Machine Learning and Feature Selection Approaches for Categorizing Arabic Text: Analysis, Comparison, and Proposal

Ayat Elnahas[*1], Mohamed Nour[*2], Nawal A. El-Fishawy[**3], Maha Tolba[**4]
[*]*Department of Research Informatics, Electronics Research Institute, Cairo, Egypt*
[**]*Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, Menoufia, Egypt*
[1]eng_ayatelnahas@yahoo.com
[2]mnour99@hotmail.com
[3]nelfishawy@hotmail.com
[4]maha_saad_tolba@yahoo.com

Abstract-*This work adopts some classification approaches for categorizing Arabic text. The approaches are operated on two datasets as test-beds. A comparative study is done to evaluate the performance of the adopted classifiers. Some feature selection methods are also analyzed, investigated, and evaluated. Selecting the most significant features is important because the huge number of features may cause performance degradation for text classification. A comparative study is done among the adopted feature selection methods for classifying Arabic documents.*
Moreover, a modification is done on the feature selection approaches by doing amalgamation for the chosen methods. A novel method is also proposed for selecting the most appropriate features. The method is based on the semantic fusion and multiple-words (SF-MW) for constructing the features. A comparison is done among the adopted feature selection methods and the proposed one.
The experimental results show that the best performance was for the SVM classifier compared to the KNN and NB classifiers. The combination among the adopted feature selection methods presents better results compared to the individual adopted ones. The proposed feature selection method (SF-MW) is promising as it reduced the features and achieved higher classification accuracy. The accuracy improvement was about 22% for the two chosen Arabic test-beds which contain 1246 and 1500 documents respectively. The proposed method is expected to be also efficient for other Arabic and English datasets.
Keywords: ***Classification Algorithms, Feature Selection, Multiple-Arabic-Words, Semantic Fusion, Datasets, and Measurable Evaluation Criteria.***

1 INTRODUCTION AND RELATED WORK

Text classification can be briefly defined as assigning document or text to predefined categories or classes based on their contents. The terms text classification, text categorization, document categorization and document classification are used interchangeably in this work. Text classification is considered one of the data mining applications. Arabic text classification became a very important task as the uploaded Arabic text and documents on the Internet are dramatically increasing. Arabic text classification is also important for several applications; examples of those applications are: e-mail filtering, opinion mining, e-mail routing, news monitoring and others [1], [2].

There are several algorithms that can be used to classify documents. Examples of such algorithms include; but not limited to; K-nearest neighbor (KNN), support vector machine (SVM), logistic regression (LR), random forest (RF), Naïve Bayes (NB), decision tree (DT), artificial neural network (ANN), and others [2],[4-15]. One of the main problems of classifying documents is the huge number of features which are describing a dataset. A huge number of features; in most cases; may reduce the efficiency of the adopted classifiers and also consume more time. So, the feature selection process is very important to choose a subset of high significant features and eliminate the non-significant ones [16]. Moreover, several research works were presented regarding text classification, machine learning algorithms, and feature selection methods. Examples of the research efforts are briefly mentioned as follows:

[2] presented a comparative study among the performance of some classification algorithms using feature selection with and without stemming. The adopted algorithms were k-nearest neighbors (KNN), Naïve Bayes (NB), and Naïve Bayes Multinomial (NBM). The adopted classifiers were operated on the BBC Arabic dataset. The results were presented in terms of precision, recall, F-measure, accuracy and training time.

[17] discussed three classification techniques which were applied on Arabic datasets. A comparative study was done for the adopted techniques to evaluate their performance. The study fixed the number of documents for all categories in training and testing phases. The experimental results reported that the support vector machine is promising.

[18] mentioned that text classification methods had emerged as a natural result of the existence of a massive amount of varied textual information written on the web. The authors presented a survey study of some research works for classifying Arabic

text using classical data representation, bag of words and conceptual representation. The data representation was based on semantic resources such as Arabic Word Net and Wikipedia.

[19] presented a text classification algorithm with semantic features. The authors adopted using the support vector machines and Word2vec. Word2Vec brings extra semantic features that help in text classification. The authors combined both Word2vec and term frequency-inverse document frequency (TF-IDF). The practical results showed that their approach outperforms TF-IDF only because Word2Vec provides complementary features.

[20] presented an overview of some feature selection methods. The objective was to provide a generic overview of variable elimination which was applied to a wide array of machine learning problems. The focus was on filter, wrapper and embedded methods. The authors applied some feature selection methods on common datasets to show the applicability of feature selection techniques.

[21] proposed a multivariate filter method for feature selection in text classification. The method focused on the reduction of redundant features using minimal-redundancy and maximal-relevancy concepts. The method didn't employ any learning algorithm to evaluate the usefulness of the selected features, so, it could be categorized as a filter method. Several experiments were implemented on three datasets to assess the effectiveness of the proposed method. From the results, the proposed method outperformed the other adopted ones.

[22] presented a comparative study to evaluate the performance of some techniques for classifying Al-Hadith Al-Shareef. This work was analyzed with some software Arabic tools such as: stem-Darwish, stem-Alex, Khoja stemmer, Quadrigrams, Trigrams and a disambiguation tool of Arabic-morph. Moreover, three classification algorithms were implemented on WEKA toolkit mainly: decision tree, Naïve Bayes, and support-vector machines. The authors used the cross-validation to evaluate the classification performance. From the results, the Khoja's stemmer outperformed the other adopted tools and the SVM classifier achieved the highest accuracy followed by the Naïve Bayes and then the decision tree classifier.

[23] mentioned that data mining involves systematic analysis of large data sets. The classification is used to manage data and predict new ones. The authors focused on using J48 algorithm which was used to create univariate decision tree. The research work discussed the idea of multivariate decision tree using more than one attribute at each internal node. The objective was to get depth knowledge and data mining techniques. The experimental work was implemented using WEKA software tool.

[24] presented a comparative study among three classifiers mainly ID3, C4.5 and C5.0. Due to the comparison, C5.0 gave better accuracy and efficiency. The authors proposed a classifier system based on C5.0 to classify the result set with high accuracy and low memory usage. The proposed system selected the relevant features which were useful in the model construction. Cross-validation gave more reliable estimate of prediction. The proposed system achieved a reasonable improvement in accuracy and reduced the error rate.

[25] mentioned that efficient and effective techniques and algorithms are important to discover useful pattern in unstructured text. Text mining is important to extract meaningful information from text. The authors in their research work described several text mining tasks and techniques including text pre-processing, classification and clustering. The authors explained text mining in biomedical and health care domains.

The organization of this work will be as follows: Section 2 briefly presents the process of Arabic text classification. Section 3 presents a brief overview of three classifiers mainly: decision tree, Naïve Bayes, support vector machine respectively. Some approaches of feature selection methods are adopted in Section 4. Section 5 presents the implementation work and performance evaluation of the adopted classifiers. Some enhancement approaches for feature selection are conducted in Section 6. Section 6 also presents a proposal of a semantic fusion method for features' selection. Finally, the discussion of results and conclusions are reported in Section 7.

2 THE PROCESS OF ARABIC TEXT CLASSIFICATION

Several research works were presented for classifying English text and other natural languages. The efforts for classifying Arabic text are considered limited compared to those done for English [26]. Arabic language is the main language in the Arab world and the secondary language in many other countries [17].

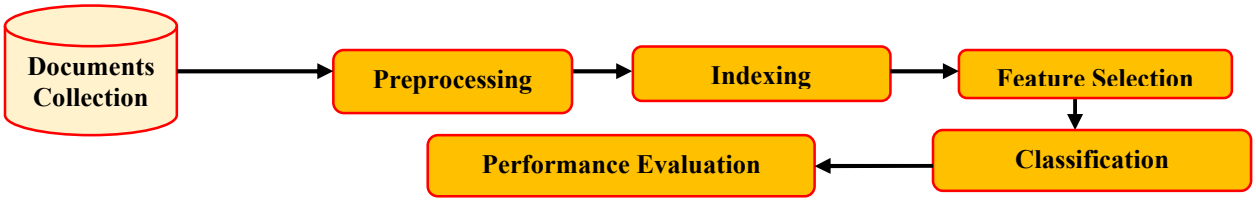

Figure.1 Text Classification Process [28] [29]

The Arabic alphabet contains 28 letters and the Arabic words are written horizontally from right to left [Laila Khreisat, 2008][27]. The Arabic language has very complex morphology due to its inflection and derivation. Representing Arabic words to their roots is very important and definitely will reduce the number of words [17]. The overall classification process can be applied to classify Arabic text/ documents. The main steps of the classification process are briefly mentioned as shown in Figure1.

A. Collection of Documents

The collection of documents is an important step for the classification process. The documents collection is considered a test-bed dataset. The chosen datasets were taken from the websites http://mlg.ucd.ie/datasets/bbc.html and https://old.datahub.io/dataset/cc-aljazeera. The first dataset is called Arabic BBC and it contains 1250 documents. The second dataset name is Aljazeera and it contains 1500 documents. For the chosen datasets, each document is represented as a vector (or instance). The instances are described by 500 features and 600 features for the BBC and Aljazeera datasets respectively. The instances of the datasets are belonging to four classes in BBC and five classes in Aljazeera respectively. Each dataset is partitioned into two parts: one for training and one for testing. The training part is used to build the training model while the other part is used for testing. The training and testing instances are 70% and 30% respectively. Each document in the training phase is assumed to be assigned to a predefined class (labeled documents). These documents are used to train the classifier to take appropriate categorization decisions [30]. The testing phase; on the other hand; comes after the training phase. It is used for classifying the unlabeled documents.

B. Preprocessing

The preprocessing steps are important for extracting the documents features. This involves many themes such as tokenization, stop-word removal and stemming. For more details, the reader can refer to [31].

C. Indexing

To prepare the indexing of the datasets' documents, each document is transformed from the full text version to a document vector. Examples of the commonly used models for document representation are: Boolean weighting model, vector space model, Latent Semantic Indexing (LSI), and TF-IDF weighting, and others [32]. In this work, the vector space model was adopted and applied.

D. Feature selection

The feature selection process is important for text classification. The feature selection process aims to select the most significant features of the original document [33]. The chosen number of features plays a vital role in the classification accuracy. More details about the feature selection methods and classification approaches are mentioned in the next sub-sections.

