Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Feature selection using an improved Chi-square for Arabic text classification

Said Bahassine [a], Abdellah Madani [b], Mohammed Al-Sarem [c], Mohamed Kissi [d],*

[a] Laboratory LIMA, Department of Computer Science, Faculty of Sciences, Chouaib Doukkali University, B.P. 20, 24000, El Jadida, Morocco
[b] Laboratory LAROSERI, Department of Computer Science Faculty of Sciences, Chouaib Doukkali University, B.P. 20, 24000, El Jadida, Morocco
[c] Information System Departement, Taibah University, Al-Madinah Al-Monawarah, Saudi Arabia
[d] Laboratory LIM, Department of Computer Science, Hassan II University Casablanca, Faculty of Sciences and Technologies of Mohammedia, B.P. 146, Mohammedia, Morocco

ARTICLE INFO

Article history:
Received 25 December 2017
Revised 17 May 2018
Accepted 21 May 2018
Available online 24 May 2018

Keywords:
Feature selection
Chi-square
Arabic text classification
Light stemming
Mutual information
Information gain
SVM
Decision tree

ABSTRACT

In text mining, feature selection (FS) is a common method for reducing the huge number of the space features and improving the accuracy of classification. In this paper, we propose an improved method for Arabic text classification that employs the Chi-square feature selection (referred to, hereafter, as ImpCHI) to enhance the classification performance. Besides, we have also compared this improved chi-square with three traditional features selection metrics namely mutual information, information gain and Chi-square.

Building on our previous work, we extend the current work to assess the method in terms of other evaluation methods using SVM classifier. For this purpose, a dataset of 5070 Arabic documents are classified into six independently classes. In terms of performance, the experimental findings show that combining ImpCHI method and SVM classifier outperforms other combinations in terms of precision, recall and f-measures. This combination significantly improves the performance of Arabic text classification model. The best f-measures obtained for this model is 90.50%, when the number of features is 900.

1. Introduction

With the fast growth of online documents, dealing with textual documents has become an important technique. On the one hand, this technique can help to find interesting relevant information. On the other hand, it is important because it facilitates understanding and organizing such documents. Labelling unseen text document to one or more predefined appropriate categories based on the document content and using machine learning is known as text classification problem. The text classification can be used in a broad range of applications in diverse areas, including digital library systems, detecting spam in opinion reviews (Hammad and El-Halees, 2015), classification of email messages (Nikhath et al., 2016), evaluation of sentiment analysis (Mostafa, 2017), analysis of movie reviews (Singh et al., 2017), text summarization (Jo, 2017), sentiment analysis for marketing and Arabic opinion mining (Cherif et al., 2016).

Although, several research studies in text classification have been conducted for natural languages, such as English (Barigou, 2016), Chinese (Ye-wang et al., 2016); (Junkai et al., 2016) , Latin and Turkish texts (Kilimci et al., 2016), the number of related works on Arabic script is still limited due to the complex nature of Arabic inflectional and derivational rules as well as its intricate grammatical rules and its rich morphology (Alghamdi and Selamat, 2019).

In the context of implementation, the text classification system can be divided into three main steps:

- Pre-processing step: where the punctuation marks, stop words, and diacritics and non-meaningful words are removed.
- Features selection step: in this step the relevant features are selected from the original text. They present the text that is input into the learning step.
- Learning step: numerous techniques have been deployed to teach systems how to divide text documents into different categories.

* Corresponding author.
E-mail address: kissim@gmail.com (M. Kissi).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2018.05.010

A major difficulty of text classification is the scope of the original data. To overcome this obstacle, a features selection method is used to remove redundant and irrelevant attributes and select the most distinct features.

In this research, we present an improved method of Chi-square feature selection method to minimize the data and produce higher classification accuracy. Then, we compare the effect of eight selection procedures using four features selection and two classifiers on Arabic text categorization.

The remainder of this paper is organized as follows: the second section introduces a review of previous work in Arabic text classification. Then, data collection and text preprocessing are presented in the third section. In section four, three traditional features selection are presented. The suggested features selection method is described in section five. In section six the experimental results are presented. Finally, we will draw some conclusions and provide suggestions for future research.

2. Previous research

In the field of Arabic text classification research, numerous methods have been used to select the relevant attributes and the optimal number of features from the original document (Mesleh, 2011). This can be done through statistical techniques that compute a score to each attribute, and then utilize the top-scoring ones in constructing the classifier.

Baraa et al. presented a novel text classification approach in Arabic, namely frequency ratio accumulation method (FRAM) (Baraa et al., 2014). It deals with features selection and classification in one process. The results revealed that the FRAM gave better results than three classifiers: Naïve Bayesian, Multi-variant Bernoulli Naïve Bayes (MBNB) and Multinomial Naïve Bayes models (MNB). It achieved 95.1% on the macro-f-measure value by using unigram word-level representation method. But the researchers did not compare the results in terms of recall, precision and f-measures.

Suhad et al. compared the accuracy of the existing Arabic stemming methods and deployed the Arabic WorldNet ontology and used it in the conceptual representation approach as a lexical and semantic technique (Suhad et al., 2015). They used BBC dataset for experimentation. The authors concluded that the position tagger with the root extractor provide the most optimal results compared to other stemming methods and that the combination of “Has Hyponym” relation with position tagger outperformed other semantic relations and led to an increase of 12.63% compared to other combinations. But the authors did not compare the results in terms of recall and precision.

