Multimedia Tools and Applications (2021) 80:10373–10390
https://doi.org/10.1007/s11042-020-10074-6

Feature selection method using improved CHI Square on Arabic text classifiers: analysis and application

Hadeel N. Alshaer[1] · Mohammed A. Otair[1] · Laith Abualigah[1] · Mohammad Alshinwan[1] · Ahmad M. Khasawneh[1]

Received: 6 June 2020 / Revised: 31 August 2020 / Accepted: 13 October 2020 /
Published online: 21 November 2020

Abstract
Text classification could be defined as the way of allocating text into predefined groups according to its contents. Over the past few years, an increase emerged in the volume of information in the varied fields on the Internet, thus making the classification of texts one of the most important, yet challenging. Text classification is commonly employed in numerous applications and for different objectives. The extensive and broad use of the Internet, particularly in the Arab world, as well as the massive number of the documents and pages which are provided in the Arabic language, raised the need for having suitable tools for classification of these pages and documents by their main categories. The aim of this paper to study the effect of the improved CHI (ImpCHI) Square on the performance of six well-known classifiers: Random Forest, Decision Tree, Naïve Bayes, Naïve Bayes Multinomial, Bayes Net, and Artificial Neural Networks. These proposed techniques are quite important for improving classification of Arabic documents and can be regarded as a promising basis for the stage of text classification because it contributes to the classification of the texts into predefined categories. This combination method takes the advantages of more than one technique, which can produce better results in the final outcomes. The dataset employed in this paper includes 9055 Arabic documents that were collected from various Arabic resources. Based on their content, these documents were divided into twelve categories. Four performance evaluation criteria were used: the F-measure, recall, precision, and Time build model. The experimental results show that the use of ImpCHI square gives better classification results than the normal CHI square method with all studied classifiers, in terms of all used performance criteria.

Keywords Text classification algorithms · Bayes net · Naïve Bayes · Random Forest · Decision tree · Artificial neural networks · CHI Square

✉ Laith Abualigah
aligah.2020@gmail.com

Extended author information available on the last page of the article

1 Introduction

Information Retrieval (IR) is a field of computer science of great importance in our time because of the increasing volume of information. This information may need to be arranged and classified so that it can be easily retrieved. Text classification (TC) a process that has been emerged importantly in various fields, especially in areas on the Internet. Text mining is a textual analysis of data in natural language text and seeks to extract useful information from textual data. Besides, text mining helps organizations extract valuable ideas from document content. The text mining process is used to increase the efficiency of the text retrieval process by discovering patterns in the text and the relationships between them to help the retrieval of texts correctly [7]. The most important applications on Text Mining are IR, Information Extraction, Classification, and Natural Language Processing (NLP). Meaning it extracts useful information from a large amount of data to replace data and information search problems [1]. In this paper, text data will be used in Arabic, the text classification is to place the text document in the appropriate category according to its content. There is a need to classify the texts because of the huge amount of text documents that are uploaded on the Internet daily [24].

Many researchers tried in various ways to study the factors that positively affect the improvement of the process of classification. Where there is a phase before the classification process has been studied and improved on them, beginning the way the data collection and try to follow the right way through the pre-processing steps: Tokenization, Normalization, Stop Word Removal, and Stemming. Then feature selection is considered an important and effective process for the text classification system. Feature selection is one of the key factors that can greatly improve the performance of the Classification process and is the removal of irrelevant and redundant data and the selection of important data aimed at reducing the complexity of the classification process [17]. The feature selection process is an effective text classification process, and the preprocessing of the dataset is very important to achieve the best classification results. The feature selection method improves the text classification in several aspects: speed of learning, effectiveness, and comprehensiveness. There are many examples of the feature selection methods, including Mutual Information (MI) Information Gain (IG) and CHI Square [29]. Also, many optimization techniques have been used to solve the feature selection problems by improving the classification process, such as in [20, 2, 12]. Furthermore, a new optimization algorithm can be employed to enhance the feature selection methods.

Document Classification faces several problems. One of which is how to choose a reasonable feature from the set of words in the source document. The CHI square method is one of the most frequently used methods for the selection of the features in the text classification process. In this paper [45], the importance of the selection process and its effect on the overall classification process is given, the researcher used the ImpCHI Square method based on the Normal CHI square method to select the best possible features that have the highest effect on the final results of the classification process and thus obtain the optimum possible classification results. CHI Square was improved and used with Chinese texts. It was found that the use of ImpCHI Square showed better results than the traditional CHI Square and was used with classification algorithms: SVM and K-NN. That this improvement led to the efficiency of the results of the classification process. Also, CHI Square was developed with the SVM Classifiers in one system based on the Arabic text in the CHI Square method for selecting the distinctive words using SVM Classifiers in a multi-stage system, one of the stages of the feature selection phase [31].

This paper study the effect of Improved CHI square (ImpCHI) and using Bayes Net (BN), Naïve Bayes (NB), Naïve Bayes Multinomial (NBM), Random Forest (RF), Decision Tree (DT), and Neural Networks (NNs) for Arabic TC problems. The Arabic text classification system is a multi-stage that is based on the successive processes of pre-processing, feature selection, and classification. In the Arabic text classification systems, the feature selection process is a highly critical step. It cuts down the document size by eliminating non-important words from the document and picking important features that foster the classification process concerning accuracy and speed in the text classification system. This Paper examines if an ImpCHI Square will offer as a feature selection on classification processes for its use with some of the existing classification algorithms. Moreover will be presented an Arabic Text Classification System, and an explanation of the work of each step in detail. Tried searching factors that positively affect the improvement of results, and then display and compare results through several criteria: precision, Recall, F-measure and Time build model, discuss and analyze these results. This combination method takes the advantages of more than one technique, which can produce better results in the outcomes. The dataset employed in this paper includes 9055 Arabic documents that were collected from various Arabic resources. Based on their content, these documents were divided into twelve categories. Four performance evaluation criteria were used: the F-measure, recall, precision, and Time build model. The experimental results show that the use of ImpCHI square gives better classification results than the normal CHI square method with all studied classifiers, in terms of all used performance criteria.

The rest of this paper is organized as follows. Section 2 presents the most related previous studies in this domain. Section 3 presents the data processing in this paper. The conventional classifiers are presented in Section 4. The Arabic text classification algorithms are presented in Section 5. Experimental results and discussions are presented in Section 6. Conclusion and future works are presented in Section 7.

