Modern Applied Science; Vol. 13, No. 5; 2019
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Arabic Text Classification: A Review

Adel Hamdan Mohammad[1]

[1] Computer Science Department, The world Islamic Sciences and Education University, Amman, Jordan

Correspondence: Adel Hamdan Mohammad, Computer Science Department, The world Islamic Sciences and Education University, Amman, Jordan. E-mail: Adel.hamdan@wise.edu.jo/Adel_hamdan@yahoo.com

Received: February 2, 2019 Accepted: April 13, 2019 Online Published: April 30, 2019

doi:10.5539/mas.v13n5p88 URL: https://doi.org/10.5539/mas.v13n5p88

Abstract

Text classification is a significant topic. The number of electronic documents available on line is massive. Text classification aims to categorise documents into a set of predefined classes. Number of researches conducted on English dataset is great in comparison with number of researches done using Arabic dataset. This research could be considered as reference for most researchers who deal with Arabic dataset. This research used the most well-known algorithms used in text classification with Arabic dataset. Besides that, dataset used in this research is large enough in comparison with most dataset for Arabic language used in other researches. In addition, this research used different selections and weighting methods for documents. I expect that all researchers who would write researches using Arabic dataset will find this work helpful. Algorithms used in this research are naïve Bayesian, support vector machines, artificial neural networks, k- nearest neighbors, C4.5 decision tree and rocchio classifier.

Keywords: text classification, arabic dataset, naïve bayesian, support vector machines, artificial neural networks, k-nearest neighbors, decision tree, rocchio classifier

1. Introduction

No doubt that the massive number of available electronic documents make text classification (TC) one of the most critical topics. The huge number of available electronic documents' forms such as research article, reports, conference papers and other forms will increase the needs for efficient text classification methods and techniques. TC aims at extracting useful and valuable information from a huge document and this valuable information is used later in identifying document content based into a set of predefined conditions. Certainly not all available documents are valuable, but the number of useful and valuable articles are not small. (Motaz,2009; Adel Hamdan,2011) Text classification is a very important topic for information retrieval. Searching for new methods and techniques for text classification is one of the main duties for almost most academic and industry researchers. Text classification try to classify a new document into a set of predefined documents based on document content. Besides that, spam detection uses several techniques for email classification. Some of these approaches are based on email content. (Adel Hamdan,2011; Raed Abu Zitar,2011; Adel Hamdan,2013) Text classification is not an easy process since sometimes there are a great number of available information in document. Besides that, this information may have a high diversity. (Mofleh, 2012; Sebastiani,2002) One of the goals of text classification is to extract useful information from a huge document and then use this extracted knowledge to identify the document based into a set of predefined documents. (Rasha,2015; Mohammad Ali,2010)

Most of text classification researches used English dataset. Actually, the number of researches conducted on English dataset are excellent. Unfortunately, the number of researches using Arabic dataset still need more investigation. There are several text classification methods such Naïve Bayesian (NB), Support vector machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (K-NN), Decision Trees (C4.5), Hidden Markov Model (HMM), Rocchio Classifier and other methods. (Sameh,2013; Adel Hamdan,2018; Abdel-Salam,2013)

A huge number of researches can be found in English dataset text classification. Actually, most of these approaches are applied several times with several researches using English dataset. (L.Borrajo,2015; Adel Hamdan,2016; Adel Hamdan, 2018) But unfortunately, the number of researches and experiments done using Arabic dataset still not enough. In this research the author applies the most well-known text classification methods and applies his experiments using Arabic dataset. Also, in this research the author uses his own built in (in-house) dataset. ctaAuthor aims at creating a good reference for all researches who uses Arabic dataset. Besides that, in this research author uses a huge built in dataset. Finally, author use different feature extraction methods.

The rest of this paper is organized as follows: the following section contains details of Naïve Bayes. Section 3 Support vector machine will be explained. Section 4 Artificial neural networks as text classifier will be explained. Section 5 K-Nearest neighbor text classifier will be shown. Section 6 will show Decision Trees classifier. Section 7 will talk about rocchio classifier. Section 8 will talk about Arabic language. Section 9 demonstrate feature selection and extraction. Section 10 will talk about related studies. Section 11 will talk about dataset used in this research. Section 12 our experiments will be shown in detail. Finally, section 13 our conclusion will be shown.