Harish et al. have conducted a comparative study of eight widely used feature selection methods namely: Term Frequency-Inverse Document Frequency (TF·IDF), Information Gain (IG), Mutual Information (MI), Chi-square ($\chi 2$), Ambiguity Measure (AM), Term Strength (TS), Term Frequency-Relevance Frequency (TF·RF) and Symbolic Feature Selection (SFS) (Harish and Revanasiddappa, 2017). They have used five different classifiers Naive Bayes, K-Nearest Neighbor, Centroid Based Classifier, Support Vector Machine and Symbolic Classifier. Experimentations were carried out on standard benchmark datasets like Reuters-21578, 20-Newsgroups and 4 University dataset. The results showed that Symbolic Feature Selection (SFS) method outperformed all the other features selection. The evaluation indicated that the SFS method had remarkable influence on improving the categorization accuracy. It achieved 94% in terms of f-measures.

Rasha et al. conducted a comparative study of the impact of four classifiers on Arabic text categorization accuracy using stemming technique (Rasha and Mahmoud, 2016): the Sequential Minimal Optimization (SMO), Naïve Bayesian (NB), Decision Tree J48 and K-Nearest Neighbors (KNN). Two approaches were used Khoja and light stemmers. The results were compared with the result of classification without using stemming. The authors collected a corpus from local and international newspaper websites. The dataset consists of 750 documents that are separated into five classes; namely economics, politics, religion, sports, and technology. All documents were preprocessed by removed punctuation marks, digits, the formatting tags and non-Arabic words. The authors implemented Khoja and light stemmers using Weka data mining tools. The results were shown in terms of precision, recall and f-measure. Findings show that light stemmer gave a better accuracy than Khoja stemmer and SMO classifier outperformed the other classifiers in the training stage, while NB classifier outperformed the other classifiers in the test stage. This is due to the fact that SMO needs bigger data to perform better. It achieved 94% in terms of f-measures when light stemming and NB are used. But the authors didn’t use feature selection to minimize the dimensionality of the data.

Attia et al. proposed a new framework for Arabic word root extraction and text classification (Attia et al., 2016). It is based on the use of Arabic patterns and extracts the root without relying on any dictionary. To investigate the performance, a corpus containing 1526 documents from six categories collected from Saudi Press Agency (SAP) was used. At the preprocessing step, stop words, non-Arabic letters, symbols and digits were removed. Then, the LibSVM with three N-gram Kernel ($N = 2,3,4$) was applied. Although, the results show that the root extraction improves the quality of classifiers in terms of recall, accuracy and f-measures, the precision slightly decreased. Accuracy and f-measure report 90.79% and 62.93% respectively.

Mahmoud et al. have proposed a new method to enhance the accuracy of the Arabic text categorization (Mahmoud et al., 2016). They proposed to mix a bag of words method with two adjacent words collected with different proportions. The Term Frequency was used as features selection and the texts were classified using frequency ratio Accumulation. Normalization and stemming were utilized in the investigation. To evaluate the method, three datasets of different categories were collected from online Arabic websites. The results indicated that the text classification using normalization outperformed text classification when normalization and stemming were not used in terms of accuracy. The results showed that applying text classification with normalization achieved the highest classification accuracy of 98.61% using dataset with 1200 documents belonging to four classes.

Harrag et al. tried to explore the impact of using decision tree method on classifying Arabic text documents (Harrag et al., 2009). They used two different corpora. The first one contains 373 documents belonging to eight classes; it was collected from the Arabian scientific encyclopedia “Do you know” (هل تعلم). The second one contained 435 documents, belonging to fourteen classes; it was collected from Hadith encyclopedia ((موسوعة الحديث الشريف. The authors have used two-third of dataset for training the text classifier and one third for testing the classifier. The performance of the improved classifier reached about 93% generalization accuracy for the scientific dataset and 91% for the literary dataset. The results showed also that the nature and the specificity of the corpus documents influenced the classifier performances.

Bahassine et al. have presented two contributions (Bahassine et al., 2014). In the first one, they developed a new hard stemming algorithm that reduces all forms of the attributes to their root. In the second contribution, they compared the impact of Khoja stemmer and the proposed stemmer on Arabic text classification. To evaluate the performance of this proposed stemmer, a dataset collected from cnnarabic.com was used. It contained 5070 documents that varied in length and fall into six categories: sport, entertainment, business, Middle East, SciTech (Science and Technology) and world. The recall was used to compare the performance of the obtained models. Two algorithms were used; Khoja stemmer and proposed stemmer after cleaning up the data. To reduce the dimensionality, the authors used Chi-square as features selection. The results showed that text classification using the new stemmer outperformed classification using Khoja stemmer. The best recall obtained for this model was 79.74%, when the number of features is 500 while the best recall obtained by Khoja algorithm was 78.44%.