2 Previous studies

Raho et al. in [36] mentioned the feature selection is necessary for efficient text classification and pre-processing of the dataset is essential to achieving an effective result and performance. Different classifiers were also compared using the feature selection process with the stemming process. Without the stemming process, BBC Arabia dataset was used and different classification algorithms were used: NB, K-NN, DT, and NBM. In addition, the results were presented and compared based on several criteria: Accuracy, precision, Recall, F-measure, and Time to build the model. Feature selection is performed in several steps: (1) Feature Generation: In this step, a subset of features is generated using some searches process, (2) Feature Evaluation: In this step, some evaluation matrices were used to measure the quality of the selected features, (3) Feature Validation: In this step, Verification is performed whether features valid or not. Feature selection has been verified with four classifiers: NB, K-NN, DT, and NBM using the dataset in Arabic it was found that the performance and accuracy of NB, DT, and NBM were higher than the performance K-NN in all cases.

Mesleh in [28] suggested the Arabic Text Classification System using Support Vector Machine (SVM) and CHI Square method as a feature selection in the pre-processing step. Compared to other classification methods, this system shows the high efficiency of the Arabic dataset. The feature selection in the classification of texts is important and deals with large

areas of data and the need to find a way to complete the feature selection process is effective. Examples of these methods are more common: Document Frequency (DF), CHI Square (CHI), Term Strength (TS), Information Gain (IG), and Mutual Information (MI). In the testing process, the Arabic dataset was used, stop words were removed, filter out of non-Arabic words, and symbols were removed and the stemming process was not applied. The performance of CHI Square as a selection feature with the SVM classifier has achieved scientifically positive results for Arabic text classification. However, Recall and Precision values have not been improved.

Hawashin et al. in [23] suggested an effective method for CHI Square feature selection to Arabic Text Classification, In Data Mining, and the feature selection process is a pre-processing step to improve the performance of the classification process. Different datasets data have been used through different works. Three aspects of the previous work have been proposed: First: A new method of effective feature selection is proposed to promote Arabic Text Classification, Second: it compares an expanded number of existing feature selection methods, Third: Adopt two publicly available dataset to encourage future businesses to adopt them to ensure fair comparisons between different businesses. Many feature selection has been used: CHI Square, Mean TF.IDF, DF, IG, Feature Sub-set Selection (FSS), and Wrapper Approach (WA). In addition, two of the dataset were used: Akhbar Alkhalij it contains (5692) Texts classified to four Classes, and Alwatan News it contains (5250) Texts classified to five Classes. A new and effective method has been proposed as a selection feature based on CHI Square statistics. This method has outperformed several existing feature selections according to their effect on Arabic TC. The proposed method has outperformed several feature selections: CHI Square, DF, and FSS with Best-Search, WA with SVM, Best-Search, and IG.

Olimat in [31] suggested Arabic Text Classification based on CHI Square as a feature selection using SVM Classifiers. The system consists of three main phases: pre-processing, feature selection, and classification. The first phase contains three steps: At the first step the documents and input texts are prepared into words through the Tokenization process, At second step the stop words, the non-Arabic words, and the special symbols are then removed by the Stop word removal process, the last step is Stemming process is then performed to remove the prefix and Suffix from words. At phase two perform the feature selection process to serve as inputs for the third phase, the classification phase. The Arabic TC system was compared to CHI Square and TF.IDF using the same pre-processing phase. The classification process was completed using the SVM, all of which are applied to the same dataset and were compared by Precision, Recall, and F1-measure. The F1-measurement values showed that the improved CHI Square method gave results to Arabic TC better than normal CHI Square and TF.IDF.

Bahassine et al. in [17] used improved CHI Square with DT Classifier for Arabic TC, mentioned that feature selection is an important and necessary process, a step that can greatly improve the performance of the classification process. The Aims to investigate a new feature selection called ImpCHI when the Light Stemming process is used. ImpCHI is an improvement of the regular CHI Square. The ImpCHI performance was evaluated using a dataset in Arabic consisting of 250 documents classified into five categories: Art and Culture, Economics, Society, Politics, and Sport. The results showed that Arabic TC using ImpCHI outperformed CHI Square in terms of Recall.

Zheng et al. in [45] suggested ImpCHI method using the Chinese language and classification using SVM Classifier. ImpCHI Square was used instead of CHI Square and therefore had several disadvantages. CHI Square has been improved and access to more advanced Feature

Selection and improved access has improved the results of the classification process. ImpCHI Square is now used as feature selection before the classification process to improve its results. The results showed that the use of ImpCHI Square on the Chinese language using SVM Classifiers led to improved performance measures. The results showed that the use of ImpCHI Square is an effective method in the feature selection process.

Mesleh in [29] mentioned FSS an important step for text classification systems, FSS is assigned to the TC functions. Using the SVM classifier and dataset is used in Arabic. The FSS has several tasks: (1) FSS improves the performance of text classification under conditions: speed of learning, effectiveness, and comprehensiveness. (2) FSS reduces the number of data dimensions plus it removes irrelevant and duplicates data. (3)The FSS provides a deep insight into the basic processes that generate data. Several FSS were used: CHI square, IG, MI, GSS score, NGL score, Odds ratio, DF, Bi-Normal separation, Power prefers frequent. The accuracy measures are Precision, Recall, Fallout, Error rate, and F1-measure. Results show that CHI square and fallout (FSS) works best with text classification tasks. Other applications that can use the feature selection techniques can find in [40–44]. An overview of the studied papers is given in Table 1.

3 Data preprocessing

Pre-processing is an important step in the classification process and supports the results and increases their efficiency; so it is important that we highlight them in order to get the best results. The Tokenization step is the process of dividing the sentences and texts into pieces called Tokens, which is to cut the sentences according to the white space followed by the Normalization step, through which to create a database containing synonyms for words in Arabic. This is followed by the Stop Word Removal step. At this stage, Stop words are removed from the text. These stop words are stored in a list until they are removed from the text. Finally, the Stemming step is returned to the root by deleting the prefix and suffix of the word if they belong to a set of letters that have an interest in

Table 1 An overview of the studied methods

Reference	Name	Evaluation measure	Method	Applications	Year
[36]	Raho et al.	Performance and accuracy	Different classifiers were also compared using the feature selection process with the stemming process	Text classification	2015
[28]	Mesleh	Recall and precision	Arabic Text Classification System using Support Vector Machine (SVM) and CHI Square method	Text classification	2007
[23]	Hawashin et al.	Performance and accuracy	An effective method for CHI Square feature selection to Arabic Text Classification	Text classification	2013
[31]	Olimat	Precision, Recall, and F1-measure	Arabic Text Classification based on CHI Square as a feature selection using SVM Classifiers	Text classification	2017
[17]	Bahassine et al	Recall	Used improved CHI Square with DT Classifier for Arabic TC	Text classification	2016
[45]	Zheng et al	Performance measures	Suggested ImpCHI method using the Chinese language and classification using SVM Classifier	Text classification	2016
[29]	Mesleh	Precision, Recall, Fallout, Error rate, and F1-measure	FSS is an important step for text classification systems, FSS is assigned to the TC functions	Text classification	2013

Such as: (ال ـ ت ـ ن ـ وا ـ و ـ ون - ين). Fig. 1 shows the steps of this phase from data collection to classification. These processing are given below.