2. Naïve Bayesian

Naïve Bayesian is a machine learning approaches used in text classification. Naïve Bayes is a probabilistic classifier based on applying Naïve Bayes theorem. (Abdel-Salam, 2013; L.Borrajo,2015; Adel Hamdan, 2016; Adel Hamdan,2018) Naïve Bayes could be considered a simple and great method in text categorization. One of the main advantages of Naïve Bayes is that it is a highly scalable method. (Russell, 2003) Naïve Bayes method assumes that the any value of a specific feature is independent of any value of other feature. (Russell,2003; Saleh Alsaleem, 2011; Rish, 2001) in Naïve Bayes there is a training stage so Naïve Bayes is a supervised learning technique. (Saleh,2011; Vladimir,1995)

Naïve Bayes could be considered one of the conditional probability models. In Naïve Bayes the probability that a given document D belongs to a given class C is calculated as follows: (Adel Hamdan, 2018; Adel Hamdan, 2016; Russell, 2003; Saleh Alsaleem,2011; Vladimir, 1995)

Prob (class | document) = Prob (class). Prob (document | class) / Prob (document)	(1)

Prob (class | document): The probability that a given document D belongs to a given class C.

Prob (document): The probability of a document.

Prob (class): The probability of a class.

Prob (document | class): The probability of document given class

In other words, bayes theorem algorithm provides a formula for calculation the posterior probability, Prob (C|X), from Prob P(C), P(X), and P(X|C). in Bayes theorem algorithm the effect of any value of a predictor (feature) (X) on a given class(C) is independent of the values of other predictor (feature).

3. Support Vector Machine

Support vector machines (SVMs) are considered one of the most well-known text classifiers. SVMs are one of the supervised machine learning techniques. In SVMs a training algorithm is used to build a model that will be used to assign a new unknown document to one category from a set of predefined categories. SVMs can be used to perform a linear and a non-linear classification (see figure 1). Besides that, SVMs can be used with supervised and unsupervised learning. (Adel Hamdan,2016; Thorsten Joachims,1998; Cristianini,2000)

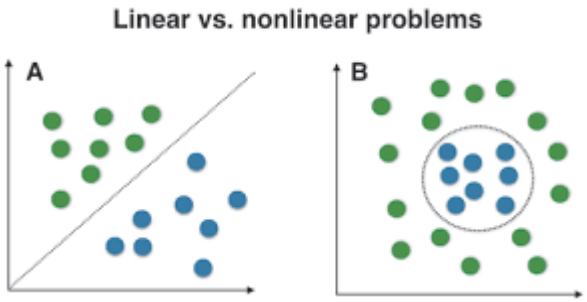

Figure 1. Linear SVM vs Nonlinear SVM

SVMs creates a hyperplane or a set of hyperplanes and these hyperplanes are used for classification or regression. In SVMs the classes are a hyperplane of the form:

$W^T X + b = 0$	(2)

W: Weight of the vector.

X: Input Vector.

b: bias.

In SVMs, for each vector X_i we have either:

W. X_i+ b >=1 for X_i having class 1 or	(3)
W. X_i+ b <=1 for X_i having class -1	(4)

The boundary between positive and negative values is called decision boundary. If we got w.x+b=0 this means we get the decision boundary.

4. Artificial Neural Networks

Artificial Neural Networks (ANNs) are one of the main tools used in machine learning. The concept of ANNs are inspired from biological human brains. In ANNs the system learns how to perform tasks after training stage. ANNs are available in different forms and shapes such as supervised and unsupervised learning. When ANNs are introduced the aim is to solve problems in the same way the human brain solves it. ANNs are used in several areas such as text classification, pattern classification, speech recognition, medical diagnosis, prediction, financial analysis, optimization, and others. (Rosenblatt,1958; M. Caudill,1992; Adel Hamdan,2016).

