection for Unstructured Text Classification

Nur Syafiqah Mohd Nafis, Suryanti Awang*
Soft Computing & Intelligent System Research Group (SPINT), Faculty of Computing, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300, Kuantan, Pahang, Malaysia

Received: 15/10/2019 Accepted: 21/1/2020

Abstract

Text documents are unstructured and high dimensional. Effective feature selection is required to select the most important and significant feature from the sparse feature space. Thus, this paper proposed an embedded feature selection technique based on Term Frequency-Inverse Document Frequency (TF-IDF) and Support Vector Machine-Recursive Feature Elimination (SVM-RFE) for unstructured and high dimensional text classification. This technique has the ability to measure the feature's importance in a high-dimensional text document. In addition, it aims to increase the efficiency of the feature selection. Hence, obtaining a promising text classification accuracy. TF-IDF act as a filter approach which measures features importance of the text documents at the first stage. SVM-RFE utilized a backward feature elimination scheme to recursively remove insignificant features from the filtered feature subsets at the second stage. This research executes sets of experiments using a text document retrieved from a benchmark repository comprising a collection of Twitter posts. Pre-processing processes are applied to extract relevant features. After that, the pre-processed features are divided into training and testing datasets. Next, feature selection is implemented on the training dataset by calculating the TF-IDF score for each feature. SVM-RFE is applied for feature ranking as the next feature selection step. Only top-rank features will be selected for text classification using the SVM classifier. Based on the experiments, it shows that the proposed technique able to achieve 98% accuracy that outperformed other existing techniques. In conclusion, the proposed technique able to select the significant features in the unstructured and high dimensional text document.

Keywords: Embedded feature selection, text classification, text mining, sentiment analysis

1.0 Introduction

Social media has been an important platform to convey information and messages nowadays. More than that, it also contains hidden knowledge that is helpful for many purposes. However, that information hidden resides in the unstructured textual data and high-dimensional data. Classifying unstructured text documents is a critical task for text mining applications, such as sentiment analysis [1], disaster prediction [2], and business analysis [3]. Text classification is one method used to extract information from text documents. It comprises several steps, including feature selection that able to facilitate selecting the significant feature subset. It also helps in improving classification accuracy, reducing computational time, and providing a better understanding of the model studied. Several researchers proved that feature selection gave an impact on their classification problems [4-6]. Feature selection approaches can be categorized into the filter, wrapper, and embedded. The filter approach selects features based on some feature matrices, for example, feature importance and feature correlation. They claim it to be the simplest and straightforward approach among all. However,

*Email:suryanti@ump.edu.my

classification accuracy cannot be guaranteed [7]. Meanwhile, the wrapper approach utilized any classifier performances to select the best features. Nevertheless, it depends on the learning algorithm, which takes a long time to search for the best features [8]. Work similar to the wrapper approach is the embedded approach. The embedded approach links feature selection with the classification stage. The link is much stronger since it included feature selection into classifier construction and optimization into consideration.

The filter approach is often used in text classification due to its fastness and simplicity in handling sparse feature space of text documents [9]. However, classification accuracy cannot be guaranteed. Thus, the learning algorithm is necessary to aid the feature selection process. Therefore, this paper proposed the two-stage embedded feature selection approach to form a new strategy to solve issues in feature selection approaches for text classification. The proposed technique provides a new strategy in selecting features for high dimensional text classification by measuring feature importance based on the classifier performance evaluation. This study demonstrates the embedded feature selection approach by utilizing Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination. TF-IDF is a well-known filter approach, which had been widely used in an information retrieval system due to its fastness, robustness and simplicity [7][10][11]. Meanwhile, SVM-RFE is an embedded approach that mostly applied in high-dimensional data classification by using the discrimination function information of SVM to remove the feature with the smallest correlation with the classifier from the original feature set [12]. This approach achieved better classification performance in terms of accuracy compared to a single TF-IDF and the two-stage hybrid of TF-IDF and SVM technique.

The structure of this paper as follows; Section 2 discusses the related works of research. Meanwhile, Section 3 describes the method. In Section 4, we discuss research findings and analysis. Last, Section 5 concludes the research.

2.0 Related Works

Feature selection has been the indispensable phase in classification. Moreover, it provides an efficient way to remove irrelevant and duplicate features from the dataset [13]. As in text classification, the aim of feature selection is to select the most important features to represent the whole text collection [14]. As mentioned in the previous section, TF-IDF is the most common feature selection techniques used in text classification. It is successfully applied for feature weighting technique using document frequency (DF) and term frequency (TF) based feature selection in text classification. Jing et.al introduced TF-IDF as early as in 2002 [15].