- **Tokenization** is a NLP which is very important, and is a preparatory stage for all stages of the pre-processing and other task of separating words from running text, which is the wholesale segmentation and the sentence is divided into a series of consecutive words, can be used spaces "White Spaces" to help with this task [13]. Tokenization is difficult for "Scripto Continua" languages such as Ancient Greek, Chinese and Thai. This complexity and difficulty because it does not contain "white spaces" between words or special characters [1]. Tokenization is an important step in NLP and is closely associated with the morphological collection process, but is often seen as an independent process [19].

- **Normalization** is a process of converting the dataset from words that sequentially to more consistent and accuracy, and the process is done by converting words into a standard model through operations that make them able to manipulate data. Normalization improves text matching by taking into account: the synonyms of some words, the type of writing, and the abbreviations. It is an important process of texts used in the process of retrieving information and improving the process to obtain the best results [14]. The performed in various pre-processing stages in order to convert differing forms of articular letter into unicode representation. For example, replacement of the un-dotted Arabic letter (ى) with the dotted letter (ی) when the former letter lies at end of Arabic word [9].

- **Stop word removal** is a process of removing the stop word from text based on a list containing stop words. The stop words in the English language comprise articles like 'an',

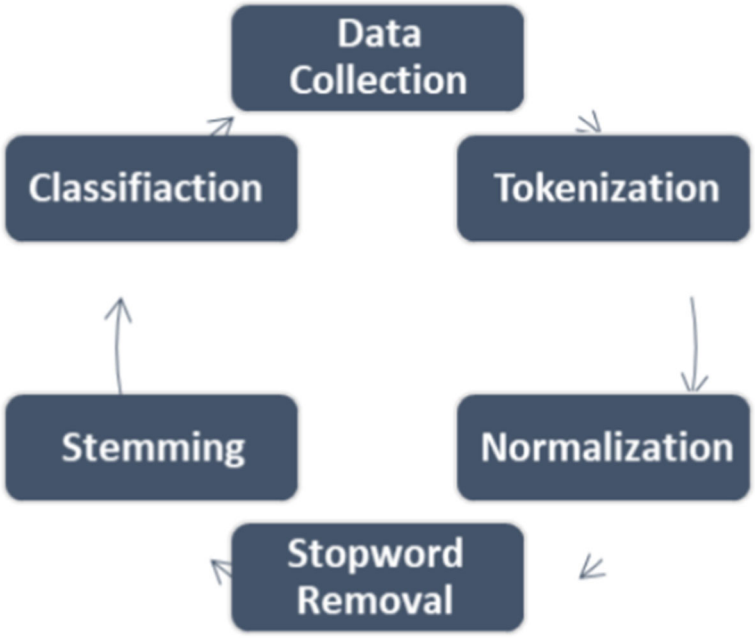

Fig. 1 Pre-processing life cycle

'the', and 'a', and demonstratives such as 'those', 'this', and 'that' [39]. Removal of these widely-appearing words from the indices cuts down the number of the words which each search term needs to be compared with, therefore greatly improving the query response time without influencing the classification accuracy. Similarly, the stop words in the Arabic language include every word that is not counted as part of the speech, namely, verb or noun. They encircle demonstratives (هذا، هذه، ...); prepositions (عن، في، على ...; adverbs (فوق، حتى، ...); and special characters ($, %,and,…) [15].

- **Stemming** is a process are used to find out the root of a word. The process of stemming converts the words into root and an example in English the following words: "User, Uses and Used" are reverted to their root "Use" [39]. The stemming process decreases size of document representation by 20–50.0% of the size of the full word representation [38]. Hence contributing to an improvement of document retrieval. Arabic stemmers is a technique aimed at finding the stem or lexical root of words in Arabic language by removing natural affixes from the word, because Arabic language contains morphological structures more complex than other languages. The algorithms of stemming can be categorized in Arabic eligible analysis degree: stem-based approach and root-based approach [33]. P-Stemmer is a new Arabic light stemmer, a modified version of one of Larkey's light stemmers. Stemmer specialist in Arabic, where his work to return the words to their roots such as "**الإقتصاد**" become "**قصد**" and based on removing prefixes and suffixes Such as "**كالصادرات**" become "**صادرات**" and "**والوحدات**" become "**وحدات**" [26].

4 The common classifiers

- **CHI SQUARE**

A statistical method by selecting random data based on two variables and independent variables are extracted from a large sample of data which is a feature selection method, used in the Data Mining process as a feature selection method. Used CHI square method in preprocessing step of the Text Classification system [28].

- **IMPROVED CHI SQUARE**

Improved CHI Square (impCHI) it is an extension of the CHI Square method by improving the properties after the basic method, ImpCHI Square was used with Chinese language and the results proved the effectiveness of this feature selection of textual data in Arabic (Zheng et al., 2016). In addition, ImpCHI square was used with the Arabic language, which was used with the Decision Tree when using the light stemming process. The results showed that the ImpCHI outperforms using normal CHI in terms of the recall measure [17].

- **TEXT CLASSIFICATION**

The classification of texts is the process of distinguishing text into any category based on specific text data for each category and trying to distinguish the text according to its characteristics. Words appearing in the same category often contain multiple morphological

structures and similar grammatical structures. In most cases, words in the same category have one or similar morphological structure (Bahassine et al., 2017). Documents are sorted into fixed number of predefined categories. The documents to be categorized may be multi-category or not suited to any category of knowledge. In this case, each set of documents with the same characteristics is in a group that is relevant to its content and makes it easier to search for a certain document [4].

5 Arabic text classification algorithms

Until the classification process is done, we need algorithms to complete the process since classification algorithms are the algorithms that perform the classification process the training process based on a categorized dataset to the process of testing for a new text and distinguish it to which category it belongs. The classification process is supervised learning, as supervised learning is a machine learning task and its function is to assign inputs to outputs. This learning creates from a set of training data. Supervised Learning Algorithms analyze training data to understand the content of testing data. Supervised learning requires mapping between a set of variables, namely the inputs of the input X and the output of the Y output. This mapping applies to predict the output of unseen data [32].

- **Bayes Net (BN)** is a model that reflects situations that are part of the world and describes how they relate to each other. This model may be linked to any Entity in this world and can be represented by BN and all existing and potential Entities that can be modeled by BN. The possibilities come from the assumption that some Entities occur frequently when another Entity exists. This model is useful because it helps us understand the world we want to model and helps us predict the results of the Entity in this world. This model is easy to represent and model the world and the Entity through it [14].