ANNs can be found in different shapes such as single layer perceptron, radial basis network (RBN), multi-layer perceptron (MLP) (see figure 2) (dtreg.com) and others. The simplest form is a network which consists of an input and an output. The simplest form was introduced by Rosenblatt (1958). (Fouzi,2010; D. Rumelhart,1986)

Multi-Layer Perceptron (MLP) is feed forward neural network. In MLP there are one or more hidden layers. MLP is passing input to output through hidden layers. Number of hidden layers and input can affect directly the output. MLP consists of three layers at least. Input layer, output layer and hidden layer. MLP use supervised learning techniques with backpropagation for training phase. Besides that, MLP can be used to separate data that not linearly separated. In this research author uses MLP. (M. Caudill,1992; Fouzi Harrag2010; D. Rumelhart, 1986)

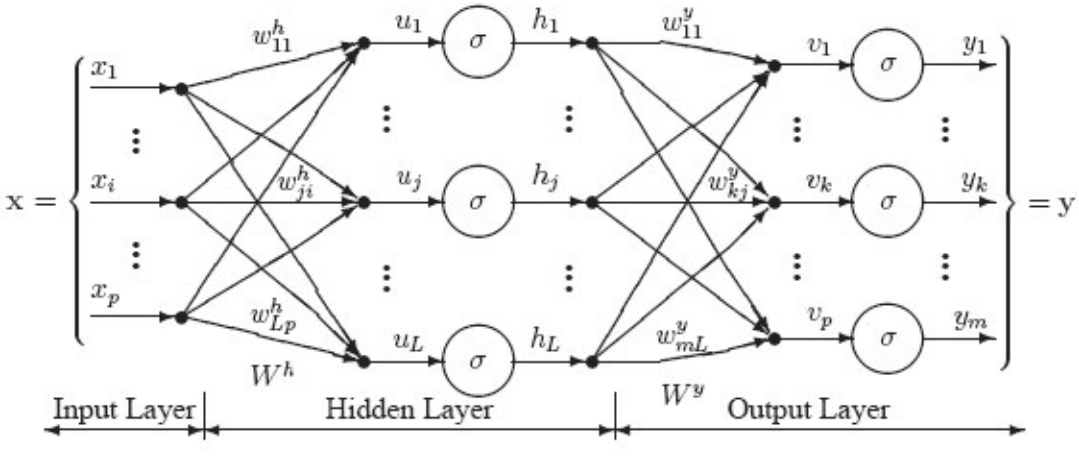

Figure 2. Multilayer Perceptron Neural Network Model

In MLP, if we have m input data (x_1, x_2,…,x_m), we call this m features then we multiply each of m features with a weight w (w_1, w_2,…, w_m) and then perform sum of them (dot product) as follows:

W. X=$w_1.x_1$+$w_2.x_2$+…+w_m. x_m=$\sum_{i=1}^{m} WiXi$	(5)
Z= $\sum_{i=1}^{m} WiXi + bias$	(6)
y'=f(z) (activation function)	(7)

5. K-Nearest Neighbor

K-Nearest neighbor (K-NN) is a statistical learning approach used in text classification. KNN is one of the simplest algorithms which used in machine learning. (Y. Yang,1999; Riyad Al-Shalabi,2006). Besides that, K-NN considered one of the top text classification methods. In K-NN there is a training phase and a testing phase. The

training phase consists of storing feature vectors and class labels of the training stage. In K-NN a commonly used distance metric for continuous variables is Euclidean distance. In K-NN when the algorithm tests a document the algorithm tries to find k nearest neighbor among the training documents. K-NN algorithm is based on feature similarity which means that documents are classified based on how it similar to its neighbors. K-NN has several merits such simple algorithm and high accuracy. Also, K-NN has several disadvantages such as computationally expensive, high memory requirements and prediction stage might be slow with big N. (Gongde Guo,2006; Li Baoli,2003; XindongWu,2007)

This research use cosine similarity to calculate similarity between documents. In cosine similarity if we assume A and B are two vectors representing documents J and K then the similarity is calculated as follows: (Gongde,2006; Li Baoli ,2003; XindongWu,2007)

SIM (A, B) $=\frac{\sum_{i=1}^{r} W_{ij} * W_{ik}}{\sqrt{\sum_{i=1}^{r} W_{ij}^2} * \sqrt{\sum_{i=1}^{r} W_{ik}^2}}$	(8)