Whereas, [16] conducted a study to investigate the impact of TF-IDF is a feature selection method on document clustering. . They conducted several experiments by dividing it into several phases. The phases are pre-processing and term selection. The term selection phase consists of TF-IDF, TF-DF, and TF-IDF*TF-DF. The highest percentages of the removed features among these three techniques are from TF-DF and TF-IDF*TF-DF compared to TF-IDF. However, when more features are eliminated, there are possibilities of data loss.

While [17], explored the term and document frequencies as feature selection matric. They examined the document frequency-based metrics of discriminative power measure and GINI index with term frequency. The proposed technique is accessed and analyzed on the Reuters 21,578 dataset. However, the experimental result reveals that the term frequency outperformed for smaller size datasets only. From deep research, it exposes the two important characteristics of term frequency, which contribute to their great performance for smaller feature sets. The smaller feature sets have a relatively larger scatter of features among the classes and accumulate information in data at a less time.

Later, [18] explored Weighted Document Frequency (WDF) for feature selection in text classification. Previous researches have stated that document frequency (DF) has been a simple but successful method for feature selection in text classification. This DF method only measures how many times the word of a term appears in the document, however, it does not measure the importance of the word or term to the document. The DF method clearly introduces too much noise. Hence, the author suggested two WDF techniques to overcome the previously mentioned problems. The techniques are WDF1 and WDF2. The WDF1 is the DF-based method, while WDF2 is based on TF-IDF. They demonstrate the experiments to measure the effectiveness of the suggested technique. The experimental results show that when the highest N-top features were selected, both WDFs

outperformed the DF technique as well as the Chi-Square technique. Nevertheless, WDF1 is more stable compared to WDF2. Both WDF1 and WDF2 as well as the conventional TF-IDF measuring feature importance for text classification. However, the TF-IDF approach is more simple and easy to understand.

The recent study of TF-IDF, suggested a modified version of the TF-IDF technique and Glasgow expression using graphical representations to minimize the size of the feature set [18]. They utilize the cumulative curve to estimate the number of features. In addition, they use the SVM classifier to test the proposed technique. The study finds that the modified version of TF-IDF and Glasgow expression are able to enhance the performance of the SVM classifier for text classification. In addition, it achieved better performance compared to the traditional term weighting expressions adopted for feature selection. Nonetheless, the proposed technique is only based on high-frequency features. Thus, there is a possibility that some low-frequency significant features might be removed.

In other text classification applications of spam filtering, DF is applied to the hybrid method (HBM) feature selection technique. It combines document frequency information and term frequency information [19]. This technique aims to solve the drawback in a single application of document frequency. In order to maintain the category discriminating ability of the selected features, an optimal document frequency-based feature selection (ODFFS) is implemented. For the remaining features, HBM will handle them by selecting features with HBM value. In addition, a parameter optimization also introduced, feature subset evaluating parameter optimization (FSEPO). Lastly, two classifiers are chosen, namely, SVM and Naïve Bayes to access the proposed methodology in four corpora. The four corpora are PU1, LingSpam, SpamAssian, and Trec2007. Among other feature selection techniques which are Information Gain, Chi-square, improved Gini-index, multi-class Odds Ratio, normalized term frequency-based discriminative power measure, and comprehensively measure feature selection, HBM shows the most significant improvement when both classifiers are applied. Nevertheless, comparing HBM with conventional TF-IDF, HBM is a complete technique, which incorporates parameter optimization for better classification performance. Thus, it motivates this paper to develop a two-stage embedded feature selection technique.

SVM-RFE is one of the SVM variants introduced by [20]. It is an embedded feature selection approach. A study was carried out to improve the feature selection technique using SVM-RFE for a multi- SVM classifier [21]. The class interval in the SVM is utilized as the evaluation criterion, and later, it eliminates features in a recursive way. Obtaining optimal SVM is a base for feature selection. Hence, the chaos particle swarm optimization (CPSO) algorithm is implemented. The improved SVM-RFE feature selection technique works well to overcome the feature selection in multi-class conditions with the help of CPSO.

In contrast, there is a research that introduced support Vector Machine- Recursive Feature Addition (SVM-RFA) [22]. SVM-RFA begins with an empty feature set and keeps adding until it meets a stopping criterion. SVM-RFA performances were tested on five established datasets ranging from 9 to 101 features which means they only test the proposed technique on the low dimensional dataset. The experimental results of the study proved that the proposed feature selection technique successfully works better than filter and wrapper as well as SVM-RFE. However, SVM-RFA does not surpass SVM-RFE in some datasets.

As a vital task in classification, researchers have put so much attention on feature selection to improve classification performances. However, the traditional feature selection provides limited contributions to classification performances. Hence, researchers had taken steps forward to enhance the capability of feature selection techniques [23-25].