- **Naïve Bayes (NB)** is a classifier of the Bayes family and is based on the Theorem Bayes. Assuming independence among beginners, the NB model is easy to build and very useful with large numbers of data, although easy to install and often outperforms its work. NB is an algorithm used in the classification process. Bayes Classifiers are widely used in automatic learning because they are easy to implement [14]. The NB approach is used to deal with document classification problems through a simple and simplified model where the NB approach is applied in a flat (liner) method and hierarchically to improve the efficiency of the results of the taxonomy model. The hierarchical classification method was found to be more effective than a flat classification. It also performs better if multiple document labels are categorized [25]. NB classifier was used with the Bayes family classifiers, which contain the following classifiers: Bayes Net, Naïve Bayes, Naïve Bayes Multinomial, Naïve Bayes Multinomial Text, Naïve Bayes Multinomial Updateable, and Naïve Bayes Updateable, The study proved outperforms to the NB classifier on the other classifiers in terms of precision, Recall and F-measure [14]. In an analysis of NB, the final classification using it is produced by combining all sources of information, and a probability is formed based on the training data to classify the document into the category to which it belongs [18].

- **Naïve Bayes Multinomial (NBM)** is a specialized version of NB, which is designed more accurately for text files, which is the process of classification after the accuracy of entering training data. It automatically classifies the categories we have introduced through the automated learning process [14]. Using a multi-dimensional probability model such as the Bayesian theorem tries to overcome the drawback of using the Bernoulli multivariate model*, which represents a text document as a binary attribute indicating words that do not appear in the document [27]. On the other hand, the probabilistic polynomial model is a uni-grams language model* with an integer of words. The document is represented by a set of word repetitions in the document. Where the number of times each word appears in the document is recorded [3]. Examines the effect of ambiguous patterns and verifies their reduced impact on the low-frequency downside, to overcome this Naïve Bayes Multinomial is used [30].

- **Random Forest (RF)** is a collection of classification algorithms that are widely used in many applications, especially with the large dataset, because of its characteristic features: Measurement of changing significance, Out of Band (OOB) detection, proximity to feature, and handling of unbalanced data. The Random Forest algorithm is used in many applications, including Network intrusion detection, Email spam detection, gene classification, Credit card fraud detection, and Text classification [45]. Random Forest is a new and powerful statistical classifier. The advantages of random forest classification are features of the classification accuracy, a new way of identifying important variables, the ability to work in the complex interaction model between statistical data analysis and an algorithm for assigning missing values [21].

- **Decision Tree (DT)** is an algorithm that performs classification, it is used in the field of Data Mining in various fields on the Internet. Also, Text Mining Algorithms, DT is used in Arabic TC and pre-processing algorithms [22]. In addition, DT was used in the Data Mining Education process, DT approach, and decision base approach. Given the global opportunities associated with global competition, even in the case of education, it is necessary to accept the best students to the maximum extent possible. Academic performance and subsequent attitudes are the best in the world [8]. The CHI square as a feature selection method was optimized and tested using the Decision Tree algorithm, using an Arabic data set of 250 documents. The results

Table 2 Results Based on Avg. precision

Algorithm	Without pre-processing	With pre-processing	Without pre-processing and CHI	With pre-processing and CHI	Without pre-processing and impCHI	With pre-processing and impCHI
BN	0.805	0.828	0.858	0.851	**0.893**	0.883
NB	0.913	0.905	0.926	0.935	**0.976**	0.944
NBM	0.915	0.912	0.935	0.926	0.952	**0.956**
RF	0.863	0.849	0.938	0.917	0.947	**0.958**
DT	0.528	0.516	0.552	0.523	**0.604**	0.577
ANNs	0.586	0.659	0.699	0.672	0.713	**0.719**

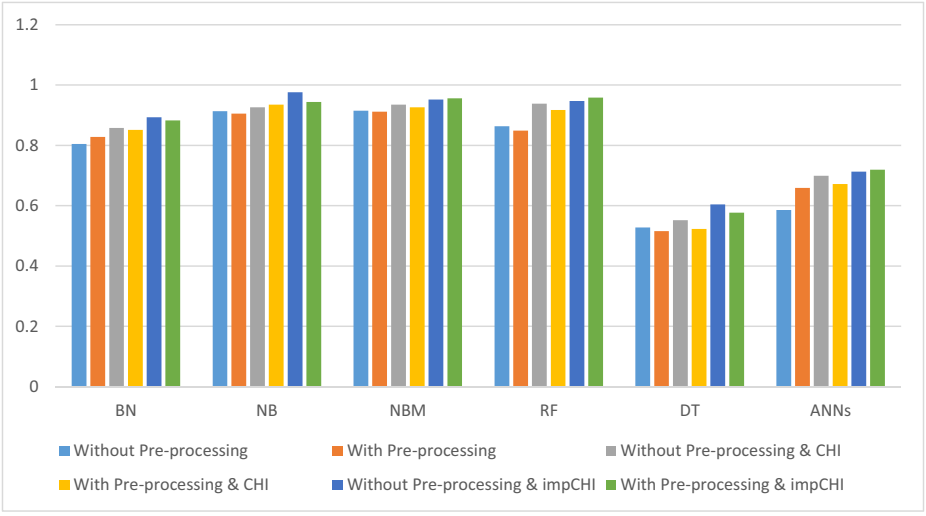

Fig. 2 Results based on Avg. precision

showed that the classification using the improved CHI square with the Decision Tree outperforms algorithm in the classification on normal CHI square [17].

- **Artificial Neural Networks (ANNs)** is a classification algorithm and is a branch of artificial intelligence. The different groups of functions were studied and their effect during the use of ANNs as classifiers and the validity of these functions is analyzed for different types of databases. The ANNs for back-propagation can be used as a successful tool to classify the data set with several training and learning functions [37]. The ANNs have been successfully applied to problems in pattern classification, function approximation, optimization, and pattern matching [34]. ANNs are usually a set of processing units such as nerves with a connection between these units. ANNs are a tool for text classification and have used a different structure from ANNs to apply text classification [35].