Afterwards, Bahassine et al. extended the first work using the same corpus (Bahassine et al., 2017). The recall, precision and f-measures were used to compare the performance of the obtained models. The results showed that text classification using the new stemmer outperformed classification using Khoja stemmer. The empirical results showed that the recall measure, precision measure and f-measure decreased whether the number of features was high or low. The sport category achieved the highest precision, recall, and f-measure values compared with other categories because the attributes in this class were distinctive compared to other classes. The entertainment category had a noticeably poor precision, recall, and f-measure. The f-measure was 92.9% in sport category and 89.1% in business category.

In another study, Bahassine et al. elaborated a new selection method (Bahassine et al. , 2016). The researchers conducted a comparative study between light stemming and hard stemming, on the one hand, and Chi-square and the proposed features selection method, on the other hand. Then, they analyzed the impact of stemming and features selection on Arabic text classification in terms of recall measures using decision tree. The authors collected a corpus that contained 250 documents from “Hespress” and “Hesport” online media. The collected documents were categorized into five categories: culture and art, economics, politics, society, and sport. In this dataset, each text was saved in a separate file. To clean up the data from non meaningful and noisy words, all documents were preprocessed by removing digits, punctuations, numbers, all non-Arabic characters and stop words as well. The results showed that combining proposed features selection method and light stemming technique greatly increased the performance of Arabic text classification in terms of recall compared to Chi-square features selection and hard stemming. It remains to compare proposed method using greater corpus in terms of precision and f-measure.

Reviewing the previous researches, we noted that they used only stemming or feature selection to shrink the number of attributes so as to optimize the performance of designed classifiers. Nevertheless, these steps are not enough to generate better classification in terms of accuracy. In this paper, both techniques will be used and the results will be compared in terms of recall, precision and f-measures. In addition to aforementioned methods, techniques such as: Naïve Bayesian (NB) (Jadon and Sharma, 2017), Decision Tree (Bahassine et al., 2014; Bahassine et al., 2017; Bahassine et al., 2016; Kissi and Ramdani, 2011) and K-Nearest Neighbors (KNN) (Jo, 2017), Neural Network (NN) (Al-Anzi and AbuZeina, 2017) and Support Vector Machine (SVM) (Al-Anzi and AbuZeina, 2017) have been used for text classification tasks. Thus, in this work, DT and SVM will be used as classifiers to compare the performance of our proposed feature selection and three others in terms of recall, precision and f-measure.

3. Data collection

A corpus or data collection can be defined as a set of text documents that can be classified under many subsets. To assess the performance quality of the features selection, one of the corpus of open source Arabic corpora (OSAC) have been used (Saad and Ashour, 2010). The corpus contains 5070 documents of different lengths. These documents consist of six classes: sports, entertainment, business, the Middle East, SciTech and world. In this data collection, each document was saved in a separate file.

The corpus is divided in two sets: The training set consists of 4057 texts. Whilst, the remain files are assigned as a test set. The distribution of the classes in the dataset is represented in Table 1.

3.1. Text pre-processing

Text pre-processing is an important step in the text classification process. This step can reduce the errors and enhance the accuracy of classification (Uysal and Gunal, 2014; Ayedh, 2016). The main objective of this endeavor is to get rid of noisy and non-meaningful words (Ayedh, 2016) in the data (Elhassan and Ahmed, 2015). Each file of the corpus was subject to the following procedure:

- delete digits, punctuation marks and numbers.
- delete all non-Arabic characters
- delete stop-words and non-useful words like: pronouns, articles and propositions.
- change the letter “”ى to “”ي.
- change the letter “”ة to “”ه.
- change the letter “”آ“ ,”إ“ ,”ؤ“ ,”ئ“ ,”أ to “”ا.

3.2. Stemming

Stemming is a process of reducing inflected words into one form (stem or root) by removing prefixes, suffixes and infixes. There are several types of stemming algorithms: statistical (Al-Shalabi and Evens, 1998), dictionary, transducer (Nehar et al., 2012), morphological (Boudchiche et al., 2017) including hard stemming (Khoja and Garside, 1999) and light one (Cherif et al., 2014). These last two algorithms were found to be the most used types. Hard stemming is based on three-letter roots for Arabic words. Light stemming deletes the common prefixes and suffixes.

It is worth noting that Khoja's algorithm is a well-known hard stemming algorithm (Khoja and Garside, 1999). It removes the largest suffix and prefix from the word, while extracting the root by comparing the rest of the word with its verbal and noun lists.

Sawalha conducted a comparative study of three stemming algorithms: Khoja's stemmer, Buckwalter's morphological analyzer and Al-Shalabi algorithm (Sawalha et al., 2008). The results showed that Khoja's stemmer performs better in terms of accuracy.

Cherif et al. have pointed out that Khoja's algorithm undergoes several flaws (Cherif et al., 2014; Sawalha et al., 2008; Cherif et al., 2015) but they have not compared the effect of stemming in Arabic text classification. However, new algorithm of stemming is proposed and compared with Khoja's one (Bahassine et al., 2017; Bahassine et al., 2016). This algorithm will be used in this paper

Table 1
Datasets description.

Categories	Number of text	Number of training set	Number of test set
Business	836	669	167
Entertainment	474	379	95
Middle East	1462	1170	292
SciTech	526	421	105
Sport	762	610	152
World	1010	808	202
All	5070	4057	1013

to extract the stem of word and the documents will be presented by vector of terms (word stem).