In conclusion, all of the above-related works implemented term frequency-based and SVM-based techniques. However, those techniques are a filter technique or an embedded technique. Thus, current techniques do not focus on how to measure the importance of the features in a document. Due to this limitation, this paper attempts to propose the enhancement of the embedded feature selection technique using TF-IDF and SVM-RFE. This proposed technique is capable to remove insignificant features and measure the importance of features in a document.

3.0 Methodology

Based on the related works, this study proposed the enhanced features selection technique for the text classification. The proposed technique embeds TF-IDF and SVM-RFE in the feature selection phase (TF-IDF+SVM-RFE). This section, explains the methodology of the proposed techniques.

Figure-1 shows a few phases involved in text classification. The process begins with data acquisition whereby the document collections are retrieved from the UCI Machine Learning Repository. The text document comprises 200 posts from Twitter known as tweets. They are labelled as negative and positive tweets. Table-1 illustrates the sample of the dataset used for this paper. The Tweet Id. represents the identification number of the sample. While the text is the post for the respected Tweet. Id. The label is denoted as "0" or "1" which refers to "0" is a negative tweet and "1" is a positive tweet. Therefore, by using this dataset and the proposed technique, it is able to classify the content of the tweets as negative or positive.

The next phase is pre-processing the raw dataset. The aim in this phase is to reduce the number of features for classification. Therefore, the proposed technique will process fewer features. The pre-processing phase consists of several activities, which are tokenization, stop-word removal, stemming and generating term-document matrix (TDM). Tokenization is a process to chunk a paragraph of sentences into separate sentences and finally into a single token known as a feature. The functional word, for example, "for", "on", "is", "the", and many more are the necessary stop-word to be removed since they do not give any significant meaning to the text classification. Stemming also will help to cut down the number of features by extracting only root words from samples. Lastly, it generates one TDM for the later process of calculating the TF-IDF score in feature selection.

The process continues by dividing the pre-processed dataset into a training dataset and testing dataset: 20% and 80%, respectively, for testing and training phases. The reason to have 80% of the training dataset is to provide more samples for the training process. With this implementation, issues of misclassification and overfitting can be avoided. Consequently, a better classification performance can be achieved. The feature selection takes place in both the training dataset and the testing dataset. For the training dataset, the feature selection will produce a set of trained features. Meanwhile, for the testing dataset, it removes unnecessary features by comparing features based on the trained features obtained from the training dataset. In feature selection, it aims to select significant features and removing redundant features to enhance the classification performances. A detail explanation of the feature selection is available in the next section.

Table 1- The sample of Twitter posts

Tweet Id	Text	Label
T1	What a waste of money and time!	0
T2	Good case, Excellent value.	1
T3	Great for the jawbone.	1
T4	I advise EVERYONE DO NOT BE FOOLED!	0
T5	The mic is great.	1