6 Experimental results and discussion

The choice of the best classifiers depends on the highest rate in the three comparison criteria: precision, Recall, and F-measure. The details of these measures can find in [5, 6, 10, 11]. As

Table 3 Results based on Avg. Recall

Algorithm	Without pre-processing	With pre-processing	Without pre-processing and CHI	With pre-processing and CHI	Without pre-processing and impCHI	With pre-processing and impCHI
BN	0.871	0.893	0.910	0.888	**0.940**	0.922
NB	0.951	0.944	0.950	0.956	0.954	**0.966**
NBM	0.911	0.922	0.945	0.947	**0.967**	0.965
RF	0.844	0.765	0.805	0.796	**0.942**	0.868
DT	0.523	0.504	0.537	0.502	0.518	**0.539**
ANNs	0.452	0.506	0.542	0.467	0.492	**0.560**

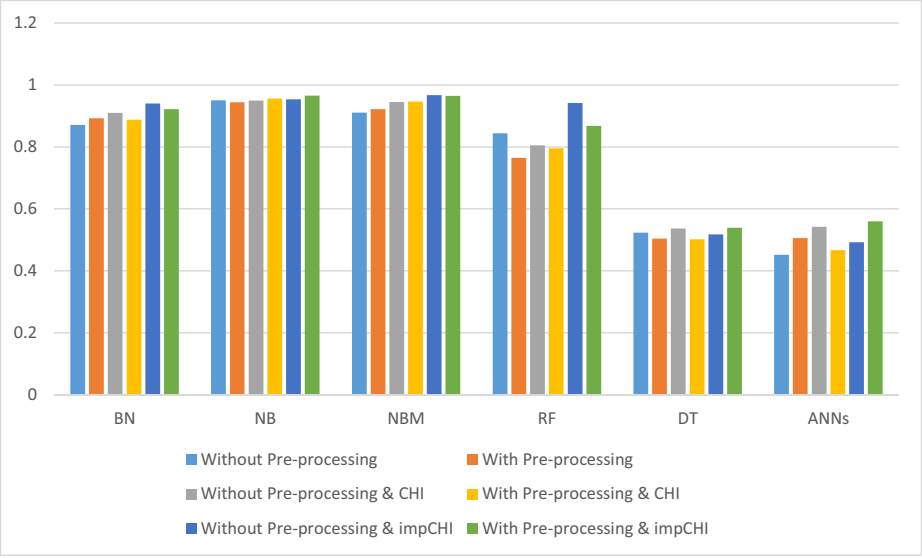

Fig. 3 Results based on Avg. Recall

for the fourth criterion, the time is preferred Classifier is the least time in the build model. In this section, the classifiers will be compared to each other according to the Average precision, Average Recall, Average F-measure, and Average Time. For a straightforward comparison, all experiments are performed utilizing MATLAB R2015b software under Windows 10, the operating system is code i7 and 16GB RAM.

We compared the Classification Classifiers according to the precision measure: BN, NB, NBM, RF, DT, and ANNs. By conducting six tests for each classifier: without pre-processing, with pre-processing, without pre-processing and CHI, with pre-processing and CHI, without pre-processing and ImpCHI, and with pre-processing and impCHI. We show that when using ImpCHI square as feature selection method, gave better results as gave better results in precision. Table 2 shows the results based on Avg. precision measure.

Whereas BN, NB, and DT gave the best results when using ImpCHI square without pre-processing. But NBM, Random Forest, and ANNs gave the best results when using ImpCHI square with pre-processing. The highest result was in Avg. precision to NB classifier the value was = 0.976. Figure 2 shows a graphical representation for Table 2. It is noticed by the bars that represent results for Avg. precision to ImpCHI square have the superiority over the

Table 4 Results based on Avg. F-measure

Algorithm	Without pre-processing	With pre-processing	Without pre-processing and CHI	With pre-processing and CHI	Without pre-processing and impCHI	With pre-processing and impCHI
BN	0.838	0.986	0.884	0.878	**0.916**	0.902
NB	0.932	0.924	0.937	0.945	**0.965**	0.952
NBM	0.910	0.916	0.885	0.933	0.959	**0.960**
RF	0.834	0.844	0.859	0.856	0.910	**0.913**
DT	0.522	0.530	0.547	0.512	**0.600**	0.568
ANNs	0.519	0.572	0.607	0.570	**0.655**	0.640

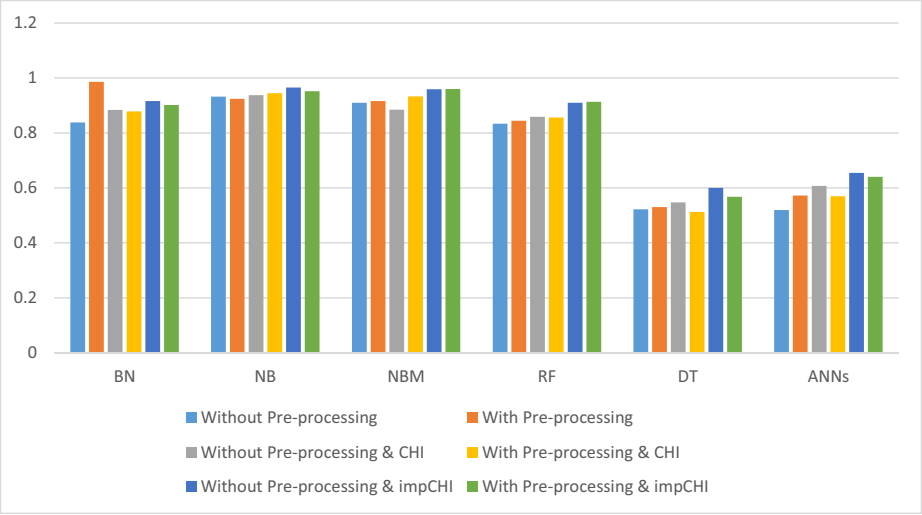

Fig. 4 Results based on Avg. F-measure

classified CHI Square. It is clear that the proposed ImpCHI without the pre-processing obtained the best results compared to all other comparative methods.

We compared the Classification Classifiers according to the Recall measure: BN, NB, NBM, RF, DT and ANNs. By conducting six tests for each classifier: without pre-processing, with pre-processing, without pre-processing and CHI, with pre-processing and CHI, without pre-processing and ImpCHI and with pre-processing and impCHI. We show that when using ImpCHI square as feature selection method, gave better results as gave better results in Recall. Table 3 shows the results based on Avg Recall.

Whereas BN, NBM and RF they gave the best results when using ImpCHI square without pre-processing. But NB, DT and ANNs they gave the best results when using ImpCHI square with pre-processing. The highest result was in Avg. Recall to NBM classifier the value was = 0.967. Figure 3 shows a graphical representation for Table 3. It is noticed by the bars that represent results for Avg. Recall to ImpCHI square have the superiority over the classified CHI Square.