4. Feature selection

Features selection is effective in the reduction of large data in text classification. It can enhance the classification process. Feature selection deletes irrelevant and noisy data and chooses a representative sub-set of all data to minimize the complexity of the classification process.

Numerous techniques of features selection can be detected in the literature such as: Mutual Information (MI) (Yang and Pedersen, 1997), Chi-square (Bahassine et al., 2014), Information Gain (IG) and Term Frequency-Inverse Document Frequency (TF-IDF) (Dadgar, et al., 2016). The present research tried to introduce a modified version of Chi-square feature selection method which will be presented hereafter.

4.1. Information gain

Information gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term t_k in a text document. The Ig for a term t_k is defined as (Mesleh, 2011; Sebastiani, 2016):

$$IG(t_k, c_i) = \sum_{c\in\{c_i,\bar{c}_i\}} \sum_{t\in\{t_k,\bar{t}_k\}} P(t,c) * \log\left(\frac{P(t,c)}{P(t) * p(c)}\right) \qquad (1)$$

4.2. Mutual information

Mutual information is a measure of dependence between variable (a term t_k and a category c_i), if the MI for a term the t_k is zero then a term t_k and a category c_i are independent. The MI is defined as (Mesleh, 2011; Sebastiani, 2016):

$$MI(t_k, c_i) = \frac{\log(P(t_k, c_i))}{P(t_k) * P(c_i)} \qquad (2)$$

4.3. Chi-square

The Chi-square statistics formula is related to information-theoretic feature selection functions which try to capture the intuition that the best terms t_k for the class c_i are the ones distributed most differently in the sets of positive and negative examples of class c_i.

$$\mathrm{Chi-square}(t_k, c_i) = \frac{\mathrm{N}(\mathrm{AD} - \mathrm{CB})^2}{(\mathrm{A}+\mathrm{C})(\mathrm{B}+\mathrm{D})(\mathrm{A}+\mathrm{B})(\mathrm{C}+\mathrm{D})} \qquad (3)$$

Eq. (3) illustrates Chi-square statistics, where: N = Total number of documents in the corpus, A = Number of documents in class c_i that contain the term t_k; B = Number of documents that contain the term t_k in other classes; C = Number of documents in class c_i that do not contain the term t_k; D = Number of documents that do not contain the term t_k in other classes.

Each feature is assigned a score in each class as described in (3). Then, all these scores are combined with a single final score max (Chi-square(t_k, c_i)). (see the example in Table 2).

The Chi-square value of the terms: برميل "barrel", تحالف "alliance", تعادل "equality", سوريا "Syria" and باحث "Researcher" are displayed in Table 1 with the six classes using (3). For example, the Chi-square value of the term تعادل "equality" in class sport is 709.373 while it is much lower in all other categories. Thus, the final score for this term is 709.373 and its final class is "sport".

Hence, the final score is used to classify all attributes in a descending order and the highest top (p) score features are selected.

5. Improved Chi-square feature selection (ImpCHI)

The Chi-square method has shown very good results but still suffers from some limitations. For instance, when the top 20 attributes are selected using Chi-square, the number of these attributes per class tends to vary accordingly (see Table 3)

The distribution of attributes per class in Table 3 is not a proportion of number of document as 9 out of 20 attributes belong to the sport category, 11 out of 20 attributes belong to the business category. Therefore, the classification of these attributes will be certainly impacted.

Table 4 shows the rate of classification per class for the top 20 attributes using Chi-square as feature selection. Tables 3 and 4 indicate that there is a correlation between the number of attributes and the f-measure. This means that the number of attributes has an impact on classification accuracy. As the results show, the number of attributes under 'sport category' is 9 with a 93.7% f-measure. The number of attributes for 'Business' class was 11 which translates into 87.6% f-measure. No attributes were found under the other classes, but 'Middle East' scored a 60.2% f-measure, 'SciTech' gave a 30.1% f-measure and other classes scored a 0%.

Despite the absence of attributes under this classes, the f-measure values of the classes 'Middle East' and 'SciTech' can be explained through the Chi-square value of the attributes that belong to other classes and that can be also commonly used in the economic jargon. This can be illustrated for example by the word 'تعادل' 'equality' which can both refer to 'sport' and 'Middle East'.

Table 3
The number of top 20 attributes per class using Chi-square.

Class	Number of attributes using Chi-square
Business	11
Entertainment	0
Middle_East	0
Scitech	0
Sport	9
World	0

Table 2
Chi-square of some terms in corpus.

Class	Term				
	برميلBarrel	تحالفAlliance	تعادلEquality	سورياSyria	باحثResearcher
Business	**254.855**	4.578	117.850	20.486	23.273
Entertainment	6.598	15.457	33.430	0.083	2.177
Middle East	13.683	8.725	125.399	**187.826**	3.209
SciTech	3.608	16.059	35.529	14.202	**555.922**
Sport	11.286	36.742	**709.373**	8.204	20.332
World	15.843	**141.742**	64.008	23.861	26.331

The input of our algorithm is the term document Matrix M ($n*w$) dimension. n refers to the number of documents and w indicates the number of terms (attributes). An entry 'm_{ij}' refers to the corresponding TF-IDF of the i^{th} document and the j^{th} term.