E. Document Classification

Documents can be classified using the machine learning approaches such as Neural Network, Decision Tree, K-nearest neighbor (KNN), Naïve Bayes, Support Vector Machines (SVMs), and others [6]. Three of such classification approaches are discussed in section 3.

F. Performance Evaluation

The performance evaluation of the whole classification process is important. Performance evaluation is done by considering a set of measurable evaluation criteria. This involves recall, precision, F-measure, accuracy, classification time and others [19].

3 THE ADOPTED CLASSIFICATION APPROCHES

A classifier is an algorithm that implements classification. The binary classifier or two-class classifier is the simplest one. When more than two different classes are used, a multi-class classifier is created to determine the classes where a document belongs to [33]. There are several classification approaches; the adopted ones are briefly mentioned as in the following.

A. Decision Trees

A decision tree classifier is a tree in which internal nodes are labeled by attributes, branches departing from them are labeled by testing the weight that an attribute has in the test document, and leafs are labeled by categories [34], [8], [23]. Decision trees are designed using a hierarchical division of the underlying data space with the use of different text features. The hierarchical division is designed to create class partitions which are more skewed in terms of their class distribution. For a given text instance, the partition is determined where that instance most likely belongs to [35], [36].

There are many algorithms for creating a decision tree [37]. J48 will be discussed and implemented in this work. The necessary steps for constructing the tree are: -

- Check whether all cases belong to the same class, then the tree is a leaf and is labeled with that class.

- For each attribute, calculate the Entropy and information gain.
- Find the best splitting attribute (depending upon current selection criterion).

Entropy is a measure of disorder of data and it is measured in bits. This is called measurement of uncertainty in any random variable [23]. Entropy is calculated as: -

$$E(S) = -\sum_{i \in X} p(i) \log_2 p(i) \quad (1)$$

where S is the current data for which entropy is being calculated; X is the set of classes in S; and
p(i) is the proportion of the number of elements in class i to the number of elements in set S.
Moreover, Information Gain (IG) is used to measure the association between inputs and outputs. It is a state to state change in information entropy [https://mariuszprzydatek.com/2014/10/31/ measuring-entropy-data-disorder-and-information-gain]. Information gain can be calculated as: -

$$IG(A,S) = E(S) - \sum_{t \in T} p(t) E(t) \quad (2)$$

where E(S) is the Entropy of set S; t is the subsets created from splitting set S by attribute A such that S= Ut €Tt ; p(t) is the proportion of the number of elements in t to the number of elements in set S; and E (t) is the Entropy of subset t.

B. Naive Bayes Classifier

Naïve Bayes (NB) classifier is a simple probabilistic classifier based on applying Bayes' theorem [38]. NB algorithm is called Naive because it makes the Naive assumption that the occurrence and frequency of the attributes are not dependent upon each other [39]. When NB classifier is used for text classification, the following equation is adopted [17] [40].

$$P(\text{class} \mid \text{document}) = \frac{P(\text{class}).P(\text{document} \mid \text{class})}{P(\text{document})} \quad (3)$$

where,
P(class | document) is the probability that a given document D belongs to a given class C.
P(document) is the probability of a document while P(class) is the probability of a class (or category) and it is computed as follows:

$$P(\text{class}) = \frac{\text{Number of documents in the category}}{\text{documents number in all categories}} \quad (4)$$

P(document | class) represents the probability of a document given class.

$$P(\text{document} \mid \text{class}) = \Pi\, p(\text{word i} \mid \text{class}). \quad (5)$$

$$P(\text{class} \mid \text{document}) = p(\text{class})\, \Pi\, p(\text{word i} \mid \text{class}). \quad (6)$$

where P(word i | class) is the probability that the i^{th} word of a given document occurs in a document from class C, and this can be computed as follows:

$$P(\text{word i} \mid \text{class}) = (T_{ct} + \lambda)/(N_c + \lambda V) \quad (7)$$

where T_{ct}: The number of times the word occurs in that category C; N_c is the number of words in category C; V is the size of the vocabulary table; and λ is the positive constant, usually 1, or 0.5 to avoid zero probability.

C. Support Vector Machine

Support vector machine (SVM) is one of the best well-known machine learning algorithms [9]. SVM can be used to analyze and categorize text data. The main principle of SVM is to determine separators in the search space which can best separate the different classes [17].
The main idea of this classifier is mapping the input points in N-dimension space into another higher dimensional space and then a maximal separating hyper plane is found. It aims to separate amounts of data based on the optimal hyper plane between vectors which are linearity separable. The separation process depends on the maximum distance between the two sides of hyper plane and the nearest vectors in the training data sample [32]. SVM can be implemented for multi-label classification [33].
The SVM is conducted using linear separable data by creating infinite number of separating line or separating plane if data is 3D and it used the best one. The best hyper plane is that one which has larger margin because it is more accurate at classifying future data tuples than the hyper plane with the smaller margin. SVM can also work using nonlinear data. The approach described for linear SVMs can be extended to create nonlinear SVMs for the classification of linearly inseparable data (also called nonlinearly separable data). Such SVMs are capable of finding nonlinear decision boundaries in input space [41]. To obtain a nonlinear SVM, there are two main steps. In the first step, the original input data is transformed into a

higher dimensional space using a nonlinear mapping. The second step searches for a linear separating hyper plane in the new space. The linear algorithm for classification is defined as:

$$f(x) = W.P + b \quad (8)$$

$$a=\text{sign}\,(f(W.P+b)) \quad (9)$$

where P is the vector of the training data-set and b is the bias to manipulate the decision boundary of the linear hyper plane. W is the weight vector for the best hyper-plane. The class of P after training (test instance) is found using the following linear decision or activation function (a).

$$a = \text{sign}\, f(x) \quad (10)$$

where sign is the Continuous Log-Sigmoid Function, sig(n) [33] [32], [41].

4 APPROCHES OF FEATURE SECTION

Machine learning approaches are using feature selection methods. A feature represents a property of a process or system that has been constructed from the original input variables [15]. It is difficult to learn good classifiers before removing the unwanted features due to the huge size of the data. Reducing the number of irrelevant or redundant features can help in getting a better insight into the underlying concept of a real-world classification problem [42]. Feature selection helps in understanding data, reducing the effect of dimensionality, reducing computation requirement and improving the predictor performance [20]. Feature selection aims to choose a subset of features to improve prediction accuracy [43].

The feature selection methods can be divided into two groups: wrapper-based and filter-based methods. The wrapper-based methods use a specific search method to find a subset of features around the feature space [21]. Such methods evaluate the usefulness of features based on the performance of a machine learning algorithm to optimize the predictive performance. The filter-based methods select the features based on evaluation metrics [44]. The filter approaches can be classified into two groups individual feature measures and group feature measures. Individual feature measures are used to evaluate the relevance of features independently. Although, these methods can effectively identify irrelevant features based on the value of this metric and ranking of the features, they are unable to remove redundant ones. The group feature measures consider correlation between features in their process, and thus can handle both irrelevant and redundant features. The performance of group feature measures is better than that of the individual feature measures but individual feature measures are more efficient in terms of running time [21], [44].

Now; some feature selection methods are briefly presented as in the following sub-sections.

A. Term Weighting (TF-IDF)

The term frequency-inverse document frequency (TF-IDF) is the common weight scheme for document representation and it is used to calculate the term weighting. Each document di is represented as a vector of terms weights wi,j as follows:

$$di = (wi,1,wi,2, \ldots, wi, j, \ldots, wi,t) \quad (11)$$

In each document, the term weighting is assigned for each term according to the term frequency. The formula of term weighting is found in [45].

B. Information Gain (IG)

Information Gain (IG) is utilized in text mining and it measures the goodness of a feature. The feature reduction methods aim to determine and apply the most useful features for distinguishing the different classes of a given feature space [4]. IG is good for an attribute's relevance. IG measure can suggest the importance of the features, by calculating the weight (relevance) of a feature in terms of the class features. The higher the weight of a feature, this feature is better. IG of a feature f is defined as the information gained by doing the split of the feature space based on that particular feature. The formula of IG is found in [21] and [4].

C. Chi-square

Chi-square is a nonparametric statistical filter method that is used to compute the lack of independence between the distributions of observed frequencies and the theoretically expected frequencies. It evaluates features individually by measuring their chi-squared statistic with respect to the classes. The value of the chi-square statistic is found in [46].

D. Gini Index (GI)

Gini index (GI) is a global feature selection method for text classification. It can be considered as an improved version of the attribute selection method used in the construction of decision tree. The formula of GI is found in [21].

5 IMPELEMENTATION AND PERFORMANCE EVALUATION OF THE ADOPTED CLASSIFIERS

The Arabic text can be classified using the previous mentioned classifiers: Decision Tree (J48), Naïve Bayes, and Support Vector Machine. Some feature selection methods also will be applied to choose that most appropriate and significant ones. Such features are based on TF-IDF, IG, GI, and Chi-Square methods. Moreover, some preprocessing steps should be done to prepare the datasets for classification. The preprocessing operations are briefly mentioned as follows:

Tokenization: This occurred by the tokenizer that discretizes and separates tokens.

Removing stop words: in this step the rejection words are removed such as punctuation marks, numbers, and words in Latin characters, abbreviation, pronouns, and single letters.

Stemming (Lucene stemmer): After removing unnecessary words from the documents the stemming technique is used to extract word root.

As mentioned before, the adopted datasets are: BBC Arabic news and Aljazeera news. The BBC Arabic dataset contains 1246 documents with size of 22.6 MB. The dataset contains four categories namely {'World News', 'economy', 'sport', 'Science'}. Figure 2 presents the distribution of the dataset. The dataset is split as follows: 70% of all documents are dedicated for training and the remaining 30% for testing. On the other hand, Aljazeera dataset contains 1500 documents classified into five predefined categories {'Politics', 'Science', 'Sport', 'Economy', 'Art'}. Aljazeera dataset is also partitioned to be 70% for training and 30% for testing. Figure 3 presents the dataset distribution. WEKA software tool was used in the implementation work. WEKA is a popular free and open source software tool for machine learning purposes written by the Java programming language. The measurable criteria such as accuracy, precision, recall, and f-measure are adopted for performance evaluation [32]. The percentage values of precision, recall, and F-measure for each class in the BBC-Arabic news were reported for the J48, Naïve Bayes and SVM classifiers respectively. This is shown in Figure 4. The features of the BBC-Arabic news were selected according to the weighted values of the features using TF-IDF. The same experiments were done for the second dataset Aljazeera-news. Figure 6 presents the percentage values of the measurable criteria for the different classes for the same adopted classifiers. The best recall, precision, and F-measure achieved by SVM and Naïve Bayes were respectively (74%, 74%, 73%) and (73%, 72%, 72%). The worst performance classifier was for the J48. When the classifiers were applied on Aljazeera dataset the results confirmed those results obtained for the BBC dataset. From Figures 4 and 5, the performance of SVM is better than the other two classifiers because SVM first transforms the non-linear data into linear and then draws a hyper plane. Moreover, Figures 6 to 11 present respectively the recall%, precision%, and f-measure% for the adopted classifiers using the two datasets.