We compared the Classification Classifiers according to the F-measure measure: BN, NB, NBM, RF, DT and ANNs. By conducting six tests for each classifier: without pre-processing, with pre-processing, without pre-processing and CHI, with pre-processing and CHI, without pre-processing and ImpCHI and with pre-processing and impCHI. We show that when using

Table 5 Results based on Avg. Time

Algorithm	Without pre-processing	With pre-processing	Without pre-processing and CHI	With pre-processing and CHI	Without pre-processing and impCHI	With pre-processing and impCHI
BN	16.7	16.1	18.20	22.5	**25.6**	20.3
NB	3.16	3.11	4.11	9.11	**9.19**	5.5
NBM	4.26	4.01	5.03	10.13	**11.22**	6.32
RF	71.48	61.32	66.41	91.21	**94.3**	70.1
DT	1.01	1.05	2.51	3.34	4.40	3.0
ANNs	36.34	39.33	43.7	39.11	44.17	**51.2**

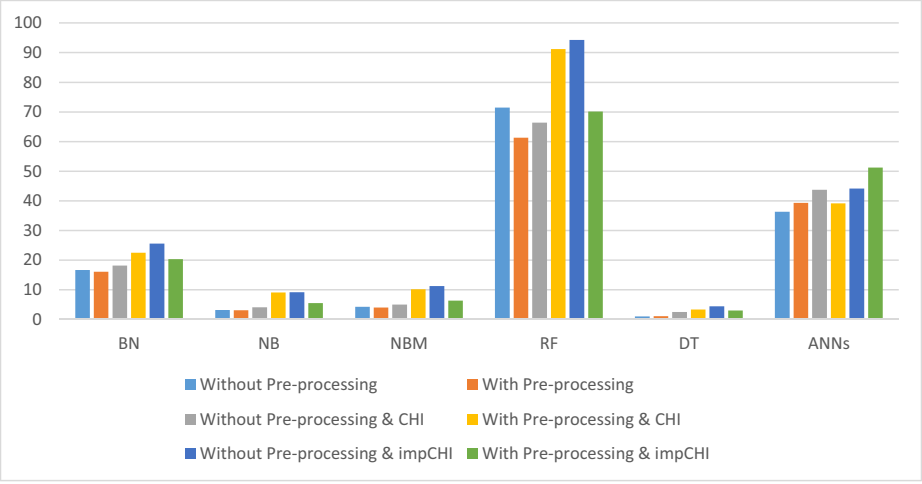

Fig. 5 Results based on Avg. Time

ImpCHI square as feature selection method, gave better results as gave better results in F-measure. Table 4 shows the results based on Avg. F-measure. It is clear that the better methods are the proposed ImpCHI without pre-processing, which got the four best results out of six. The second best method is the proposed ImpCHI with pre-processing, it got the two best results out of six.

Whereas BN, NB, DT and ANNs they gave the best results when using ImpCHI square without pre-processing. But Naïve Bayes Multinomial and Random Forest they gave the best results when using ImpCHI square with pre-processing. The highest result was in Avg. F-measure to Naïve Bayes classifier the value was = 0.965. Figure 4 shows the graphical representation for Table 4. It is noticed by the bars that represent results for Avg. F-measure to ImpCHI square have the superiority over the classified CHI Square. It is noticed by the bars that represent results for Avg. f-measure to ImpCHI square have the superiority over the classified CHI Square. It is clear that the better methods are the proposed ImpCHI without pre-processing, which got the four best results out of six (i.e., BN, NB, DT, and ANNs). The second best method is the proposed ImpCHI with pre-processing, it got the two best results out of six (i.e., NBM, and RF).

We compared the Classification Classifiers according to the Time measure: BN, NB, NBM, RF, DT and ANNs. By conducting six tests for each classifier: without pre-processing, with pre-processing, without pre-processing and CHI, with pre-processing and CHI, without pre-processing and ImpCHI and with pre-processing and impCHI. We show that when using ImpCHI square as feature selection method, gave worse results as gave worse results in Time build model. Table 5 shows the results based on Avg. Time. It is noticed by the bars that

Table 6 Enhancement Percentage on CHI (without pre-processing)

	BN	NB	NBM	RF	DT	ANNs
precision Enhancement	4.9%	4.4%	2.8%	3.3%	**15.5%**	6.1%
Recall Enhancement	5.8%	0.73%	2.2%	**18.3%**	3.2%	5.3%
F-measure Enhancement	4.3%	2.53%	2.8%	6.3%	**17.2%**	6.14%

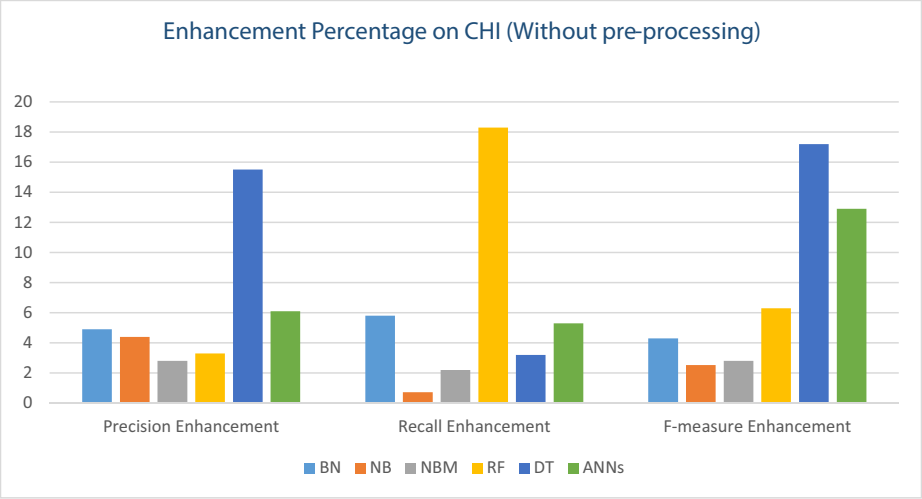

Fig. 6 Enhancement Percentage on CHI (without pre-processing)

represent results for Avg. time to ImpCHI square have the superiority over the classified CHI Square. It is clear that the better methods are the proposed ImpCHI without pre-processing, which got the four best results out of six (i.e., BN, NB, NBM, and RF). The second best method is the proposed ImpCHI with pre-processing, it got the two best results out of six (i.e., DT, and ANNs).

Whereas BN, NB, NBM, RF and DT they gave the worse results when using ImpCHI square without pre-processing. But ANNs they gave the worse results when using ImpCHI square with pre-processing. The less result was in Avg. Time to Random Forest classifier the value was = 94.3 s. Moreover, the proposed IimpCHI method with pre-processing got the best results in almost all the tested cases (i.e., it got four best cases of six). Figure 5 Shows the graphical representation for Table 5. It is noticed by the bars that represent results for Avg. F-measure to ImpCHI square have the superiority over the classified CHI Square.