6. Decision Tree (C4.5)

Decision tree is a predictive model used in learning. The main goal of decision tree is to create a model that has the capability to predict the value of target variables. In decision tree the variables created in training phase are used to predict target variables. A decision tree is one of simplest representation used in classification. Decision tree applies simple and straightforward ideas to solve classification problems. In general decision tree is built from a set of attributes. Decision tree has several forms such as ID3 and C4.5. (Abdullah,2012; Nidhi,2011; Badr,2014; Edy Budiman,2018)

C4.5 is one of the most successful classifiers in data mining. C4.5 is a statistical classifier. (XindongWu,2007; Abdullah, 2012). C4.5 builds decision tree from a set of training data. C4.5 uses gain ration and information gain to rank possible tests. Basically C4.5 consists of four steps: (XindongWu,2007; Edy Budiman,2018).

I. Choose attribute as a root.

II. Generate branch every value.

III. Put dataset in branch.

IV. Repeat the second process until every class have the same value.

Formulas used in C4.5 is shown bellows:

Entropy (S)$\sum_{i=1}^{n} -Pi * \log_2 pi$	(9)

S: Entropy.

P: class proportion in the output.

Gain (S,A)= Entropy (S) $\sum_{i=1}^{n} \frac{\lvert Si \rvert}{\lvert S \rvert} * Entropy\ (S)$	(10)

S: Set of case.

A: Attribute of case.

|Si|: number of cases to i.

|S|: number of cases in the set.

7. Rocchio Classifier

Rocchio classifier is an information retrieval algorithm. Rocchio classifier is a linear classifier. Rocchio methods is based on relevance feedback. Rocchio classifier developed based on Vector Space Model. Rocchio classifier works as follows: Dataset is divided into two main categories training data and test data. Rocchio Classifier algorithm gives a benchmark dataset. Besides that, all documents or samples in the initial classes must be labeled. Rocchio classifier selects a class C_i and W_{Ci} is the representative vector. In the training stage every sample must have its own weight vectors (W_{Ci}). (Anping,2011; F. Sebastiani ,2002; F. Sebastiani,1999; M. M. Syiam,2006; Gongde Guo,2006).

in rocchio classifier, classifying new instance requires computing the inner product between the new instance and the generalized instances.

In rocchio classifier, given a training dataset Tr, this means that in straight method we can compute a classifier Ci

(W_{i1},W_{i2},….W_{in}) for category Ci. Finally, the weighted average of a group is computed as follow: (Gongde Guo,2006; Tarek Fouad,2009]

$$W_{ik} = \beta \sum_{dj \in POSi} \frac{Wjk}{|POSi|} - \gamma . \sum_{dj \in NEGi} \frac{Wjk}{|NEGi|} \quad (11)$$

W_{ik}: the weight of term t_k in document d_j.

POS_i and $NEG_{i:}$ document d_j belonging to (or not belonging to) category ci.

β, γ: control parameters that allow setting of positive and negative examples.

8. Arabic Language

Arabic language is spoken by more than 250 million. Letters of Arabic language consist of 28 letters plus hamza. Arabic language letters are written from right to left. One of the main characteristics of Arabic language is that its letters has different forms and shapes depending on the position of the letter.one of the excellent merits of Arabic language is that majority of Arabic word has a root. Besides that, majority of Arabic root words are consisting of three letters. Representing words with its root helps in reducing the number of words. (Rehab Duwairi,2005; Eldos,2003; Adel Hamdan,2016; Adel Hamdan,2019).

9. Feature Extraction, Selection and Terms Weighting

Cleaning the documents from unnecessary characters is essential in text classification. Cleaning a document means removing all prepositions, auxiliary verbs, special characters, tags and others useless letters. One of the main advantages of Arabic language is that Arabic language has its built-in filtering instruments. Data preprocessing is critical in text classification. Data preprocessing aims to reduce the size and number of words in documents. Cleaning documents include two main steps feature extraction and feature selection (Rehab,2007; Liu,1998).

Feature extraction is used to clean documents from all unnecessary words and letters. Feature extraction aims at reducing the amount of data required to describe a huge document. (Liu,1998; Wang,2005) In other words feature extraction aims at reducing the amount of input data by refining its representative attributes. Besides that, we can say that feature extraction means deriving new feature from its original feature. One of the significant goals of feature extraction is that keeping only key words which have the highest score or values according to a set of predefined criteria. (Zi-Qiang,2005; Montanes,2003)

Feature selection means selecting only useful subset form extracted features. No doubt that not all extracted data from feature extraction are useful. Selecting only subset of extracted feature is done in feature selection.