$N(c_i)$: the number of documents belonging to the class c_i .

Output: p relevant terms (attributes) will be selected using ImpCHI).

To make up for the chi-square problem which caused the absence of attributes under some classes, like in 'entertainment', 'business', 'Middle East', 'SciTech' and 'world' classes, we suggest here then the use of ImpCHI algorithm to balance the selection of the number of attributes per class.

The pseudo code of the algorithm is given below:

Algorithm

```
D={$d_1$,$d_2$,$d_3$,. . ... . .....,$d_n$} set of n documents
T={$t_1$,$t_2$,$t_3$,. . ... . .....,$t_w$} set of w terms (attributes)
C={$c_1$,$c_2$,$c_3$,. . ... . .....,$c_k$} set of k classes (k<=n)
M={$m_{ij}$=tf-idf($d_i$,$t_j$)} the data matrix
p: number of pertinent terms to select for all classes
N($c_i$): the number of documents belonging to class $c_i$
$ch_i$: Chi-squre($t_i$,$c_i$) (see equation 3)
L: list of triplets ($t_i$,$ch_i$,$c_i$)
function NumSelbyClass(p,$c_i$)
    // the number of relevant terms belonging to class $c_i$ that
    will be selected by ImpCHI
return round(p*N($c_i$)/n) // round(x) rounds the elements of x
    to the nearest integers.
function TermsSelbyClass(L, $c_i$, p)
    // the list of relevant terms belonging to class $c_i$ that will be
    selected by ImpCHI.
    i = 0
    count = 0
    LT=[][]
    limite = NumSelbyClass(p,$c_i$)
    while (count < limite)
        if L[i][2]=$c_i$ then
            LT = LT+L[i][0]
            count++
        endif
    i++
    endwhile
return LT
begin
    L=[][]
    Tr=[] // the list of relevant terms belonging to class $c_i$ that
    will be selected by ImpCHI
    for each $t_i$ in T
        $ch_i$,$c_i$= maxChiClass($t_i$,C) // $ch_i$ the maximal chi-square of
    the term $t_i$ for all classes, and // $c_i$ the correspondent class of
    $ch_i$ that gives the maximum value of chi-square
        L = L+[($t_i$,$ch_i$,$c_i$)]
        L = SortbyChi(L) // sorted L by chi-square value $ch_i$ (2nd
    element in triplet)
for $c_i$ in C
        Tr = Tr+ TermsSelbyClass(L,$c_i$,p)
end
```

For example, to select the top twenty first terms (p = 20) using ImpCHI algorithm. First, Chi-square value is computed for each term and for all classes. Then, **maxChiClass(t_i,C)** is kept and becomes an term (attribute) of class c_j (see Table 2). Attributes belonging to the same class are sorted by Chi-square value. The top (**NumSelbyClass(p,c_i)**) attributes of every class is proportion by number of the documents belonging to the same class.

Table 4
Resuls of classification per class for top 20 attributes using ch-squqre and DT.

Class	Precision	Recall	F-Measure
Business	93.2	82.6	87.6
Entertainment	00	00	00
Middle East	44	95.2	60.2
SciTech	38.2	24.8	30.1
Sport	90.2	97.4	93.7
World	00	00	00
Average	45.6	58.3	49

Table 5
The distribution of top P relevant attributes By classes using impchi algorithm.

P	20	100	600
Business	3	16	99
Entertainment	2	9	56
Middle East	6	29	173
SciTech	2	10	62
Sport	3	15	90
World	4	20	120
Sum	20	99	600

Finally, the top p attributes are selected.

As Table 5 shows, the distribution of top p relevant attributes using ImpCHI. The distribution of attributes per class in Table 5 is proportion of number of document. 6 out of 20 attributes belong to the 'Middle East' category, 4 out of 20 attributes belong to the word category, and 2 out of 20 belong to the entertainment and 'SciTech' classes.

6. Results and analysis

The value of selected features was varied from 20 to 1400 in order to carry out an optimal comparison of the impact of the previously mentioned features selection Chi-square, MI, IG and ImpCHI using DT and SVM classifiers on Arabic documents classification. The findings were compared using the commonly used evaluation metrics precision, recall and f-measure as outlined below:

The recall measure is the ratio of the relevant data among the retrieved data. It is defined as follows:

$$r = \frac{tp}{tp + fn} \tag{5}$$

where: tp : true positive; fn: false negative

The precision measure is the ratio of the accurate data among the retrieved data. Its formula is given as follows:

$$p = \frac{tp}{tp + fp} \tag{6}$$

where: fp: false positive; The F-measure of the system is defined as the weighed harmonic means of its precision and recall. It is defined as follows:

$$F = \frac{2rp}{r + p} \tag{7}$$

where: r is recall which is given in (5); p is precision which is given in (6)

The variation of the number of selected attributes has facilitated the analysis of the performance of the four features selection methods and the two classifiers in the comparison phase. The results indicate that the precision, recall and f-measure decrease when the number of features is lower than 60, which can be interpreted by the fact that selected features are insufficient.