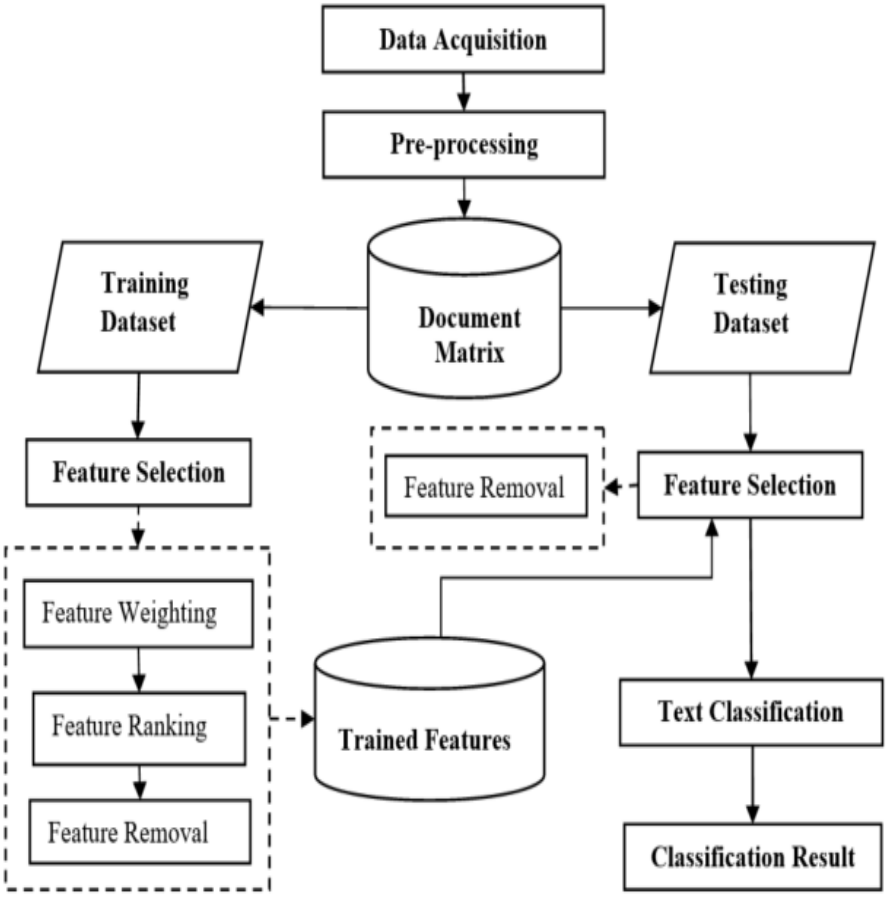

Figure 1- Text classification methodology

3.1 Feature Selection

Feature selection is one of the most vital phases in classification. It helps in improving the classification performances by reducing the number of dimensionalities. A major problem that arises in text classification is having a high-dimensional feature space [25]. The curse of dimensionality is one of the problems mentioned in [26]. Hence, selecting the best feature subset in the high-dimensional feature space is a challenging task. Thus, in this research, an enhanced feature selection technique is proposed to select the best features in a high-dimensional space.

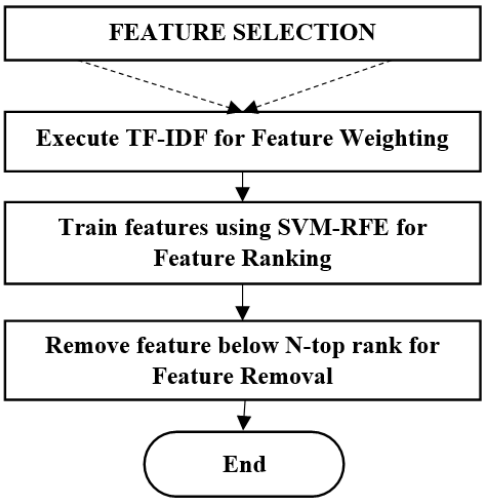

Figure 2– Feature selection flowchart.

The flowchart in Figure-2 summarized the whole processes taken for the feature selection in the training dataset. It consists of three steps. Firstly, it computes the TF-IDF score each pre-processed feature of the sample. The higher the TF-IDF value, the more important the feature. Next, it creates a readable vector matrix for 'LibSVM' and SVM-RFE is applied to the training dataset for feature ranking. This process produced a list of feature ranks. The top of the list indicates that the most important feature. However, it selects only the top rank feature for later the classification phase. The percentage (N-top features) of selected top rank features is pre-defined before the classification process. Finally, to measure the effectiveness of the proposed technique, SVM is chosen as a classifier. The next sub-sections explain the detail process that involved in this feature selection.

A. Feature Weighting using Term Frequency-Inverse Document Frequency

TF-IDF is a well-known filter feature selection approach. It measures the importance and relevance of a feature of a large document collection. The TF-IDF formula is written as follow;

$$TF-IDF(t,d,D) = TF(t,d) * IDF(t,D) \quad (1)$$

where, let $D = \{d1, d2, d3, ...dn\}$ be a collection of documents and t be a term that appears in the collection. $TF(t,d)$ represents the frequency of the term t in document d. It can be represented in the formula as follows;

$$TF = \frac{TN}{ND} \quad (2)$$

where, TN is a total number of the term in a document, and ND is the number of times a term appears in a document. In short, the term frequency (TF) represents the total number of terms that appear in a document. Meanwhile, $IDF(t, D)$ is the inverse document frequency, where t represents the frequency of the term that appears in D. D is the number of the document in the collection. The inverse document frequency (IDF) determines the importance of a term in the whole document collection. It can be represented in the formula as below;

$$IDF = \log_2 \frac{NDT}{TD} \quad (3)$$

where, NDT is a number of the document with term t in them, and TD is a total number of documents. Overall, TF-IDF defines that the TF-IDF score increases proportionally with the frequency of a word appears in a document compared to the inverse proportion of the frequency of the same word in the whole document collections. The feature is more representative if it has a larger TF-IDF value.