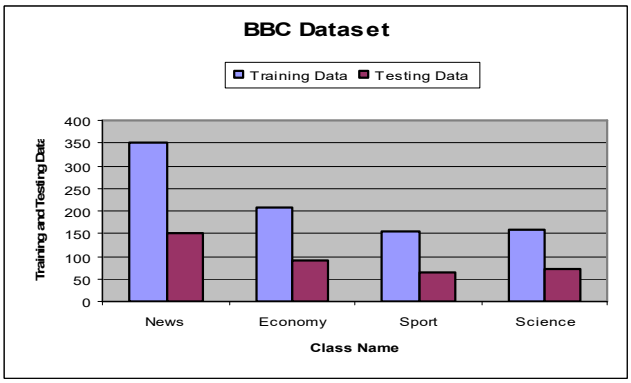

Figure 2: BBC Dataset: Training and Testing Data

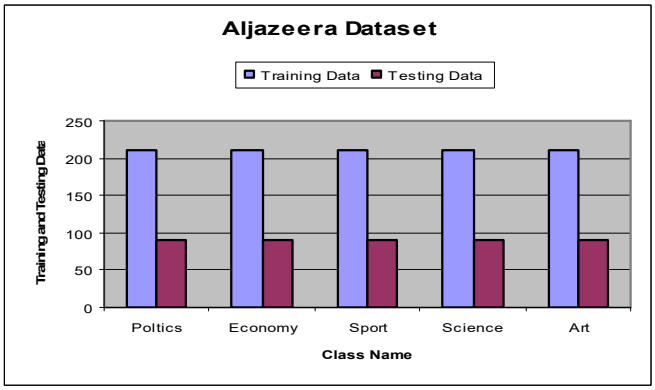

Figure 3: Aljazeera Dataset: Training and Testing Data

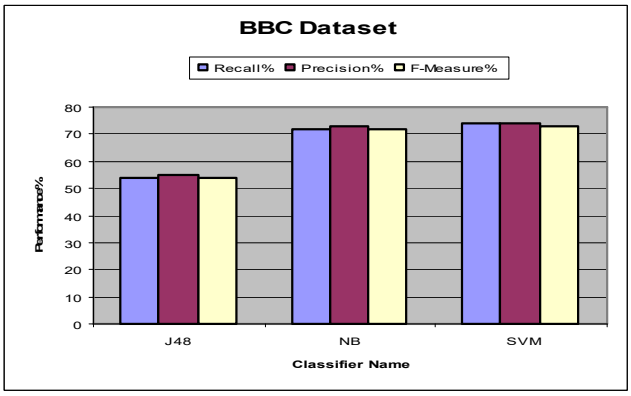

Figure 4: Classifiers Operated on BBC Dataset

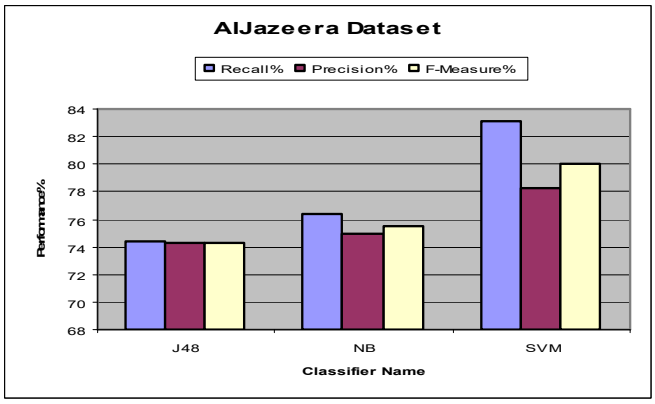

Figure 5: Classifiers Operated on Aljazeera Dataset

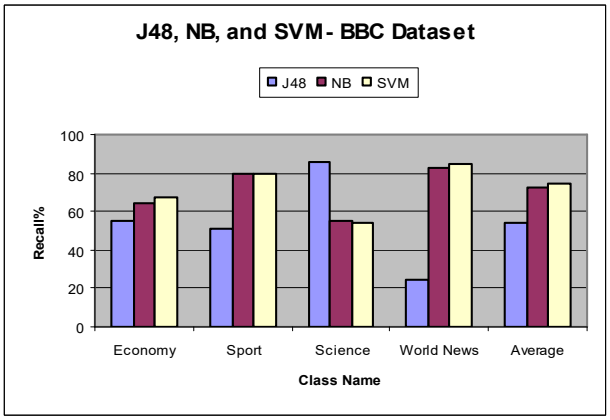

Figure 6: Recall% using BBC Dataset

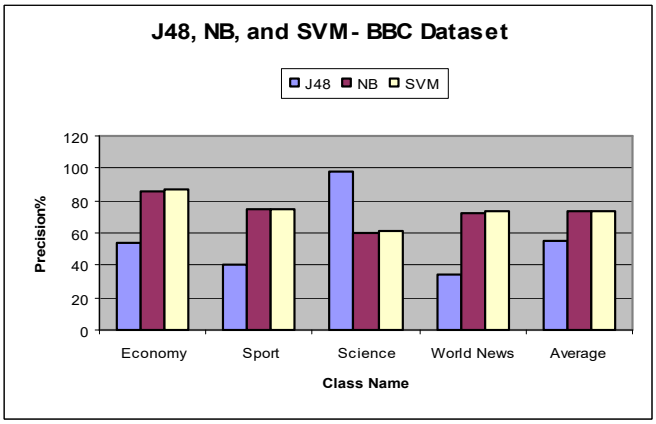

Figure 7: Precision% using BBC Dataset

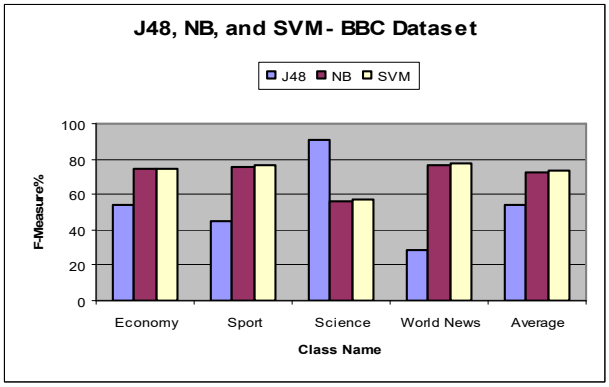

Figure 8: F-measure% using BBC Dataset

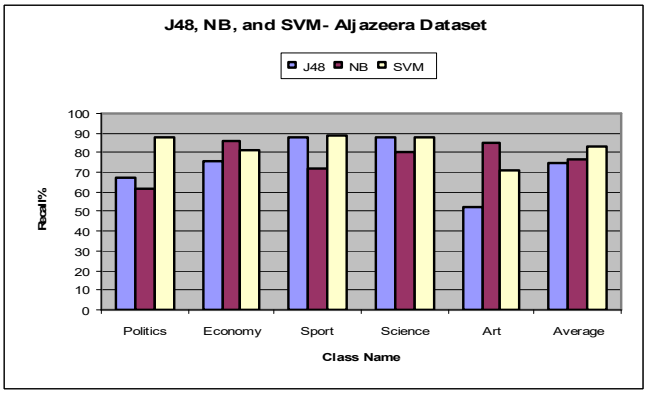

Figure 9: Recall% using Aljazeera Dataset

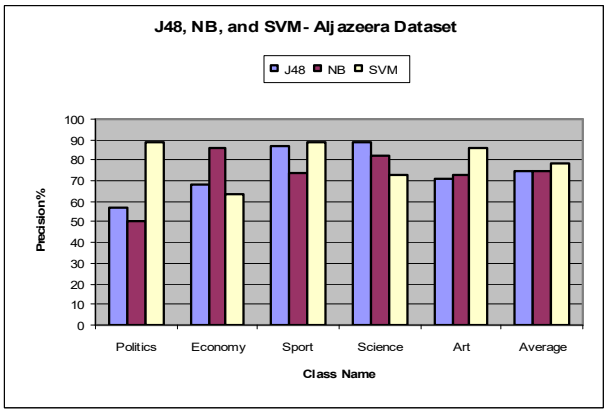

Figure 10: Precision% using Aljazeera Dataset

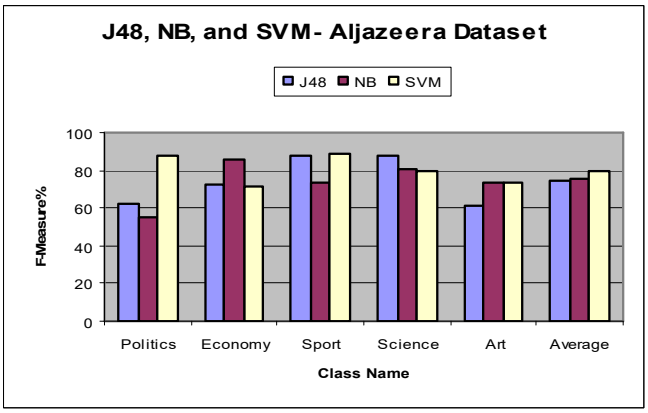

Figure 11: F-measure% using Aljazeera Dataset

6 THE PROPOSED ENHANSMENT APPROCHES FOR FEATURE SELECTION USING SVM CLASSIFIER

From the previous experiments, we found that there are several features with no effect in determining the classes. This may reduce the classification accuracy. When TF-IDF feature selection method was used, it reduced some features but the accuracy is still low. This was the reason to motivate us to combine some other feature selection methods.