Feature selection has three different forms which are filter methods, wrapper methods and embedded methods. (L. Ladha,2013; Zaghoul,2013; Franca,2003) Filter methods apply statistical procedures to assign score for each feature. One of the main examples of filter methods is Chi-Square method, information gain and correlation coefficient scores. In wrapper methods the subset selection is based on learning algorithm used to train the model. Example of wrapper methods is recursive feature elimination. Finally, embedded methods which combines filter and wrapper methods. One popular example on embedded method is LASSO (Least Absolute Shrinkage and Selection Operator). (L. Ladha, 2013; F. Sebastiani, 2002; F. Sebastiani1999; Johannes Furnkranz, 1998; Abu-Errub, 2014)

The final stage of documents cleaning is that each document must be weighted as a vector. Weighting is crucial for the next stages. There are many methods and algorithms for weighting documents vector such as term frequency method, inverse document frequency, normalized term frequency inverse document frequency, term frequency inverse document frequency and others. (F. Sebastiani,2002; F. Sebastiani1999; Johannes Furnkranz,1998; Abu-Errub,2014)

10. Related Studies

In this part author is going to talk about only researchers who conducted their researches using Arabic dataset. No doubt that the number of researches done using Arabic dataset is very small in comparison with English dataset, but there are a few researchers who perform an accepted experiment with Arabic dataset.

Adel Hamdan (Adel Hamdan & Tariq Alwada'n, 2016; Adel Hamdan & Omar, 2016). In these researches, authors perform several experiments using Naïve Bayesian, K-Nearest neighbours, Neural Network, Rocchio classifier, C4.5, and Support vector machine. The experiments done in theses researches are conducted with Arabic dataset. Experiments in this research are done with small dataset in comparison to this research. Besides that, experiments in (Adel Hamdan & Tariq Alwada'n, 2016; Adel Hamdan & Omar, 2016) are done using one weighting technique

(tf.idf). Experiments conducted show that those methods are performed well, and the results show an acceptable precision and recall measures.

Aisha Adel (Aisha, 2014). In this research authors make a comparative study of combine feature selection with Arabic dataset. This research investigates the performance of five feature selections named chi-square, correlation, Galavotti-Sebastiani-Simi (GSS) coefficient, information gain and relief F. experiments conducted using Naïve Bayesian and support vector machines. Experiments show that combination of multiple feature selection gives better results than using only one method.

Wail Hamood (Wail Hamood,2014). In this research author apply K-NN and naïve Bayesian classifiers. This research is also conducted using Arabic dataset. Experiments in this research are done using Weka Toolkit. Authors use precision and recall criterion. Experiments show that improved K-NN improves performance and accuracy. Besides that, the experiments show that naïve Bayesian gives the best accuracy.

Riyal Al-Shalbi (Riyad Al-Shalabi,2006). In this research authors implement K-NN with Arabic dataset. In this research authors use document frequency as a method for feature extraction and selection. Authors reach 0.95 micro average precision and recall. Dataset used in this experiment consist of 621 documents which belongs to six categories.

Tarek Fouad (Tarek Fouad,2009). In this research authors applied support vector machines classifier with Arabic dataset. Authors in this research make a comparison with other classifiers such as naive Bayesian, K-NN and rocchio classifier. Dataset in this research consists of 1132 documents. In this research the best results appear with rocchio classifier when the size of feature set is small. Besides that, SVM gives the best results when the size of feature set is large.

M. Syiam (M. M. Syiam, 2006). In this research authors use k-nearest neighbor and rocchio classifier as text classifiers. Self-collected dataset is used in this research. In this research many algorithms for stemming and feature selection are used. Results in this research show that hybrid approach of document frequency and information gain gives the best results. Besides that, authors mention that rocchio classifier has advantages in classification process over k-nearest neighbor. Authors in this research recommend using statistical n-gram stemmers for document preprocessing, information gain for feature selection and normalized tf.idf for term weighting.