B. Train Features using SVM-RFE for Feature Ranking

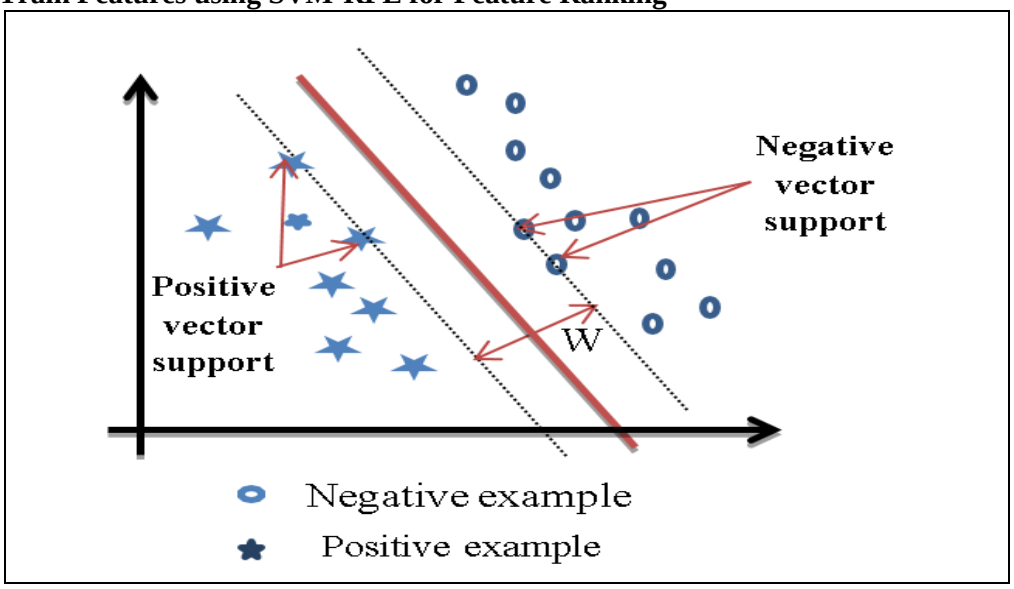

Figure 3– SVM hyperplane concept

Support Vector Machine (SVM) works well in categorizing text documents. SVM classifies binary class problem by finding the separation between hyperplanes defined by classes of data shown in Figure-3. Assume there is a given set, S, of points $x_i \in R^{\wedge}(n)$ with $i=1, 2, 3....N$. Each point x_i belongs

to either of two classes with a given a label $y_i \in \{-1,1\}$. The objective is to establish the equation of a hyperplane that divides *S* leaving all the points of the same class on the same side. SVM commits classification by developing an N-dimensional hyperplane that optimally splits the data into two categories. SVM score, *W* can be written as the formula below;

$$W = \sum_{i=1}^{n} a_i y_i x_i \quad (4)$$

where *i* is the number of terms ranging from 1 to *n*, a_i is Lagrangian Multiplier estimated from the training set; x_i is term vector for sample *i* and, y_i is the class label of sample *i*. Weighted vector or SVM score is defined by the sum square of the weight vector *W* of the SVMs using formula (4).

Meanwhile, SVM-RFE is commonly implemented for high dimensional data, for instance, micro-array gene expression [27-29]. In the SVM-RFE algorithm, it eliminates the irrelevant and redundant terms, as well as noises, in a sequential iterative process. The Algorithm-1 illustrates the algorithm of SVM-RFE. This algorithm is trained by a linear SVM and the features are removed recursively using the smallest ranking criterion. The input or initial subset is randomly selected from the vector space. Then, it trained by SVM using the initial feature subset. At the end of the process, the algorithm ranks the feature based on the SVM score, *W*, features with the smallest SVM score will be removed in a recursive manner.

Algorithm-1; Support Vector Machine- Recursive Feature Elimination

Input: Initial feature subset, *S*= {*1, 2, 3…..n*}
1: Set *R*= {};
2: **repeat**
3: Train SVM using *S;*
4: Compute the Weight Vector using (5);
5: Compute the Ranking Criteria, $Rank = W^2$*;*
6: Rank the features as in a sorted manner;
New_{rank} *= sort (Rank);*
7: Update the feature Rank List;
Update $R = R + S\ (New_{rank})$*;*
8: Eliminate the feature with the smallest rank;
Update S = S - $S\ (New_{rank})$
until S is not empty
Output: Ranked list according to the smallest weight criterion, *R*

4.0 Results and the Discussion

This section presents the findings and analysis of the study. The main objective of this research is to study the impact of the embedded TF-IDF and SVM-RFE as the new feature selection technique in text classification accuracy. The proposed technique performances are evaluated based on the accuracy result obtained from the classification process. A linear SVM is implemented as a classifier to observe the result. The reason for using this classifier is that it is suitable for text classification problems that are linearly separable; also, it is good when there are high dimension features, whereby, a text document is known as unstructured and having high-dimensional features. Furthermore, it is a simple and faster algorithm since there is a fewer number of the parameter to optimize.