A. Enhancement Using Combination of Chi Square and TF-IDF Feature Selection

Figures 12 and 13 show the changing values of precision%, recall%, and F-measure% when combining both the Chi Square and TF-IDF feature selection. This was applied on the BBC and Aljazeera Arabic datasets respectively. It is shown that the classification performance is improved when amalgamating both the chi-square and TF-IDF methods. The values of recall %, precision %, and F-measure % after combination are better than their corresponding values without combination. This is clear when applying the SVM classifier on the adopted datasets. During the experimental work, several iterations were run where the weight's threshold value is decremented by 5% in each iteration. i.e the weight threshold values in the adopted experiments were respectively 95%, 90%, 85%, 80%, 75%, and 70%. The best values of recall%, precision%, and F-measure% occurred when the threshold values were 70% and 80% respectively for the BBC Arabic and Aljazeera datasets.
Moreover, the highly weighted features are used while the other are rejected. This is done by adopting a threshold value. The TF in TF-IDF shows the relative frequency of a certain term or feature appearing in a document. The TF-IDF weight for each term feature is calculated by the formula shown in equation (12) [47]

$$TF - IDF(t_i, d_j) = tf(t_i, d_j) * \log \frac{N}{N(t_i)} \quad (12)$$

where N is the number of all documents; $N(t_i)$ is the number of documents in the collection in which the term t_i occurs at least once; and $tf(t_i,d_j)$ is the frequency of the term t_i in the document d_j.

The chi-square was combined with the TF-IDF method to obtain the most significant features. The Chi-square is a good statistical approach for determining the relevant category for a feature word. The Chi-square formula is shown in equation (13) where a higher Chi value means that a feature word has a stronger ability to identify a category [48].

$$x^2(t,c) = \frac{N' * (AD - BC)^2}{(A+c)(A+B)(B+D)(c+D)} \quad (13)$$

where N' is the size of the training set; A is the number of documents that belong to class c and contain the word t; B is the number of documents that do not belong to class c but contain the word t; C is the numbers of documents that belong to the class c but do not contain the word t, and D is the number of documents that do not belong to class c and do not contain word t.
Although the classification performance was improved when using the chi-square, the average values of recall%, precision%, and F-measure% did not reach 90%. Although high frequency words that appear in all categories have higher chi-values, they do not necessary make a sense for category distinctions. The chi-square takes into account the appearance of a word but not the frequency of the word in a document. The feature vectors selected by the chi-square may have high dimensionality.

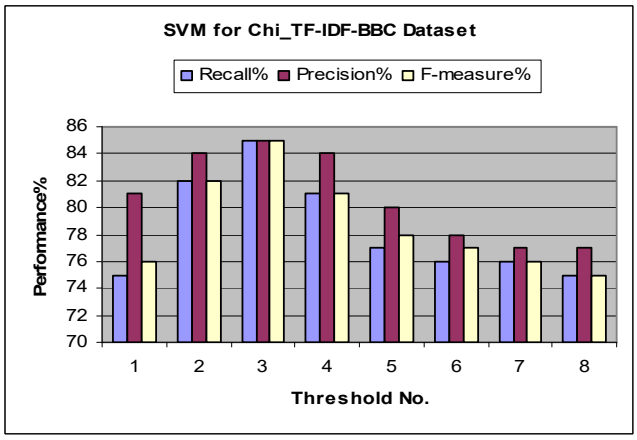

Figure 12: Feature Selection-1 using BBC Dataset

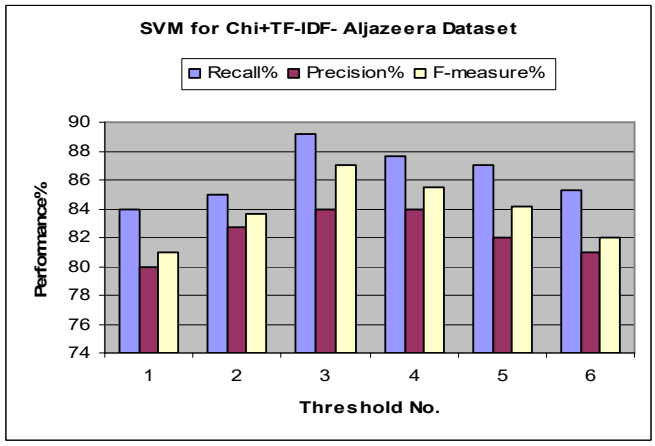

Figure 13: Feature Selection-1 using Aljazeera Dataset

B. Enhancement Using Gini Index and TF-IDF Feature Selection

Figures 14 and 15 show the changing of the performance metric values (recall%, precision% and F-measure%) when combining Gini index to TF-IDF feature selection. This was applied on the two chosen datasets. From the experiments it is shown that the classification performance using the combination outperforms that one using TF-IDF only. Several experiments were run by iteratively decrementing the weight's threshold by 5%. i.e the weight's threshold values were 95%, 90%, 85%, 80%, 75%, 70%, 65%, and 60%. The best values of the performance occurred at threshold values 70% and 80% for the BBC Arabic and Aljazeera datasets respectively. Moreover, Gini Index (GI) improved the performance as it enhanced the classification accuracy in selecting the most appropriate attributes. The GI measure; as defined before; can be written as in equation (14) [21]

$$GI(t) = \sum_{i=1}^{M} p(t|c_i)^2 \, p(c_i|t)^2 \tag{14}$$

where p(t|ci) is the probability of term t given ci, p(ci|t) is the probability of ci in the presence of t, and M is the number of classes.

The Gini Index measure is considered the impurity of feature towards classification. The higher impurity means bad features and such features should be rejected. The amount of computation in the Gini Index is low compared to the other adopted feature selection approaches. Another advantage of Gini Index is that it is appropriate and suitable for selecting features because it is relevant to the high dimensional data.

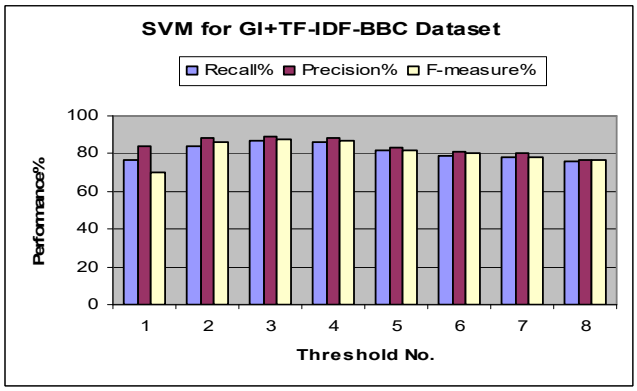

Figure 14: Feature Selection-2 using BBC

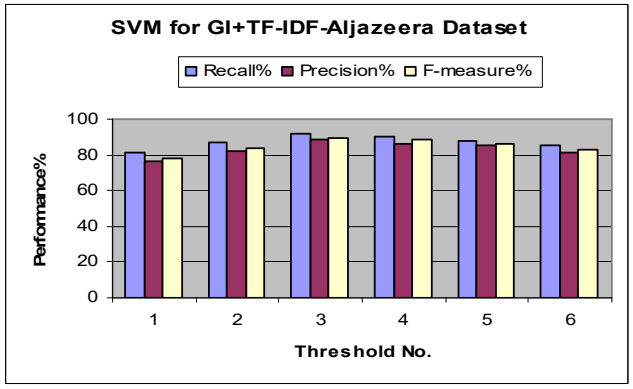

Figure 15: Feature Selection-2 using Aljazeera

C. Enhancement Using Information Gain and TF-IDF Feature Selection

Figures 16 and 17 show the changing of the values of (recall %, precision % and F-measure %) when combining Information Gain (IG) to TF-IDF feature selection methods. This was applied on the BBC Arabic and Aljazeera datasets. IG is a good statistical measure for feature selection. IG is used to indicate how significant each of the attribute is by calculating the weight of each attribute in terms of the class attributes. If the weight of an attribute is high, it is considered a significant feature. IG is used to measure the number of bits of information obtained for category prediction. This is done by determining the presence or absence of a term t in text documents. IG can be written as shown in equation (15):

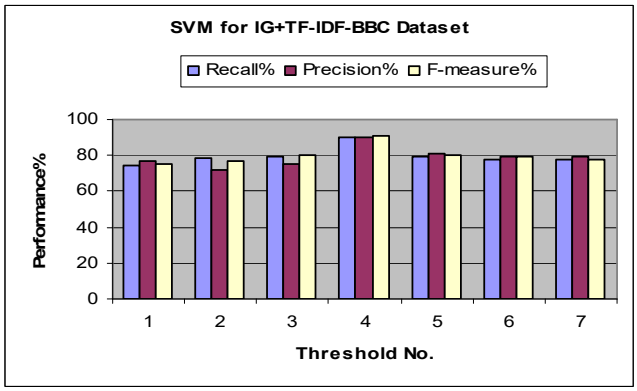

Figure 16: Feature Selection-3 using BBC Dataset

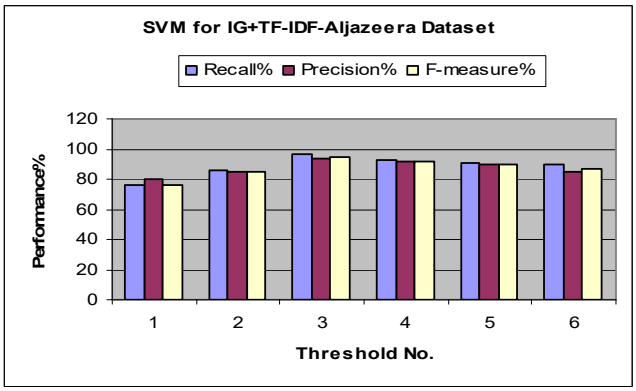

Figure 17: Feature Selection-3 using Aljazeera Dataset

$$IG(t) = -\sum_{i=1}^{M} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{M} p(c_i|t)\log p(c_i|t) + p(\bar{t})\sum_{i=1}^{M} p(c_i|\bar{t})\log p(c_i|\bar{t}) \quad (15)$$

where M is the number of categories; $p(c_i)$ is the probability of category c_i. p(t) and $p(t^-)$ are the probabilities of presence and absence of term t respectively. $P(c_i|t)$ and $P(c_i|t^-)$ are the conditional probabilities of class c_i considering presence and absence of t respectively [21], [7], [49], and [4]. IG is used to reduce the entropy caused by partitioning the objects according to an attribute. i.e The IG value of each feature is calculated based on the entropy of classes and feature values. The features are sorted in an ascending order according to their information gain values. The features with low information gain values are removed from the feature set. This approach is good to capture the features with high statistical quality from the documents. The values of recall%, precision%, and F-measure% for the combined work (TF-IDF and IG) outperform that one using only TF-IDF. From the experiments, the best values occurred at 80% of the total number of features for the BBC Arabic and Aljazeera datasets respectively.