Majed Ismail (Majed,2011). In this research authors use several algorithms to implement text classification. This research uses the Sequential Minimal Optimization (SMO), J48 (C4.5) and Naïve Bayesian (NB). In this research authors use Weka program. Experiments in this research show that SMO achieves the best accuracy. Besides that, results show that the time needed to build SMO model is the best.

Al-Harbi (Al-Harbi,2008). In this research authors present SVM and C5.0 classifiers. Dataset used in this research consists from seven corpora. Chi-squared is applied in this research. Results show that C5.0 outperformed the SVM algorithm.

Aymen Abu-Errub (Aymen,2014). In this research author proposed a method for Arabic text classification. In this research a document is compared with a predefined document category. The content of the document is used to identify it. Arabic dataset is sed in this research. Term frequency inverse document frequency and chi-square are used in this research.

Wail Hamood (Wail Hamood,2014). In this research authors implement traditional, proposed K-NN and Naïve Bayesian classifiers. Arabic dataset is used in this research. Results show that improved K-NN improves the accuracy performance. Besides that, experiments in this research show that NB has the best accuracy.

11. Dataset

One of the main problems of text classification for both English and Arabic language in general is lacking the availability of general dataset which can be used as benchmark. There is no general Arabic dataset which can be used by different authors as benchmark. Most of Arabic text classification researchers build their own dataset. In this research author used his own dataset which he used in several researches about Arabic text classification. Authors update and enlarge dataset used in this research to be consisted of 4000 documents. Dataset in this research are collected from several web resources such as Aljazeera site (www.aljazeera.net), al-hayat site (www.alhayat.com), and Saudi press agency. (www.spa.gov.sa) see table (1).

Table 1. (in-house) data set

Category	Total Number of Doc.	Training. (#of doc)	Testing. (# of doc)

Politics	500	300	200
Economics	500	300	200
Culture	500	300	200
Sports	500	300	200
Art	500	300	200
Technology	500	300	200
Science	500	300	200
Education	500	300	200
Total	4000	2400	1600

12. Experiments and Analysis

Author in this research applies several algorithms used in text classification. Documents in these experiments belong to 8 categories. Documents are collected from several resources as mentioned in section 11. The most familiar evaluation measures used in text classification are precision, recall. To understand these measures, see table 2 and formulas 12,13.

Table 2. The symbols that are required to calculate the three evaluation measures

Iteration	Relevant Document	Irrelevant Document
# of retrieved Documents	a	b
# of un-retrieved Documents	c	d

$$\text{Precision}= \frac{a}{a+b} \tag{12}$$

$$\text{Recall}= \frac{a}{a+c} \tag{13}$$

a: the relevant retrieved documents

b: the irrelevant retrieved document.

c: the relevant documents that haven't been retrieved

d: the irrelevant document that haven't been retrieved.

In this research author uses information gain, gain ratio and Chi-square as feature selection methods. Also, in this research authors use different number of input layers when dealing with multi-layer perceptron (MLP). Author applies 200,400,600,800 and 1000 input layers. The best results are shown with 800 input layers, so author demonstrates only MLP with 800 input layers. Related to K-NN, the value of K is very important and affects the results directly. Author applies several experiments using different values of K. In these experiments when K=12, best results are appearing.

Tables 3 and 4 demonstrate the author's results using information gain as a feature selection, then tables 5 and 6 demonstrate results using gain ratio and finally tables 7 and 8 show experiment results using chi-square.

Table 3. NB, SVM and ANN (Information Gain)

Classifier	Naïve		SVM		ANN	
Categories	Precision	Recall	Precision	Recall	Precision	Recall
Politics	0.82	0.78	0.67	0.71	0.69	0.51
Economics	0.81	0.79	0.69	0.82	0.68	0.57
Culture	0.79	0.79	0.69	0.81	0.64	0.51
Sports	0.68	0.76	0.79	0.49	0.69	0.58
Art	0.79	0.81	0.6	0.59	0.64	0.67
Technology	0.58	0.69	0.65	0.58	0.58	0.71
Science	0.79	0.68	0.69	0.57	0.6	0.79
Education	0.85	0.68	0.54	0.72	0.54	0.52
Average	0.76375	0.7475	0.665	0.6613	0.6325	0.6075

Table 4. NB, SVM and ANN (Information Gain)