Tables 6, 7 and 8 show the findings for precision, recall and f-measures values of four feature selection methods MI, IG, Chi-square and ImpCHI when SVM and DT classifiers are used.

Experimental results in this paper reveal that ImpCHI performed better than other features selection for most features.

Table 6 shows the precision values of different feature selection methods when SVM and DT are used. It can be seen that, when SVM is used, the precision values of the proposed method are generally higher than those of the other methods, except when the numbers of feature exceed 900, the average value of the precision was 85.29%. When DT is used, IG shows the best performance, the highest value of precision is 79%, but the average value of precision for different size of features of the proposed methods is 75.43% higher than IG only 74.37%.

Table 7 shows the recall values of four feature selection methods using SVM and DT classifiers. When SVM is used, MI shows the worst performance. The recall values of the proposed method are generally higher than other methods; the highest value of recall is 90.50% when the numbers of feature are 900 features. When DT is used, the highest value of recall is 79.50% when the number of feature is 300 features.

Table 8 shows the f-measure values of MI, IG, Chi-square and proposed method using SVM and DT classifiers. When SVM is used, the proposed method obtained better results when the number of features between 40 and 900, the best value of f-measure was 90.50% and it was obtained for 900 features, the average value of f-measure is 84.93%. When DT is used, the average value of f-measure using the proposed method was 74.54%.

Another interesting result is that the highest rate of precision, recall and f-measures are attained when ImpCHI is deployed with SVM classifier.

SVM classifier gave the better results in terms of precision, recall and f-measure compared to DT for all features selection at different size of features except when the number of features was 20. However, decision tree provides an easy interpretable result by non export, it can help us to identify important and pertinent terms for every classes, while SVM is a black box that is difficult to interpret the results.

IG feature selection obtains better results when the number of features exceed 900, the maximum value of f-measure was 89.50% and it was obtained for 1000 features. For a small number of features this method does not select the best attributes, but when the number of attributes increases, this method attains better results.

MI feature selection increase as the number of selected features. The highest value of f-measure was 83.50%, the average value is 51.28%. It obtained the worst value compared to others feature selections using SVM and DT classifiers.

ImpCHI feature selection obtains better results when the number of features between 40 and 900, the best value of f-measure was 90.50% and it was obtained for 900 features.

Chi-square, which achieved the best f-measure compared to sixteen features selection on Arabic text classification (Mesleh, 2011), does not give the best result compared to ImpCHI.

The average value of the proposed method achieved better performance with an increase of 10.52% compared to the average value of f-measure (Bahassine et al., 2017) using the same corpora.

Table 6
Precision values for SVM and DT classifier with the four FS at different sizes of features.

Classifier	FS	20	40	100	300	500	700	900	1000	1400	Avg
DT	Chi-square	45.60	63.40	72.70	75.40	74.70	76.00	73.30	73.40	74.60	69.90
	IG	61.30	70.00	76.80	79.00	78.80	76.00	75.90	76.40	75.10	74.37
	ImpCHI	70.80	74.90	76.70	78.90	75.20	76.00	76.50	74.90	75.00	75.43
	MI	15.60	15.60	37.80	48.70	74.40	74.50	74.50	77.80	74.20	54.79
SVM	Chi-square	53.70	70.00	80.70	87.20	88.40	88.80	88.10	88.10	87.60	81.40
	IG	68.00	74.50	82.30	87.30	87.30	88.90	88.80	89.70	89.20	84.00
	ImpCHI	69.70	78.60	83.80	88.90	88.90	89.40	90.80	89.30	88.20	85.29
	MI	16.90	17.70	42.30	65.50	79.30	82.00	81.00	81.20	84.50	61.16

Table 7
Recall values for SVM and DT classifier with the four FS at different sizes of features.

Classifier	FS	20	40	100	300	500	700	900	1000	1400	Avg
DT	Chi-square	58.30	66.20	73.00	76.80	74.50	75.70	74.30	74.40	74.20	71.93
	IG	61.40	71.20	77.40	79.50	78.40	76.40	76.10	76.40	74.50	74.59
	ImpCHI	70.40	75.70	77.60	79.40	76.00	76.10	76.00	74.60	74.90	75.63
	MI	30.10	30.10	40.60	55.40	68.60	72.20	71.60	74.20	73.90	57.41
SVM	Chi-square	57.70	66.90	79.30	87.30	88.20	88.60	88.40	880	87.70	81.34
	IG	65.10	74.90	82.10	87.10	87.60	89.00	88.90	89.70	89.20	83.73
	ImpCHI	68.00	79.10	84.10	89.00	88.90	89.50	90.50	89.20	88.20	85.17
	MI	29.90	29.90	41.00	57.60	69.70	74.70	76.70	79.40	84.40	60.37

Table 8
F-measures values for SVM and DT classifier with the four FS at different sizes of features.