D. Proposal of a Semantic Fusion Method (SF-MW) for Features' Selection

To classify documents, each document should be represented in an appropriate form; like a vector; to be easily handled and processed. The 'bag-of-words' is one of the common methods for representing the documents. That method uses a set of words and the number of occurrences of the words in a document. i.e the representation includes information about the terms and their corresponding frequencies in a document. As a result, each document can be represented in a vector space containing a set of weighted terms.

The 'bag-of-words' method is considered a traditional approach as it doesn't take the multiple words and word senses into consideration during the construction of feature selection.

Now, we propose a feature selection approach based on using two axioms namely: the multiple words and the word senses as shown below.

Using multiple words in feature construction or selection means the possibility to replace the single words by word sequences like that concept known as phrase extraction or n-gram features. The previous research works showed that n values may be up to three words (tri-gram) and this is efficient for most of the classification process.

Moreover, amalgamating two or three sequence of words in a document instead of using such individual words will reduce the number of features. For example, the three individual words 'جمهورية', 'مصر', 'العربية' can be concatenated together to form one feature in a phrase form or a multiple-words form. Also, the two words features like 'كرة', 'القدم' can be replaced by the combined feature 'كرة القدم'. Reducing the number of features may improve the classification accuracy.

As mentioned before, the collection of documents (or the dataset) is represented in a matrix with size N×S where N is the number of instances (or documents) while S is the number of features describing the documents. If two or three individual terms are used as features and those individual terms are included in a long multiple word feature, the individual features are eliminated from the feature set. This will reduce the number of features describing the document's dataset.

It is worth mentioning that the feature set may contain some related features. i.e., the features may be related with each other in the form of semantic relation: hyponym, hypernym, synonymous,…etc. This means that considering the relationship between the features is important to construct the significant features. For this reason, the similarity among the features is important. After computing the similarity among the individual features the result will be put in a matrix of size N×S. A higher similarity value means a higher relationship between two features. The highly related features can determine the same category or class; so one of such features can be eliminated while keeping the other one as a feature in the feature set. Figure 18 shows the similarity matrix among the features. The S features describing the dataset can be considered as S vectors. The similarity measure $Sim(f_i,f_j)$ between any two features f_i and f_j can be computed as shown in equation (16).

$$Sim(f_i, f_j) = \frac{\sum_{i,j=1}^{N} f_i f_j}{\sqrt{\sum_{i=1}^{N} f_i^2}\sqrt{\sum_{j=1}^{N} f_j^2}} \quad (16)$$

As a result, the matrix in Figure 18 represents the similarity values among the individual features.

	f_1	f_2	f	$f_3$3..... f_s>>	. . . . .
f_1	1	Sim_{12}	Sim_{13}…..	Sim_{1S}	
f_2		1	Sim_{23} …..	Sim_{2S}	
f_3			1 …….	Sim_{3S}	
.					
f_s				1	

Figure 18: The Similarity Matrix among Features

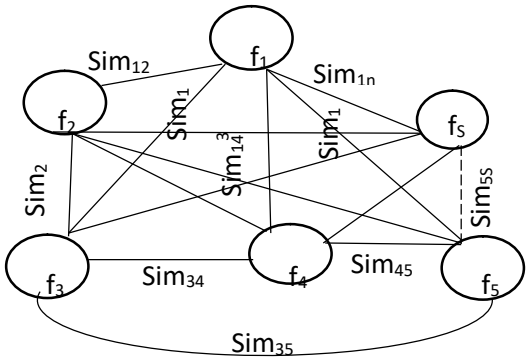

Figure 19: The Graph Similarity

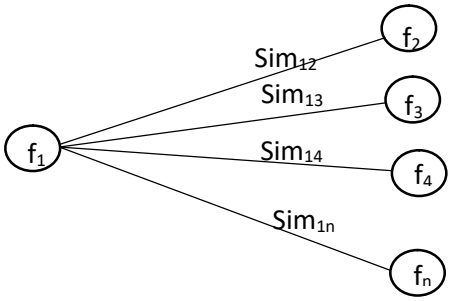

Figure 20: Similarity Values among the Individual Features

The similarity values among the features can be represented in a graph consisting of S nodes connected together by links as shown in Figure 19. The nodes represent the features while a link between any two nodes represents the similarity value between them. The graph similarity is used to construct the most significant features. i.e the original number of features can be reduced and/or fused to improve the classification accuracy. The number of features can be reduced by adopting the fusion concept as in the following steps.

a. Let us assume a starting threshold value T_s. The similarity values between f_1 and all the other features are investigated as shown in Figure 20. If the similarity value, say, Sim_{12}, or Sim_{13}, ….., or Sim_{1S} is less than the threshold value T_s then those corresponding features are taken into consideration. All the connected features with a certain feature say f_1 with similarity values greater than or equal T_s can be eliminated. In this case f_1 is considered a combined feature and it is added to the features set. i.e the number of features becomes the original S features without those eliminated ones.
b. The same process is repeated for the same threshold but between the resulted number of features and f_2. All the features with similarity values ≥ threshold T_s with f_2 are eliminated. i.e the number of features becomes the obtained number of features from the previous step without those features having similarity values ≥ T_s. So, the number of features is reduced.
c. This process is repeated till the last feature f_S.
d. The resulted number of features will be the input to the classifier. Reducing the number of features may enhance the classification accuracy.
e. Other experiments are run and operated for other threshold values. The same previous steps are operated for each experiment. The number of obtained features will be changed and the accuracy values will be also changed. During the implementation of this work, five experiments are operated for different threshold values namely 0.92, 0.93, 0.94, 0.95 and 0.96 respectively. This was done as the maximum similarity value in the feature set was 0.96.

Now, the previous steps can be summarized as shown in the following algorithm.

……..

Algorithm: Combined Semantic Features Fusion

Input: Matrix (N,S) / * N= the no. of documents and S is the no. of features*/

Output: No. of features after combining or fusing the features

Steps:

Compute the similarity values among the individual features
Set a starting threshold T_s to a specific value
Set K=0 /* K is the number of eliminated features*/
Set R=0
While ($T_s < Sim_{max}$)
For i=1 to N
For j=i+1 to (N-R) Do
If Sim (i,j) < T_s
Then K=K+1
Next j
R=K
K=0
Next i
End While
The no. of features after fusion= S-R

……..

As mentioned before, WEKA software tool was used in the implementation work. WEKA is an open source software tool for machine learning purposes written by Java programming language.

Figures 21, 22, and 23 show the values of the measurable criteria for each class name of the BBC dataset for the adopted feature selection methods and the proposed one. The same experiments were done as shown in Figures 24, 25, and 26 for Aljazeera dataset. The performance of the proposed feature selection method based on the semantic fusion approach outperforms the other adopted feature selection methods. This was valid for the two datasets. Figure 27 shows the number of extracted or selected features using the adopted and proposed feature selection methods. The number of selected significant features for the proposed method was less than its corresponding value of either the IG or GI methods and slightly larger than that value of Chi-square method. Figure 28 shows the performance of the SVM classifier using the adopted and proposed feature selection methods. The Semantic Fusion (SF) approach outperforms the other adopted selection methods in terms of recall%, precision%, F-measure% respectively. Moreover, Figures 29, 30, and 31 present respectively the recall%, precision%, and F-measure% for the original features based on term weighting and the proposed feature selection based on semantic fusion. It is shown that the classification performance using the proposed feature selection method is better than that approach based on term weighting. The same concluding remark is also achieved for Aljazeera dataset as shown in Figures 32, 33, and 34.