A text collection is retrieved from an established repository namely, UCI Machine Learning Repository. It is a collection of 200 twitter posts. The twitter posts have been annotated as 0 for the negative tweet and 1 for positive tweets, as shown in Table-2. 80 samples, are labeled as negative tweets and 120 are positive tweets samples. The samples are split into 80% of training (160 samples) set and 20% (40 samples) for testing.

Table 2- Example of features in documents

Feature Id / Tweet Id	**F1**	**F2**	**F3**	**F4**	**F5**	**……..**	**F547**	**F548**	**F549**	**Label**
T1	0.11	0.47	0.61	0.15	0.53	……..	0.58	0.41	0.00	1
T2	0.29	0.33	0.00	0.81	0.00	……..	0.00	0.00	0.31	0

T3	0.13	0.00	0.44	0.66	0.00	……..	0.61	0.17	0.00	0
T4	0.21	0.51	0.33	0.00	0.19	……...	0.12	0.91	0.00	0
T5	0.99	0.00	0.77	0.91	0.25	……...	0.33	0.11	0.10	1
⋮ ⋮ ⋮ .	⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮	……..	⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮
T200	0.00	0.21	0.10	0.32	4	……..	0.00	0.00	0.00	1

In this study, four sets of experiments were set up to access how the number of features affects the classification performances. The 549 features were extracted from the feature extraction phase prior to the feature selection phase as shown in Table-2. The selected features obtained from the proposed feature selection technique on the training dataset are grouped into four categories based on N-top rank; 10% of top rank, 25% of top rank, 50% of top rank, and 75% of top rank. 100% top rank is not implemented in these experiments since it is impossible in one Twitter post to have all the features extracted. These four groups of ranks are used in text classification. The reason for grouping the features into the N-top rank is because only the high score features will be considered for text classification. It is also to observe the impact of feature selection on the classification accuracy.

Table-2 shows an example of the pre-processed features of the tweet samples. For instance, there are 200 Twitter post samples indicated by T1 to T200. The features are written from F1 to F549, which indicates that there are 549 features extracted. The features are words exist in the tweets, for example, 'waste', 'tough', 'convert' etc. Whereas, the numbers in the row, for example, 0.11, 0.47, 0,61, etc. are the TF-IDF score. It is calculated to represent the importance of feature for the respected to the sample or tweet. In the feature selection phase, the SVM-RFE ranked the pre-processed features with the TF-TDF score. Later in the testing phase, some of the ranked features will be removed based on the N-top rank grouped as mentioned earlier. For example, 75% top rank, which is only 75% of the top-ranked features, will be considered for classifying the testing dataset. The top-ranked features are based on the SVM score calculated and ranked by the SVM-RFE algorithm as shown in Figure-3.

As a comparison, the experiment with the same the same parameter setup as in previous techniques is conducted. The aim is to observe if this proposed technique is able to enhance the accuracy performance compared to the previous techniques. The previous techniques that have been compared are single TF-IDF and embedded of TF-IDF and SVM. For TF-IDF, there is no feature selection technique is implemented. Therefore, all of the extracted features are considered in the classification. However, for the embedded of TF-IDF and SVM technique, the feature selection technique is implemented by using N-top ranked approach. In this approach, it selects four groups of N-top features from the feature list. The feature list is calculated and ranked by SVM during the selection process.

Table 3– The Accuracy Performances of The Proposed Technique

N-top rank (%)	**Accuracy (%)**		
	TF-IDF	**TF-IDF + SVM**	**TF-IDF+SVM-RFE**
10	-	60.0	76.0
25	-	82.0	88.0
50	-	92.0	94.0
75	-	96.0	98.0
All features	84.5	-	-

Table-3 summarises the experimental results obtained for all the techniques tested. The TF-IDF technique achieves the accuracy performance of only 84.5% when the same classifier is used. Since no feature is removed, it leads to unpromising results due to the presence of insignificant features in the dataset that distract the accuracy performance. At this point, the advanced feature selection technique is required to enhance the classification accuracy.

Later, the embedded of TF-IDF and SVM is tested. Overall, the results are better than the previous technique, which is more than 90% of accuracies are achieved except for 25% and 10% N-top rank with 82% and 60% accuracy, respectively. The low accuracies obtain due to many important features

that have been removed. The highest accuracy of 96% is obtained with 75% of N-top-rank features is chosen. At this point, if a feature matrix evaluation is involved, better classification accuracy is foreseen.

For the proposed technique, which is the TF-IDF+SVM-RFE feature selection technique, achieves better classification accuracies compared to the other two techniques in each N-top rank. It obtains the highest accuracy from 75% of N-top rank features with 98% accuracy. It can be concluded that Recursive Feature Elimination (RFE) as a feature matrix evaluation assist in increasing the classification accuracy by generating a more significant feature weighted ranking. The number of features plays an important role in classification. A large number of features does not guarantee the best classification performances and vice versa. Yet, the optimum number of features will generate optimum classification accuracy.