Classifier	FS	20	40	100	300	500	700	900	1000	1400	Avg
DT	Chi-square	49.00	61.80	72.50	75.70	73.90	73.60	72.30	72.40	73.30	69.39
	IG	59.30	70.40	76.60	79.10	77.80	74.20	73.90	74.20	73.30	73.20
	ImpCHI	69.60	75.00	76.50	78.10	74.50	75.10	74.00	73.80	74.30	74.54
	MI	15.30	15.30	30.80	47.30	65.70	71.50	70.40	72.20	73.00	51.28
SVM	Chi-square	47.00	62.60	79.40	87.20	88.20	88.50	88.10	87.80	87.50	79.59
	IG	62.10	74.70	82.20	87.00	87.20	88.80	88.60	89.50	89.00	83.23
	ImpCHI	67.30	78.50	83.80	88.90	88.90	89.40	90.50	89.10	88.00	84.93
	MI	15.00	15.00	31.20	51.10	67.50	73.20	74.70	77.60	83.50	54.31

It is not easy to choose the features selection that is always most effective for all features selection at different sizes of features. However, ImpCHI is mostly among the best one.

Based on these findings, it can be inferred that ImpCHI algorithm and SVM classifier significantly improve the categorization accuracy of Arabic text classification in terms of precision, recall and f-measures.

7. Conclusion

The present study reports the results of an improved feature selection algorithm combined with decision three at one interval and with SVM at another interval on text classification and compares the impact of this approach with results of text classification using Chi-square, MI and GI. The results indicated that the use of ImpCHI and SVM in text classification has achieved better findings than the use of Chi-square, MI and GI. In future research, we will try to apply this method to data in other languages in order to improve the generalizability of our improved algorithm. Besides, we will attempt to generalize the notion of balancing of attributes per class in other features selection algorithms.

References

Al-Anzi, F.S., AbuZeina, D., 2017. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. J. King Saud Univ. – Comput. Inform. Sci. 295 (2), 189–195. https://doi.org/10.1016/j.jksuci.2016.04.001.

Alghamdi, H.M., Selamat, A., 2019. Arabic Web page clustering: A review. J King Saud University – Comput. Inform. Sci. 31 (1), 1–14. https://doi.org/10.1016/j.jksuci.2017.06.002.

Al-Shalabi, R., Evens, M., A Computational Morphology System for Arabic, in: the Proceedings of the Workshop on Computational Approaches to Semitic Languages, Montreal, Quebec, Canada, 1998. Doi: 10.3115/1621753.1621765

Attia, N., Djelloul, Z., Hadda, C., 2016. Rational kernels for Arabic root extraction and text classification. J. King Saud Univ. – Comput. Inform. Sci. 28 (2), 157–169. https://doi.org/10.1016/j.jksuci.2015.11.004.

Ayedh, A.A., uanzheng, T., TAN, A., Alwesabi, K., Rajeh, R.H., 2016. The effect of preprocessing on arabic document categorization. Algorithms 9 (2). https://doi.org/10.3390/a9020027.

Bahassine, S. Kissi M. Madani, A. New Stemming For Arabic Text Classification Using Feature Selection and Decision Trees, in: Proceedings of the 5th International Conference on Arabic Language Processing (CITALA) Oujda, Morocco, (2014) pp. 200–205.

Bahassine, S., Madani, A., Kissi, M., 2016. An improved Chi-sqaure feature selection for Arabic text classification using decision tree, 11th International Conference on Intelligent Systems: theories and Applications (SITA). Mohammedia, 1–5. https://doi.org/10.1109/SITA.2016.7772289.

Bahassine, S., Madani, A., Kissi, M., 2017. Arabic text classification using new stemmer for feature selection and decision trees. J. Eng. Sci. Technol. 12 (6), 1475–1487.

Baraa, S., Nazlia, O., Zeyad, S., 2014. An automated Arabic text categorization based on the frequency ratio accumulation. Int. Arab J. Inform. Technol. 11 (2), 213–221.

Barigou, F., 2016. improving k-nearest neighbor efficiency for text categorization. Neu. Network World 26 (1), 45–65. https://doi.org/10.14311/Nnw.2016.26.003.

Boudchiche, M., Mazroui, A., Ould Abdallahi, O.B., Lakhouaja, A., Boudlal, A., 2017. AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer. J. King Saud Univ. – Comput. Inform. Sci. 29 (2), 141–146. https://doi.org/10.1016/j.jksuci.2016.05.002.

W. Cherif, A. Madani, M. Kissi, Building a Syntactic Rules-Based Stemmer to Improve Search Effectiveness for Arabic Language, in: 9th International Conference in Intelligent Systems: Theories and Applications (SITA), (2014) pp. 1–6. doi: 10.1109/SITA.2014.6847295

Cherif, W., Madani, A., Kissi, M., 2015. New Rules-Based Algorithm to Improve Arabic Stemming Accuracy. Int. J. Knowl. Eng. Data Min. 3, 315–336. https://doi.org/10.1504/IJKEDM.2015.074082.

Cherif, W., Madani, A., Kissi, M., 2016. A hybrid optimal weighting scheme and machine learning for rendering sentiments in tweets. Int. J. Intell. Eng. Inform. (IJIEI) 4 (3/4), 322–339. https://doi.org/10.1504/IJIEI.2016.080527.

Dadgar, S.M.H., Araghi, M.S., Farahani, M.M. A novel text mining approach based on TF-IDF and Support Vector Machine for news classification, in: IEEE International Conference on Engineering and Technology (ICETECH), Coimbatore, (2016) pp. 112–116. doi: 10.1109/ICETECH.2016.7569223.