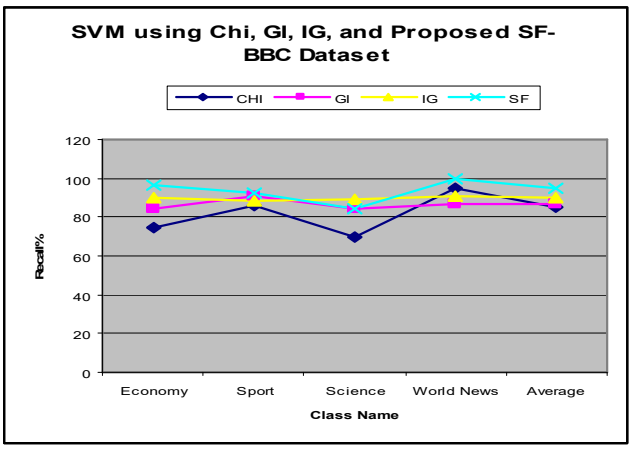

Figure 21: Recall% using BBC Dataset

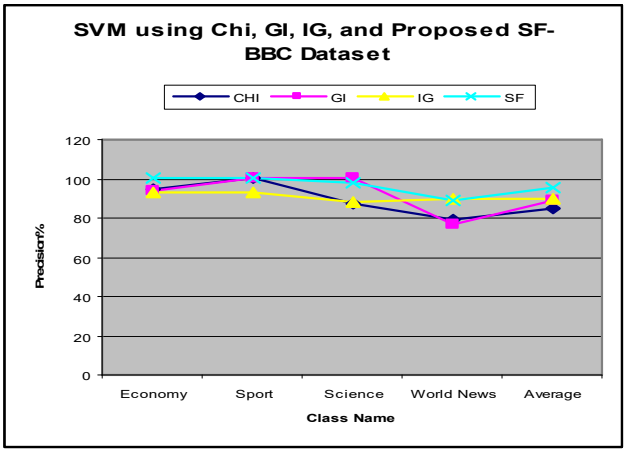

Figure 22: Precision% using BBC Dataset

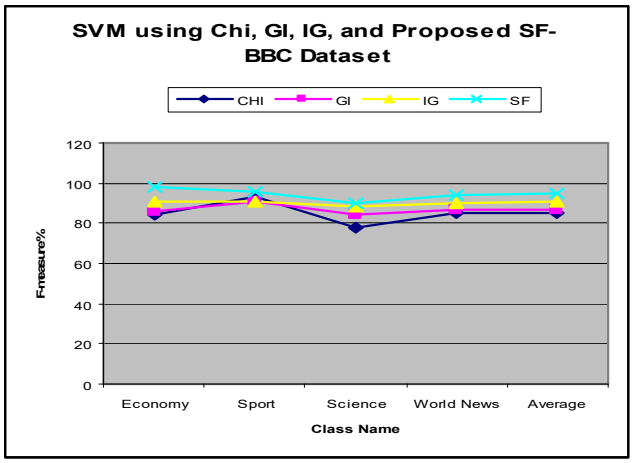

Figure 23: F-measure% using BBC Dataset

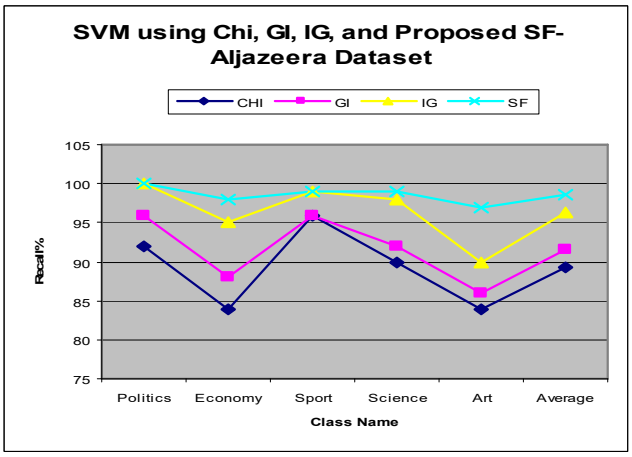

Figure 24: Recall% using Aljazeera Dataset

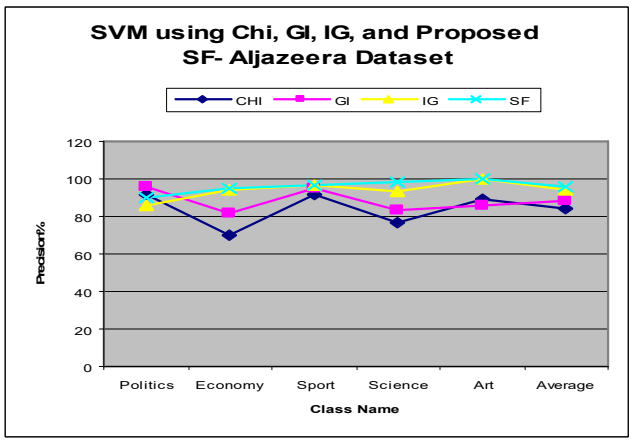

Figure 25: Precision% using Aljazeera Dataset

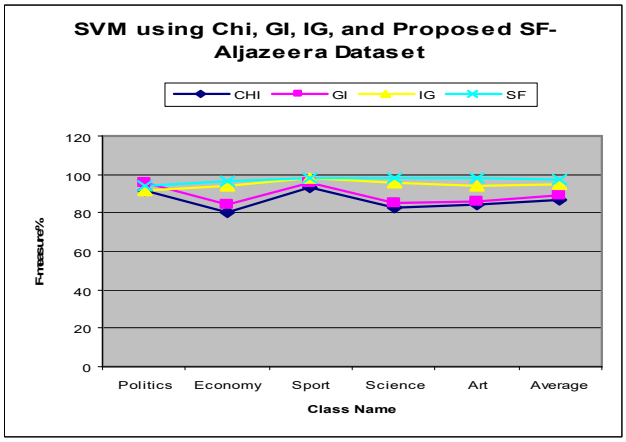

Figure 26: F-measure% using Aljazeera Dataset

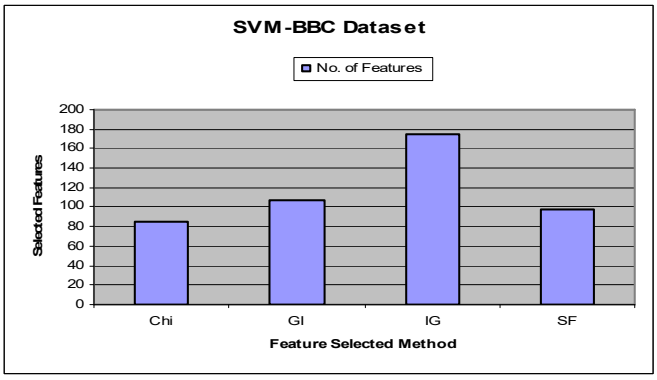

Figure 27: Selected Features for the Selection Methods

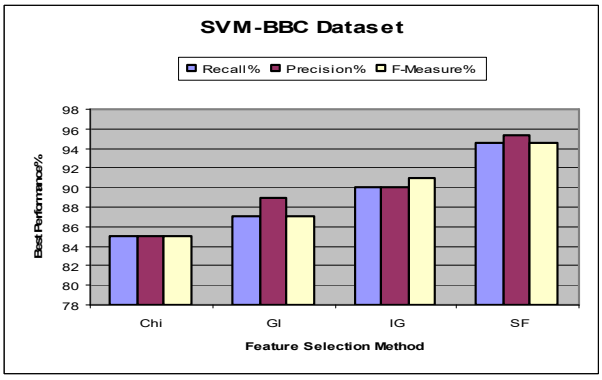

Figure 28: Performance for the Significant Features

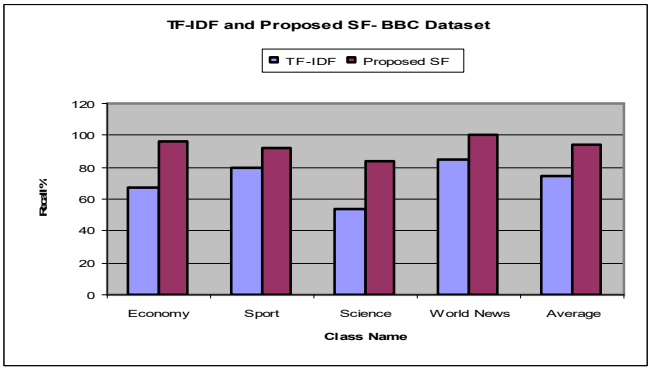

Figure 29: Recall% for TF-IDF and Proposed SF

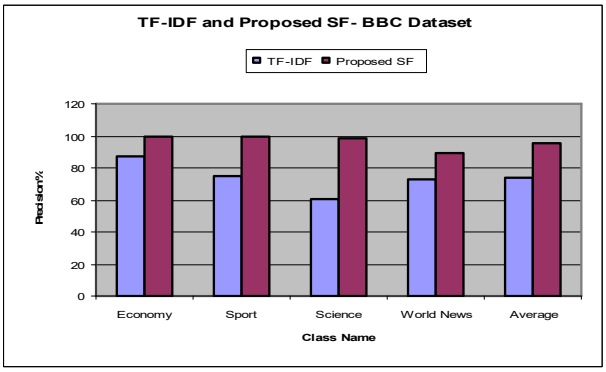

Figure 30: Precision% for TF-IDF and Proposed SF

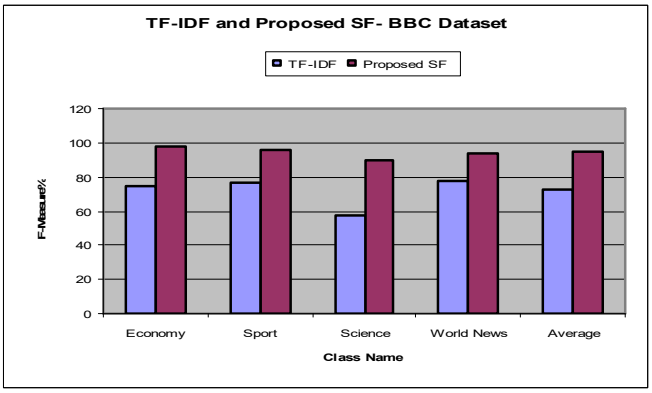

Figure 31: F-Measure% for TF-IDF and Proposed SF

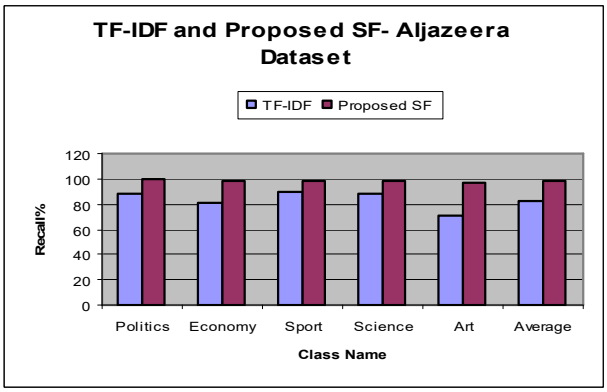

Figure 32: Recall% for TF-IDF and Proposed SF

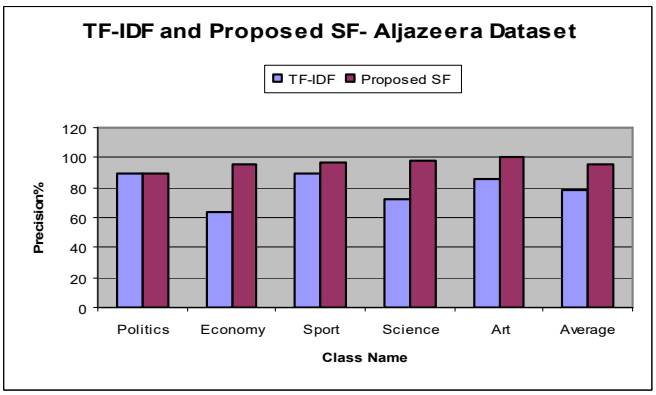

Figure 33: Precision% for TF-IDF and Proposed SF

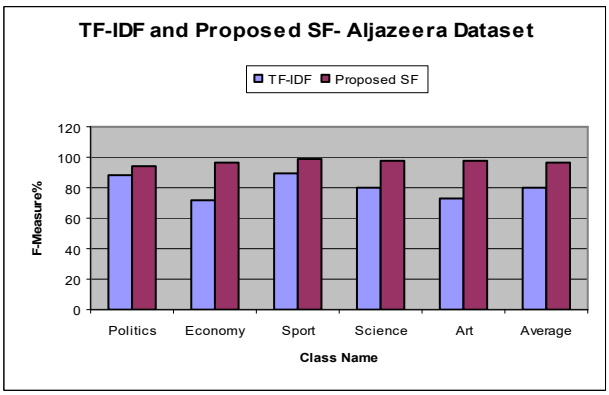

Figure 34: F-Measure% for TF-IDF and Proposed SF

7 DISCUSSION OF RESULTS AND CONCLUSION

This research work was focused on classifying Arabic text. Three types of classifiers were analyzed, operated and tested. The classifiers are decision tree (DT), Naïve Bayes (NB) and support vector machines (SVM). The classifiers were tested using two Arabic datasets mainly BBC Arabic and Aljazeera datasets. Each dataset contains a set of documents. Any document was represented as an instance containing a set of features. i.e., each of the adopted dataset was represented as a matrix containing a set of instances (rows) and a set of features (columns) that are describing those features. A weighted value was calculated for each feature for each document. A threshold value was chosen and then the whole dataset was represented in a way easy to be processed by the classifiers.