After testing for ImpCHI square by classification algorithms: BN, NB, NBM, RF, DT and ANNs. We calculated enhancement percentage without pre-processing the results ranged from 0.75% to 18.3% as shown in Table 6. Where through precision enhancement percentage was the highest value for DT classifier it was = 15.5%, as for the Recall enhancement percentage was the highest value for RF classifier it was = 18.3%, as for the F-measure enhancement percentage was the highest value for DT classifier it was = 17.2%. Figure 6 shows the graphical representation for Table 6. It is noticed by the bars that represent results for Enhancement Percentage on CHI (without pre-processing). The enhancement ratio in terms of the Precision measure is the best using DT classifier. In terms of the Recall measure, the RF got the most enhancement ration compared to others methods. Finally, According to the F-

Table 7 Enhancement Percentage on CHI (With pre-processing)

	BN	NB	NBM	RF	DT	ANNs
Precision Enhancement	2.9%	**4.8%**	3.3%	2.13%	4.5%	2.9%
Recall Enhancement	1.3%	1.7%	2.11%	**7.8%**	0.4%	3.3%
F-measure Enhancement	2.1%	1.6%	**9.1%**	6.3%	3.8%	5.4%

measure, the DT obtained the best enhancement ration compared to other comparative methods. So, the DT classifier got the most suitable results in almost all test cases.

After testing for ImpCHI square by classification algorithms: BN, NB, NBM, RF, DT and ANNs. We calculated enhancement percentage with pre-processing the results ranged from 0.4% to 9.1% as shown in Table 7. Where through precision enhancement percentage was the highest value for NB classifier it was = 4.8%, as for the Recall enhancement percentage was the highest value for RF classifier it was = 7.8%, as for the F-measure enhancement percentage was the highest value for NBM classifier it was = 9.1%. Figure 7 shows the graphical representation for Table 6. It is noticed by the bars that represent results for Enhancement Percentage on CHI (with pre-processing). The enhancement ratio in terms of the Precision measure is the best using NB classifier. In terms of the Recall measure, the RF got the most enhancement ration compared to others methods. Finally, According to the F-measure, the NBM obtained the best enhancement ration compared to other comparative methods.

7 Conclusions and future work

In this paper, an impact on ImpCHI square is studied on Arabic Text Classifiers and the effect of the feature selection process using improved CHI square on the results of the classification process in terms of precision, Recall, F-measure and Time build the model. The test was carried out using dataset in Arabic, pre-processing and comparing the results of the classification with 6 tests: (1) Test without pre-processing, (2) Test with pre-processing, (3) Test without pre-processing and Using CHI Square, (4) Test with pre-processing and Using CHI Square, (5) Test without pre-processing and Using ICHI Square and (6) Test with pre-processing and Using ICHI Square.

After testing the classification algorithms, we compared them through four criteria: precision, Recall, F-measure and Time build the model. When comparing algorithms through Avg. precision was the best result for the Naïve Bayes classifier when we compared it through Avg.

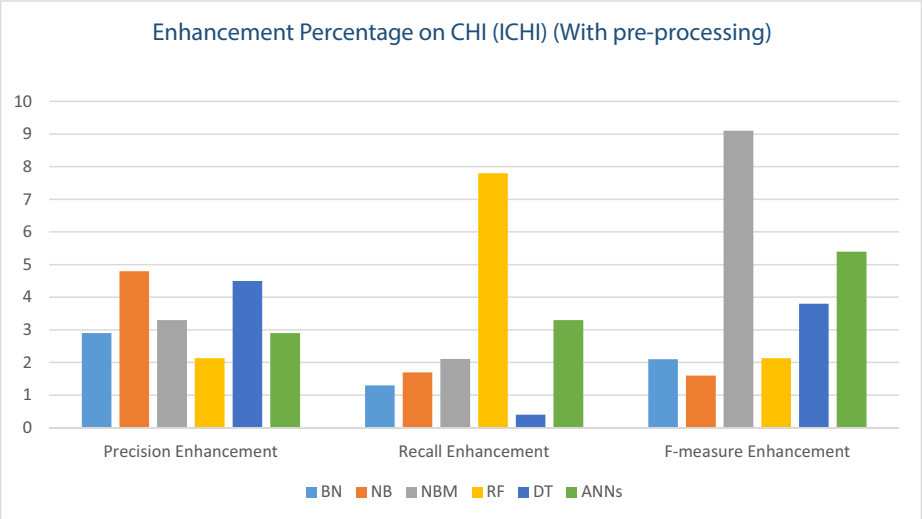

Fig. 7 Enhancement Percentage on CHI (with pre-processing)

Recall it was the best result for Naïve Bayes Multinomial classifier when we compared it through Avg. F-measure was the best result for the Naïve Bayes classifier. After comparison through Time build a model it was less time for Decision Tree classifier, so as that Naïve Bayes classifier the best results outweighed by two measures: precision and Recall, this means that the Naïve Bayes classifier is the best algorithm we've compared. We come to Enhancement Percentage when testing without pre-processing the results ranged from 0.75% to 18.3%. But when testing with pre-processing the results ranged from 0.4% to 9.1%. This shows us that the Enhancement Percentage was better in the case without pre-processing meaning without Tokenization, Normalization, Stop word removal, and Stemming.

In future work, you will test ImpCHI square on other types of learning such as unsupervised learning. We will try to test clustering algorithms and examine the impact of ImpCHI square as feature selection on clustering processes and other applications such as, text clustering, data classification, feature selection, gene expression, cancer classification image feature selection, etc. Furthermore, a new optimization algorithm can be employed to enhance the feature selection methods.