Elhassan, R., Ahmed, M., 2015. Arabic text classification review. Int. J. Comput. Sci. Software Eng. 4 (1), 1–5.

Hammad, A.A., El-Halees, A., 2015. An approach for detecting spam in arabic opinion reviews. Int. Arab J. Inform. Technol. 12, 9–16.

Harish, B.S., Revanasiddappa, M.B., 2017. A Comprehensive survey on various feature selection methods to categorize text documents. Int. J. Comput. Appl. 164 (8), 1–7. https://doi.org/10.5120/ijca2017913711.

Harrag, F., El-Qawasmeh, E., Pichappan, P., 2009. Improving arabic text categorization using decision trees, First International Conference on Networked Digital Technologies. Ostrava, 110–115. https://doi.org/10.1109/NDT.2009.5272214.

Jadon, E., Sharma, R., 2017. Data mining: document classification using naive bayes classifier. Int. J. Comput. Appl. 167 (6), 13–16. https://doi.org/10.5120/ijca2017913925.

Jo, T., 2017. K nearest neighbor for text summarization using feature similarity. Int. Conf. Commun. Control, Comput. Electron. Eng. (ICCCCEE), Khartoum, 1–5. https://doi.org/10.1109/ICCCCEE.2017.7866705.

Junkai, Y., Guang, Y., Jing, W., 2016. Category discrimination based feature selection algorithm in chinese text classification. J. Inform. Sci. Eng. 32 (5), 1145–1159.

Khoja, S., Garside, R., 1999. Stemming Arabic Text. Lancaster University, UK Computing Department.

Kilimci, Z.H., Akyokus, S., Omurca, S.I., 2016. The effectiveness of homogenous ensemble classifiers for Turkish and English texts. Int. Symp. INnov. Intell. SysTems Appl. (INISTA), 1–7. https://doi.org/10.1109/INISTA.2016.7571854.

Kissi, M., Ramdani, M., 2011. A hybrid decision trees-adaptive neuro-fuzzy inference system in prediction of anti-HIV molecules. Expert Syst. Appl. 38 (5), 6376–6380. https://doi.org/10.1016/j.eswa.2010.11.011.

Mahmoud, H., Hamdy, M.M., Rouhia, M.S., 2016. Arabic text categorization using mixed words. Int. J. Inform. Technol. Comput. Sci. (IJITCS) 8 (11), 74–81. https://doi.org/10.5815/ijitcs.2016.11.09.

Mesleh, A.M.d., 2011. Feature sub-set selection metrics for Arabic text classification. Patt. Recog. Lett. 32, 1922–1929. https://doi.org/10.1016/j.patrec.2011.07.010.

Mostafa, A.M., 2017. An evaluation of sentiment analysis and classification algorithms for Arabic textual data. Int. J. Comput. Appl. 158 (3), 29–36. https://doi.org/10.5120/ijca2017912770.

Nehar, A., Ziadi, D., Cherroun, H. Guellouma, Y., An efficient stemming for Arabic Text Classification, International Conference on Innovations in Information Technology (IIT), Abu Dhabi, (2012) 328-332. doi: 10.1109/INNOVATIONS.2012.6207760.

Nikhath, A.K., Subrahmanyam, K., Vasavi, R., 2016. Building a K-nearest neighbor classifier for text categorization. Int. J. Comput Sci. Inform. Technol. (IJCSIT) 7 (1), 254–256.

Rasha, M., Mahmoud, A., 2016. Arabic text stemming: comparative analysis, Conference of Basic Sciences and Engineering Studies (SGCAC). Khartoum, 88–93. https://doi.org/10.1109/SGCAC.2016.7458011.

Saad, M., Ashour, W., OSAC: Open Source Arabic Corpora, in: EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus. (2010) pp. 118–123. DOI: 10.13140/2.1.4664.9288

Sawalha, M., Atwell, E.S., 2008. Comparative Evaluation of Arabic Language Morphological Analysers and Stemmers. Coling Organizing Committee, Manchester, pp. 107–110.

Sebastiani, F., 2016. Machine learning in automated text categorization. ACM Comput. Surv. 34 (1), 1–47. https://doi.org/10.1145/505282.505283.

Singh, V., Saxena, P., Singh, S., Rajendran, S., 2017. Opinion mining and analysis of movie reviews. Indian J. Sci. Technol. 10 (19).

Suhad, I.E., Yousif, A., Venus, W., Samawi Zantout, R., 2015. The effect of combining different semantic relations on arabic text classification. World Comput. Sci. Inform. Technol. J. (WSCIT) 5 (6), 112–118.

Uysal, A.K., Gunal, S., 2014. The impact of preprocessing on text classification. Inf. Process. Manage. 50, 104–112. https://doi.org/10.1016/j.ipm.2013.08.006.

Yang, Y., Pedersen, J.O., 1997. A comparative study on feature selection in text categorization. Proceed. Fourteenth Int. Conf/ Mach. Learn., 412–420

Ye-wang, C., Qing, Z., Wei, L., Ji-Xiang, D., 2016. Classification of Chinese texts based on recognition of semantic topics. Cognit. Comput. 8 (1), 114–124. https://doi.org/10.1007/s12559-015-9346-8.