From the experimental results, the support vector machines presented a better performance than the other two classifiers. The feature selection plays a vital role in the performance of the classification process. A set of feature selection approaches was analyzed and applied. The approaches are term weighted (or TF-IDF), information gain, Gini index, and chi-square. Feature

selection is important theme to remove the irrelevant features. i.e., the most significant features which describe the dataset were taken while the other irrelevant ones were discarded.

To better choose the most significant features, amalgamation between the feature selection approaches were done. A set of experiments were run using combination of the feature selection methods and the term weighting approach. Due to the amalgamation steps the classification performance was improved due to the changing number of the selected features. i.e., after combining the feature selection methods with the term weighting, the accuracy, recall and F-measure were improved compared to those approaches without combination. This was done for all the experiments where the number of selected features was changed due to the change in the threshold value. Another approach was presented taking into account the multiword features and semantic feature fusion. i.e., a multiword phrase or sequence was used as one feature instead of using the individual words as many features. Also, some features in the feature space were fused. The fusion was done based on the similarity between the individual features. Considering different threshold values of similarity, the number of extracted features was changed. i.e., the fusion was done according to the semantic relationship between the individual features. The feature fusion process reduced the number of selected features. Several experiments were operated and run due to the change of the threshold similarity values. The experiments were operated using the support vector machine as its behavior was better than the other two classifiers.

Finally, a comparative study was done between the performance of the SVM using the adopted approaches of feature selection methods and the proposed one which is based on multiple words and semantic fusion (SF-MW). From the experimental results, the classification accuracy using the combined feature selection was improved by about 14% for the adopted datasets compared to those methods without combination. Using the proposed approach (SF-MW), the classification accuracy% was improved up to 22% compared to some other results published in the literatures as shown in Table 1. The accuracy here is represented in terms of recall, precision, and f-measure. The improvement was achieved for the two adopted datasets. The proposed approach is expected to be also promising for other test-bed datasets.

TABLE 1: COMPARATIVE RESULTS BETWEEN THE PROPOSED APPROACH AND SOME PREVIOUS PUBLISHED WORKS

	BBC Arabic News Dataset			Aljazeera Arabic Dataset		
Previous Literature	Recall%	Precision %	F-Measure%	Recall%	Precision %	F-Measure%
[Ibrahim Abuhaiba and Hassan Dawoud, 2017], [12]	88.2	87.9	88.1	--	--	--
[Ahmed T. Abdulameer et.al, 2017], [14]	72.0	71.0	71.0	--	--	--
[Hamza Mohammed Naji, 2016], [32]	82.9	83.4	85.2	89.7	84.7	86.8
[W.A. Awad, 2012], [13]	--	--	--	86.1	78.1	81.9
[Adel Hamdan, et.al, 2016], [17]	--	--	--	77.4	77.8	77.5
[Mayy M. Al-Tahrawi, 2016], [11]	--	--	--	84.0	85.0	84.3
Combining χ^2 +TF-IDF	**85.0**	**85.0**	**85.0**	**89.2**	**84.0**	**87.0**
Combining GI + TF-IDF	**87.0**	**89.0**	**87.0**	**91.6**	**88.4**	**89.4**
Combining IG + TF-IDF	**90.0**	**90.0**	**90.0**	**96.4**	**94.0**	**94.9**
Proposed SF-MW	**94.5**	**95.4**	**94.5**	**98.6**	**96.0**	**96.9**

χ^2 = Chi-Square, GI = Gini Index, IG = Information Gain, TF-IDF = Term Frequency-Inverse Document Frequency, SF-MW= Semantic Fusion- Multiple Words

REFERENCES

[1] Sumaia Al-Ghuribi and Saleh Alshomrani, "A Comparative Survey on Web Content Extration Algorithms and Techniques", The International Conference on Information Science and Applications (ICISA), PP. 1-6, USA, 2013.

[2] Ghazi Raho, Riyad Al-Shalabi, Ghassan Kanaan, and Asmaa Nassar, "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. 192-195, 2015.

[3] Fawaz S. Al-Anzi and Dia AbuZeina, "Toward an Enhanced Arabic Text Classification Using Cosine Similarity and Latent Semantic Indexing", Journal of King Saud University–Computer and Information Sciences, Vol. 29, No. 2, PP. 189–195, 2016.

[4] Thabit Sabbah, Mosab Ayyash and Mahmood Ashraf, "Support Vector Machine-based Feature Selection Method for Text Classification", The International Arab Conference on Information Technology, Tunisia, PP.1-8, December 22-24, 2017.
[5] Wei Zhang and Feng Gao, "An Improvement to Naive Bayes for Text Classification", Published by Elsevier Ltd, 2011.
[6] Rasha Elhassan and Mahmoud Ahmed, "Arabic Text Classification on Full Word", The International Journal of Computer Science and Software Engineering (IJCSSE),Vol. 4, No. 5, PP. 114-120, May 2015.
[7] Said Bahassine, Abdellah Madani, Mohammed Al-Sarem and Mohamed Kissi, "Feature Selection Using an Improved Chi-square for Arabic Text Classification", Journal of King Saud University Computer and Information Sciences, https://www.sciencedirect.com/science/article/pii/S131915781730544X, 2018.
[8] Tina R. Patil and S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", The International Journal of Computer Science and Applications, Vol. 6, No. 2, pp. 256-261, 2013.
[9] Durgesh K. Srivastava and Lekha Bhambhu,"Data Classification Using Support Vector Machine", Journal of Theoretical and Applied Information Technology, Vol. 12, No. 1, PP. 1-7, 2010.
[10] Mayy M. Al-Tahrawi and Sumaya N. Al-Khatib, "Arabic Text Classification using Polynomial Networks", Journal of King Saud University –Computer and Information Sciences, Vol. 27, No. 4, PP. 437-449, September 2015.
[11] Mayy M. Al-Tahrawi, "Polynomial Neural Networks versus Other Arabic Text Classifiers", Journal of Software, Vol. 11, No. 4, PP. 418-430, April 2016.
[12] Ibrahim Abuhaiba and Hassan Dawoud, "Combining Different Approaches to Improve Arabic Text Documents Classification", Intelligent Systems and Applications, Vol. 4, No. 1, PP. 39-52, 2017.
[13] W.A. Awad, "Machine Learning Algorithm in Web Page Classification", The International Journal of Computer Science and Information Technology, Vol. 4, No. 5, PP. 93-101, 2012.
[14] Ahmed T. Abdulameer, Israa S. Ahmed, Dalia A. Abdulameer, "Arabic Text Classification: An Improved Model using New Relations-Based Features", Journal of College of Education/ Wasit, Vol. 1, No. 27, PP. 455-472, 2017.
[15] Jorge R. Vergara and Pablo A. Este, "A Review of Feature Selection Methods Based on Mutual Information", Neural Computing & Applications, Vol. 24, No. 1, PP. 175–186, 2014.
[16] Aisha Adel, Nazlia Omar, and Adel Al-Shabi, "A Comparative Study of Combined Feature Selection Methods for Arabic Text Classification", The Journal of Computer Science, Vol. 10, No. 11, PP. 2232–2239, 2014.
[17] Adel Hamdan Mohammad, Tariq Alwada'n and Omar Al-Momani, "Arabic Text Categorization Using Support Vector Machine, Naïve Bayes and Neural Network", GSTF Journal on Computing (JOC), Vol. 5, No. 1, pp. 108-115, 2016.
[18] Musab Mustafa Hijazi, Akram M. Zeki and Amelia Ritahani Ismail, "Arabic Text Classification: Review Study", Journal of Engineering and Applied Sciences, Vol. 11, No. 3, PP. 528-536, 2016.
[19] Joseph Lilleberg, Yun Zhu and Yanqing Zhang," Support Vector Machines and Word2vec for Text Classification with Semantic Features", IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing, Beijing, China, 2015.
[20] Girish Chandranshekar and Ferat Sahin, "A Survey on Feature Selection Methods", Computers and Electrical Engineering, Vol. 40, No. 1, PP. 16-28, 2013.
[21] Mahdieh Labani , Parham Moradi , Fardin Ahmadizar and Mahdi Jalili, "A Novel Multivariate Filter Method For Feature Selection in Text Classification Problems", Engineering Applications of Artificial Intelligence, Vol. 70, PP. 25- 37, 2018.
[22] Kaouther Faidi, Raja Ayed, Ibrahim Bounhas and Bilel Elayeb, "Comparing Arabic NLP tools for Hadith Classification", The International Journal on Islamic Applications in Computer Science and Technology, Vol. 3, No. 3, PP. 1-12, 2015.
[23] Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria,"Decision Tree Analysis on J48 Algorithm for Data Mining", The International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, No. 6, pp. 1114-1119, 2013.
[24] Rutvija Pandya and Jayati Pandya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning", The International Journal of Computer Applications, Vol. 117, No. 16, PP. 18-21, May 2015.
[25] Mehdi Allahyari, Seyedamin Pouriyeh and Mehdi Assefi, "A Brief Survey of Text Mining :Classification, Clustering and Extraction Techniques", KDD Bigdas, Halifax, Canada, 2017.
[26] Rami Ayadi, Mohsen Maraoui and Mounir Zrigui, "A Survey of Arabic Text Representation and Classification Methods", Research in Computing Science, Vol. 117, No.1, PP. 51-62, 2016.
[27] Laila Khreisat, "Arabic Text Classification Using N-Gram Frequency Statistics: A Comparative Study", International Conference on Data Mining, Las Vegas, Nevada, USA, June 26-29, 2008.
[28] Bhumika, Sukhjit Singh Sehra and Anand Nayyar, "A Review Paper on Algorithms Used for Text Classification", The International Journal of Application or Innovation in Engineering & Management, Vol. 2, No. 3, PP. 90-99, 2013.
[29] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes and Donald Brown, "Text Classification Algorithms: A Survey", Downloaded From https://arxiv.org/abs/1904.08067, 2019.
[30] Pradnya Kumbhar and Manisha Mali, "A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification", International Journal of Science and Research, Vol. 5, No. 5, PP. 1-9, 2016.