References

1. Abualigah L, Alfar HE, Shehab M, Hussein AMA (2020) Sentiment analysis in healthcare: a brief review. In: Recent advances in NLP:the case of arabic language. Springer, Cham, pp 129–141
2. Abualigah L, Alsalibi B, Shehab M, Alshinwan M, Khasawneh AM, Alabool H (2020) A parallel hybrid krill herd algorithm for feature selection. Int J Mach Learn Cybern:1–24
3. Abualigah L, Bashabsheh MQ, Alabool H, Shehab M (2020) Text summarization: a brief review. In: Recent advances in NLP: the case of arabic language. Springer, Cham, pp 1–15
4. Abualigah L, Diabat A, Geem ZW (2020) A comprehensive survey of the harmony search algorithm in clustering applications. Appl Sci 10(11):3827
5. Abualigah LMQ (2019) Feature selection and enhanced krill herd algorithm for text document clustering. Springer, Berlin, pp 1–165
6. Abualigah LMQ, Hanandeh ES (2015) Applying genetic algorithms to information retrieval using vector space model. Int J Comput Sci Eng Appl 5(1):19
7. Abualigah LM, Khader AT (2017) Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. J Supercomput 73(11):4773–4795
8. Abualigah LM, Khader AT, Al-Betar MA, Alomari OA (2017) Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. Expert Syst Appl 84:24–36
9. Abualigah LM, Khader AT, Hanandeh ES (2018) Hybrid clustering analysis using improved krill herd algorithm. Appl Intell 48(11):4047–4071
10. Abualigah LM, Khader AT, Hanandeh ES (2018) A new feature selection method to improve the document clustering using particle swarm optimization algorithm. J Comput Sci 25:456–466
11. Abualigah LM, Khader AT, Hanandeh ES (2018) A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. Eng Appl Artif Intell 73:111–125
12. Abualigah L, Shehab M, Diabat A, Abraham A (2020) Selection scheme sensitivity for a hybrid Salp swarm algorithm: analysis and applications. Eng Comput 1–27
13. Aliwy AH (2012) Tokenization as preprocessing for arabic tagging system. Int J Inform Educ Technol (IJET) 2(4):348
14. Alshaer H, Alzwahrah B, Otair M (2017) Arabic text classification using Bayes classifiers. Int J Inform Syst Comput Sci
15. Ayedh A, Tan G, Alwesabi K, Rajeh H (2016) The effect of preprocessing on arabic document categorization. Algorithms 9(2):27
16. Bahassine S, Madani A, Al-Sarem M, Kissi M (2020) Feature selection using an improved chi-square for Arabic text classification. J King Saud Univ Comp & Info Sci 32(2):225–231
17. Bahassine S, Madani A, Kissi M (2016) An improved chi-sqaure feature selection for Arabic text classification using decision tree. In 2016 11th international conference on intelligent systems: theories and applications (SITA), IEEE, pp. 1–5

18. Bawaneh MJ, Alkoffash MS, Al Rabea AI (2008) Arabic text classification using K-NN and naive Bayes. J Comput Sci 4(7):600–605
19. Chanod JP, Tapanainen P (1996) A non-deterministic tokeniser for finite-state parsing. In: Proceedings of the workshop on extended finite state models of language (ECAI'96)
20. Chen Y, He F, Li H, Zhang D, Wu Y (2020) A full migration BBO algorithm with enhanced population quality bounds for multimodal biomedical image registration. Appl Soft Comput:106335
21. Cutler D, Edwards C, Beard K, Cutler A, Hess K, Gibson J, Lawler J (2007) Random Forest for classification in ecology. Ecology 88:2783–2792
22. Gharib TF, Habib MB, Fayed ZT (2009) Arabic text classification using support vector machines. Int J Comput Their Appl 16(4):192–199
23. Hawashin B, Mansour A, Aljawarneh S (2013) An efficient feature selection method for Arabic text classification. Int J Comput Appl 83(17)
24. Hmeidi I, Al-Ayyoub M, Abdulla NA, Almodawar AA, Abooraig R, Mahyoub NA (2015) Automatic Arabic text categorization: A comprehensive comparative study. J Inf Sci 41(1):114–124
25. Jadon E, Sharma R (2017) Data mining: document classification using naive Bayes classifier. Int J Comput Appl 167(6):13–16
26. Kanan T, Fox EA (2016) Automated arabic text classification with P-S temmer, machine learning, and a tailored news article taxonomy. J Assoc Inf Sci Technol 67(11):2667–2683
27. McCallum A, Nigam K (1998) A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization 752(1):41–48
28. Moh'd A, Mesleh A (2007) Chi square feature extraction based SVMs arabic language text categorization system. J Comput Sci 3(6):430–435
29. Mesleh A (2011) Feature sub-set selection metrics for Arabic text classification. Pattern Recogn Lett 32: 1922–1929
30. Mohana R, Sumathi S (2014) Document classification using multinomial Naïve Bayesian classifier. Int J Sci Eng Technol Res(IJSETR) 3(5):1557–1563
31. Mohammad AH, Alwada'n T, Al-Momani O (2016) Arabic text categorization using support vector machine, Naïve Bayes and neural network. GSTF Journal on Computing (JoC) 5(1):108
32. Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J (2017) Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT) 48(3):128–138
33. Otair MA (2013) Comparative analysis of Arabic stemming algorithms. J Inf Technol Manag 5(2):1–13
34. Parekh R, Yang J, Honavar V (2000) Constructive neural-network learning algorithms for pattern classification. IEEE Trans Neural Netw 11:436–451
35. Patra A, Singh D (2013) Neural network approach for text classification using relevance factor as term weighing method. Int J Comput Appl 68(17):37–41
36. Raho G, Al-Shalabi R, Kanaan G, Nassar A (2015) Different classification algorithms based on Arabic text classification: feature selection comparative study. International Journal of Advanced Computer Science and Applications (IJACSA) 6(2):23–28
37. Saravanan K, Sasithra S (2014) Review on classification based on artificial neural networks. International Journal of Ambient Systems and Applications (IJASA) 2(4):11–18
38. Sembok TMT, Ata BA, Bakar ZA (2011) A rule-based Arabic stemming algorithm. Proceedings of the European Computing Conference, pp 392–397
39. Sharma D, Jain S (2015) Evaluation of stemming and stop word techniques on text classification problem. International Journal of Scientific Research in Computer Science and Engineering (IJSRCSE)) 3(2):1–4
40. Xu Q, Li M (2019) A new cluster computing technique for social media data analysis. Clust Comput 22(2): 2731–2738
41. Xu Q, Li M, Li M, Liu S (2018) Energy spectrum CT image detection based dimensionality reduction with phase congruency. J Med Syst 42(3):49
42. Xu Q, Wang Z, Wang F, Li J (2018) Thermal comfort research on human CT data modeling. Multimed Tools Appl 77(5):6311–6326
43. Xu Q, Li M, Yu M (2019) Learning to rank with relational graph and pointwise constraint for cross-modal retrieval. Soft Comput 23(19):9413–9427
44. Xu Q, Wang F, Gong Y, Wang Z, Zeng K, Li Q, Luo X (2019) A novel edge-oriented framework for saliency detection enhancement. Image Vis Comput 87:1–12
45. Zakariah M (2014) Classification of large datasets using random Forest algorithm in various applications: survey. International Journal of Engineering and Innovative Technology (IJJEIT) 4(3))

Affiliations

Hadeel N. Alshaer[1] · Mohammed A. Otair[1] · Laith Abualigah[1] · Mohammad Alshinwan[1] · Ahmad M. Khasawneh[1]

Hadeel N. Alshaer
HadeelAlshaer94@outlook.com

Mohammed A. Otair
Otair@aau.edu.jo

Mohammad Alshinwan
mohmdsh@aau.edu.jo

Ahmad M. Khasawneh
a.khasawneh@aau.edu.jo

[1] Faculty of Computer Sciences and Informatics, Amman Arab University, Amman, Jordan