Besides, a comparison of the accuracy performance with the related works is summarising in Table 4. This table consists of the highest classification performance reported in the related works and our tested and proposed technique. From Table 4, the proposed technique by [17] achieves the lowest classification accuracy compared to others. They improved the TF-IDF by solving the confusion issues when the uneven class distribution exists in the dataset. However, it is tested on the biggest number of samples with a relatively fewer number of features. Hence, it might cause poor classification accuracy. Meanwhile, this proposed technique is tested on a relatively small sample. Thus, the classification accuracy is somewhat promising.

In the work done by [18], they implemented TF-IDF+SVM that similar to our tested technique. The difference is they measure the classification performance using F-measure. F-measure combines recall and precision evaluation with equal weight. Whereas, for LFW+DDR+HA done by [19]. Length Feature Weight (LFW) is a new feature weighting technique introduced to overcome some drawbacks in the TF-IDF technique while DDR is a new dynamic dimension reduction. It reduces the number of features used in the text clustering which assists to improve the performance of the tested feature selection algorithms (Genetic algorithm (GA), harmony search (HS) algorithm, and particle swarm optimization (PSO) algorithm). The combination of LFW, DDR, and HA obtained the best classification performance in terms of accuracy and F-measure. The proposed technique is tested in eight datasets. Nevertheless, the highest accuracy achieved is only 78.91%.

Table 4– The Accuracy Performances Comparison

Technique	Dataset	Number of Samples	Number of Features	Accuracy (%)	F-measure (%)
TF-IDF +SVM [30]	BBC News	5070	Not specified	-	97.84
LFW+DDR+HA [31]	Web pages	333	Not specified	78.91	72.81
Improved TF-IDF [11]	20-Newsgroups	20000	2000	69.50	-
TF-IDF	Twitter	200	543	84.50	-
TF-IDF+SVM	Twitter	200	407	96.00	-
Proposed technique	Twitter	200	407	98.00	-

5.0 Conclusion and Future Work

In this paper, it generates feature sets from a task call feature selection. The proposed feature selection technique is the two-stage enhanced embedded feature selection consists of TF-IDF and SVM-RFE. TF-IDF will extract the feature using a feature-weighted approach. Later, SVM-RFE will evaluate the feature subset from the previous task using recursive feature elimination producing a feature ranking. The SVM-RFE will rank the remaining features based on the SVM score. For the later text classification process, the proposed method only selects N-top rank features. Lastly, the SVM classifier is applied. In conclusion, when selecting 75% of the top-rank feature, it shows the optimum classification accuracy. As for future work, the proposed technique will be evaluated on the bigger and multiple datasets to access its capability on more high-dimensional data. In addition, it also will be evaluated using other performance measures such as precision, recall, and F-measure.

Acknowledgment

The authors would like to express their gratitude to the Ministry of Higher Education of Malaysia, under the Fundamental Research Grant Scheme (FRGS/1/2019/ICT02/UMP/02/1) for supporting this study