[31] Alaa M. El-Halees, "Arabic Text Classification Using Maximum Entropy", The Islamic University Journal, Vol. 15, No. 1, PP. 157-167, 2007.
[32] Hamza Mohammed Naji, "A New Model in Arabic Text Classification Using BPSO/REP-Tree", M.Sc. Thesis Presented to the Faculty of Engineering, The Islamic University, Gaza, 2016.
[33] Jurgen Kleverwal, "Supervised Text Classification of Medical Triage Reports", M.Sc. Thesis Presented to the Faculty of Electronic Engineering, Mathematics and Computer Science, University of Twente, 2015.
[34] Fouzi Harrag, Eyas El-Qawasmeh and Pit Pichappan, "Improving Arabic Text Categorization Using Decision Trees", The 1st International Conference on Networked Digital Technologies, Ostrava, Czech Republic, 2009.
[35] Charu Aggarwal and ChengXiang Zhai, "Mining Text Data", Chapter 6, Springer Science, 2012.
[36] Vasile Paul Bresfelean, "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment", Conference on Information Technology Interfaces, Croatia, PP. 51-56, 2007.
[37] Sofien Lazreg, "Using Information Extraction and Text Classification in an Effort to Support Systematic Literature Reviews", M.Sc. Thesis Presented to the Department of Computer and Information Science, Norwegian University of Science and Technology, 2012.
[38] Aditya Chainulu Karamcheti, "A Comparative Study on Text Categorization", M.Sc. Thesis Presented to Bachelor of Technology in Information Technology Jawaharlal Nehru Technological University, India, May 2007.
[39] Adam Holmlund, "A Comparison of Machine Learning Algorithms for Classification of Curiosity in Text", Thesis Presented to the Faculty of Science and Technology, Umeå University, 2016.
[40] Saleh Alsaleem, "Automated Arabic Text Categorization Using SVM and NB", International Arab Journal of Information Technology, Vol.2, No.2, PP. 124-128, 2011.
[41] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques", Elsevier, 2012.
[42] Jiliang Tang, Salem Alelyani and Huan Liu,"Feature Selection for Classification: A Review", Data Classification: Algorithms and Applications, https://pdfs.semanticscholar.org/ 310e/a531640728702fce6c743c1dd680a23d2ef4.pdf, 2014.
[43] M. Dash and H. Liu, "Feature Selection for Classification", Intelligent Data Analysis, Vol. 1, PP. 131–156, 1997.
[44] Aytugˇ Onan and Serdar Korukogˇ lu, "A Feature Selection Model Based on Genetic Rank Aggregation for Text Sentiment Classification", Journal of Information Science, Vol. 43, No.1, PP: 25–38, 2017.
[45] Laith Mohammad Abualigah and Ahamad Tajudin Khader, "Unsupervised Text Feature Selection Technique Based on Hybrid Particle Swarm Optimization Algorithm with Genetic Operators for the Text Clustering", Journal of Super Computing, Vol. 73, PP. 4773–4795, 2017.
[46] Heba Abusamra, "A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data", M.Sc. Thesis presented to King Abdullah University of Science and Technology, Kingdom of Saudi Arabia, 2013.
[47] Saeed Parseh and Ahmad Baraani, "Improving Persian Document Classification Using Semantic Relations between Words", Download From https://arxiv.org/abs/1412.8147, 2020.
[48] Haipeng Yao, Chong Liu, Peiying Zhang, and Luyao Wang, "A Feature Selection Method Based on Synonym Merging in Text Classification System", EURASIP Journal on Wireless Communications and Networking, PP. 1-8, 2017, Download From https://www.jwcn-eurasip/journals.springer.open.com/articles/10.1186/s13638-017-0950-z.
[49] Wen Zhang, Taketoshi Yoshida, and Xijin Tang, "Text Classification Based on Multi-word with Support Vector Machines", Knowledge Based Systems, Vol. 21, No. 1, PP. 879–886, 2008.

BIOGRAPGHY

Prof. Dr. Nawal El-Fishawy received the Ph.D. degree in mobile communications, Faculty of Electronic Eng., Menoufia University, Menouf, Egypt, in collaboration with Southampton University in 1991. Now she is the head of Computer Science and Engineering Dept., Faculty of Electronic Eng. Her re-search interest includes computer communication networks with emphasis on protocol design, traffic modeling and performance evaluation of broadband networks and multiple access control protocols for wireless communications systems and networks. Now she directed her research interests to the developments of security over wireless communications networks (mobile communications, WLAN, Bluetooth), VOIP, and encryption algorithms. She has served as a reviewer for many national and international journals and conferences.

Prof. Dr. Mohamed Nour Elsayed is a professor of computer engineering at the Electronics Research Institute, Cairo. He was graduated from the Computer Department at the Faculty of Engineering, Ain Shams University in 1980. He obtained his M.Sc. and Ph.D. in 1987 and 1993 respectively. He taught more than twenty-years ago at the American University in Cairo (AUC) as a part-time instructor. He taught also five years ago at Princess Nourah University, Riyadh, KSA. He was the head of the Informatics Department as well as the Vice-President of the Electronics Research Institute, Cairo. He is an IEEE member and a reviewer of some national and international computer journals. The areas of his interest include; but not limited to; high performance computing, artificial intelligence, computational linguistics, and others.

Dr. Maha Saad Tolba is a lecturer of Computer Engineering at the Faculty of Electronic Engineering, Menofia University. She was graduated from the Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University in 1997. She obtained her M.Sc. and Ph.D. in 2006 and 2011 respectively. The areas of her interest include; but not limited to; computer networks, information security, and information technology.

Eng. Ayat Elnahaas is a research assistant at the Department of Research Informatics, Electronics Research Institute, Cairo. She was graduated from the Faculty of Electronic Engineering, Menofia University in 2013. Currently; she is working in her M.Sc. in the area of Arabic text processing. The research areas of her interest are: computational linguistics, and information technology.

TRASLATED ABSTRACT

طرق تعلم الآلة واستخلاص الصفات لتصنيف النصوص العربية: تحليل ودراسة مقارنة وإقتراح

آيات النحاس*[1]، محمد نور*[2]، نوال الفيشاوى**[3]، مها طلبة**[4]
*معهد بحوث الإلكترونيات- القاهرة- جمهورية مصر العربية**
*كلية الهندسة الإلكترونية- منوف- جامعة المنوفية- جمهورية مصر العربية***
[1]eng_ayatelnahas@yahoo.com
[2]mnour99@hotmail.com
[3]nelfishawy@hotmail.com
[4]maha_saad_tolba@yahoo.com

الملخص:
يقدم هذا العمل البحثى تحليلا وتدقيقا لبعض أنواع تصنيف النصوص والتى تم اختبار آدائها على مجموعتين من النصوص العربية كعينات إختبارية. تم إجراء دراسة مقارنة بين آداء أنواع التصنيفات المختلفة التى تم تبنيها، كما تم دراسة وتحليل وتقييم عدد من طرق استخلاص الصفات فى النصوص العربية أيضا لكون تلك الصفات تأثير كبير على دقة عملية التصنيف. فاستخدام عدد هائل من الصفات قد يؤدى بدورة إلى تقليل من كفاءة ودقة عملية التصنيف، الأمر الذى يتطلب استخلاص الصفات التى لها أهمية كبيرة فى تحديد التصنيف السليم. وعليه فقد تم مقارنة الآداء لطرق استخلاص الصفات التى تبنيها هذه الدراسة.
هذا وقد تم إجراء بعض التعديلات على طرق استخلاص الصفات من خلال تعظيم الاستفادة من مزايا كل طريقة والقيام بعملية صهر لتلك الطرق المختلفة. كما تم إقتراح طريقة مستحدثة لاستخلاص الصفات اعتمادا على وجود الكلمات المتعددة التى يمكن معالجتها ككلمة أو كصفة واحدة، وكذا إنتقاء الصفات من خلال عملية الصهر الدلالى لبعضها Semantic Fusion. وفى هذا الصدد تم مقارنة آداء طرق استخلاص الصفات مع آداء الطريقة المقترحة.

ومن خلال التجارب العملية التى أجريت، فقد أظهر أسلوب تصنيف المتجة الآلى المدعم SVM لتصنيف النصوص آداء أفضل من كل من مصنفى KNN and NB. كما أظهرت عملية الصهر لطرق استخلاص الصفات دقة أفضل لتصنيف النصوص عن تلك الطرق التى تم تبنيها منفردة. إضافة لما تقدم فقد أظهر المقترح الجديد لاستخلاص النصوص المعتمدة على الكلمات المتعددة والصهر الدلالى للكلمات Multiple-words and Semantic Fusion آداء أفضل وأكثر دقة من تلك الطرق المختلفة محل الدراسة. وتشير النتائج العملية إلى تفاوت درجة التحسين فى الطريقة المقترحة عن نظيراتها محل الدراسة بحوالى22% عند تطبيق ذلك على عينتين من النصوص العربية كعينات اختبارية والتى احتوت إحداها على عدد 1246 مستندا بينما احتوت الأخرى على 1500 مستندا. هذا ومن المتوقع أن تعطى الطريقة المقترحة أيضا نتائج واعدة عند تجربتها على نصوص لأخرى سواء عربية أو إنجليزية.

الكلمات الدالة:
خوارزميات التصنيف، استخلاص الصفات، الكلمات العربية المتعددة، الانصهار الدلالى، مجموعات النصوص الاختبارية، معايير تقييم الآداء.