References

1. A. AL-Saffar, B. Sabri, H. Tao, S. Awang, M. Abdul Majid and W. Al-Saiagh. **2016**. Sentiment analysis in Arabic social media using association rule mining. *Journal of Engineering and Applied Sciences*, **11**: 3239-3247. 2016
2. V. Bhaskar. **2017**. "Mining Crisis Information : A Strategic Approach for Detection of People at Risk through Social Media Analysis," *Int. J. Disaster Risk Reduct.*, 2017.
3. C. Bucur. **2015**. "Using Opinion Mining Techniques in Tourism," *Procedia Econ. Financ.*, **23**(October 2014): 1666–1673, 2015.
4. N. S. M. Nafis and S. Awang. **2020**. The Impact of Pre-processing and Feature Selection on Text Classification. In *Advances in Electronics Engineering* (pp. 269-280). Springer, Singapore. 2020.
5. W. AL-Saiagh, S. Tiun, A. AL-Saffar, S. Awang and A. S. Al-Khaleefa. **2018**. Word sense disambiguation using hybrid swarm intelligence approach. *PloS one*, **13**(12), e0208695. 2018.
6. T. Al-Moslmi, S. Gaber, A. Al-Shabi, M. Albared, and N. Omar. **2015**. "Feature selection methods effects on machine learning approaches in Malay sentiment analysis," *Proc. 1st ICRIL-Int. Conf. Inno. Sci. Technol. (IICIST)*, no. October, pp. 1–2, 2015.
7. J. Tang, S. Alelyani, and H. Liu. **2014**. *Feature Selection for Classification: A Review*. 2014.
8. A. Jovic, K. Brkic, and N. Bogunovic. **2015**. "A review of feature selection methods with applications," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron.*, pp. 1200–1205, 2015.
9. S. Gunal. **2012**. "Hybrid feature selection for text classification," *Turkish J. Electr. Eng. Comput. Sci. ·*, **20**(2): 1296–1311, 2012.
10. B. Trstenjak, S. Mikac, and D. Donko. **2014**. "KNN with TF-IDF based framework for text categorization," *Procedia Eng.*, **69**: 1356–1364, 2014.
11. W. Dai. **2018**. "Improvement and Implementation of Feature Weighting Algorithm TF-IDF in Text Classification," **147**(Ncce): 583–587, 2018.
12. Z. Yin, Z. Fei, C. Yang, and A. Chen. **2016**. "A novel SVM-RFE based biomedical data processing approach : basic and beyond," in *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, 2016, pp. 7143–7148.
13. Z. M. Hira and D. F. Gillies. **2015**. "A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data," **2015**(1), 2015.
14. H. Uğuz. **2011**. "A two-stage feature selection method for text categorization by using information gain, principal component analysis, and genetic algorithm," *Knowledge-Based Syst.*, **24**(7): 1024–1032, 2011.
15. L.-P. Jing, H.-K. Huang, and H.-B. Shi. **2002**. "Improved Feature Selection Approach Tfidf in Text Mining," in *Proceedings of the First International Conference on Machine Learning and Cybernetics*, 2002, no. November, pp. 4–5.
16. L. H. Patil and M. Atique, "A Novel Approach for Feature Selection Method TF- IDF in Document Clustering," pp. 858–862, 2012.
17. N. Azam and J. Yao. **2012**. "Comparison of term frequency and document frequency-based feature selection metrics in text categorization," *Expert Syst. Appl.*, **39**(5): 4760–4768, 2012.
18. Y. An. **2015**. "Weighted Document Frequency for Feature Selection in Text Classification," 2015.
19. Z. Liu, S. Liu, L. Liu, J. Sun, X. Peng, and T. Wang. **2015**. "Sentiment recognition of online course reviews using multi-swarm optimization-based selected features," *Neurocomputing*, pp. 1–10, 2015.
20. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. **2002**. "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, **46**(1–3): 389–422, 2002.
21. J. Wang, G. Shan, X. Duan, and B. Wen. **2011**. "Improved SVM-RFE Feature Selection Method for Multi-SVM Classifier," pp. 1592–1595, 2011.
22. T. Hamed, R. Dara, and S. C. Kremer. **2014**. "An accurate, fast embedded feature selection for SVMS," *Proc. - 2014 13th Int. Conf. Mach. Learn. Appl. ICMLA 2014*, pp. 135–140, 2014.

23. J. Miao and L. Niu. **2016**."A Survey on Feature Selection," *Procedia Comput. Sci.*, **91**(Itqm): 919–926, 2016.
24. J. Cai, J. Luo, S. Wang, and S. Yang. **2018**. "Feature selection in machine learning: A new perspective," *Neurocomputing*, **300**: 70–79, 2018.
25. S. Vora and H. Yang. **2017**. "A comprehensive study of eleven feature selection algorithms and their impact on text classification," *2017 Comput. Conf.*, no. July, pp. 440–449, 2017.
26. C. H. P. Ferreira, D. M. R. de Medeiros, and F. O. de França. **2018**. "DCDistance: A Supervised Text Document Feature extraction based on class labels," 2018.
27. J. Apolloni, G. Leguizamón, and E. Alba. **2016**. "Two-hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments," *Appl. Soft Comput. J.*, **38**: 922–932, 2016.
28. P. A. Hassan. **2013**. "A Hybrid Feature Selection approach of ensemble multiple Filter methods and wrapper method for Improving the Classification Accuracy of Microarray Data Set," **3**(2): 185–190, 2013.
29. N. D. Cilia, C. De Stefano, F. Fontanella, S. Raimondo, and A. Scotto. **2019**. "An Experimental Comparison of Feature-Selection and Classification Methods for Microarray Datasets," pp. 1–13, 2019.
30. S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani. **2016**. "A novel text mining approach based on TF-IDF and support vector machine for news classification," *Proc. 2nd IEEE Int. Conf. Eng. Technol. ICETECH 2016*, no. March, pp. 112–116, 2016.
31. L. M. Abualigah, A. T. Khader, M. A. Al-Betar, and O. A. Alomari. **2017**. "Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering," *Expert Syst. Appl.*, **84**: 24–36, 2017.