Classifier	K-Nearest Neighbours		Decision Tree (C4.5)		Rocchio Classifier	
Categories	Precision	Recall	Precision	Recall	Precision	Recall
Politics	0.79	0.78	0.78	0.67	0.69	0.68
Economics	0.79	0.79	0.71	0.62	0.66	0.78
Culture	0.75	0.71	0.71	0.62	0.71	0.71
Sports	0.65	0.71	0.65	0.78	0.65	0.74
Art	0.52	0.68	0.65	0.61	0.79	0.75
Technology	0.63	0.64	0.69	0.61	0.71	0.75
Science	0.59	0.62	0.79	0.79	0.79	0.79
Education	0.53	0.63	0.79	0.69	0.81	0.61
Average	0.65625	0.695	0.72125	0.67375	0.72625	0.72625

Table 5. NB, SVM and ANN (Gain Ratio)

Classifier	Naïve		SVM		ANN	
Categories	Precision	Recall	Precision	Recall	Precision	Recall
Politics	0.87	0.81	0.77	0.79	0.71	0.54
Economics	0.84	0.89	0.78	0.85	0.69	0.59
Culture	0.79	0.87	0.79	0.91	0.65	0.52
Sports	0.69	0.81	0.85	0.59	0.72	0.59
Art	0.89	0.79	0.59	0.58	0.65	0.68
Technology	0.59	0.72	0.68	0.59	0.59	0.78
Science	0.89	0.72	0.78	0.57	0.58	0.81
Education	0.91	0.69	0.59	0.78	0.57	0.59
Average	0.80875	0.7875	0.72875	0.7075	0.645	0.6375

Table 6. K-NNN, C4.5 and Rocchio (Gain Ratio)

Classifier	K-Nearest Neighbours		Decision Tree (C4.5)		Rocchio Classifier	
Categories	Precision	Recall	Precision	Recall	Precision	Recall
Politics	0.84	0.81	0.79	0.69	0.78	0.69
Economics	0.82	0.82	0.78	0.68	0.77	0.78
Culture	0.81	0.76	0.77	0.66	0.75	0.72
Sports	0.69	0.72	0.69	0.87	0.69	0.75
Art	0.67	0.69	0.68	0.64	0.8	0.78
Technology	0.65	0.68	0.79	0.71	0.81	0.79
Science	0.61	0.68	0.81	0.8	0.82	0.88
Education	0.64	0.65	0.82	0.71	0.84	0.67
Average	0.71625	0.72625	0.76625	0.72	0.7825	0.7575

Table 7. NB, SVM and ANN (Chi-Square)

Classifier	Naïve		SVM		ANN	
Categories	Precision	Recall	Precision	Recall	Precision	Recall
Politics	0.89	0.82	0.78	0.81	0.71	0.55
Economics	0.85	0.91	0.79	0.87	0.71	0.6
Culture	0.81	0.91	0.81	0.9	0.57	0.52
Sports	0.71	0.84	0.86	0.61	0.74	0.61
Art	0.91	0.81	0.61	0.61	0.69	0.69
Technology	0.61	0.73	0.69	0.62	0.61	0.88
Science	0.91	0.73	0.81	0.59	0.55	0.83
Education	0.9	0.71	0.61	0.81	0.59	0.63
Average	0.82375	0.8075	0.745	0.7275	0.64625	0.6638

Table 8. K-NNN, C4.5 and Rocchio (Chi-Square)

Classifier	K-Nearest Neighbours		Decision Tree (C4.5)		Rocchio Classifier	
Categories	Precision	Recall	Precision	Recall	Precision	Recall
Politics	0.85	0.82	0.81	0.71	0.81	0.71
Economics	0.85	0.84	0.81	0.71	0.81	0.81
Culture	0.84	0.79	0.82	0.68	0.79	0.74
Sports	0.71	0.75	0.71	0.89	0.71	0.79
Art	0.59	0.71	0.71	0.68	0.83	0.79
Technology	0.59	0.74	0.81	0.76	0.84	0.81
Science	0.69	0.71	0.84	0.85	0.89	0.87
Education	0.78	0.59	0.85	0.76	0.86	0.71
Average	0.7375	0.74375	0.795	0.755	0.8175	0.77875

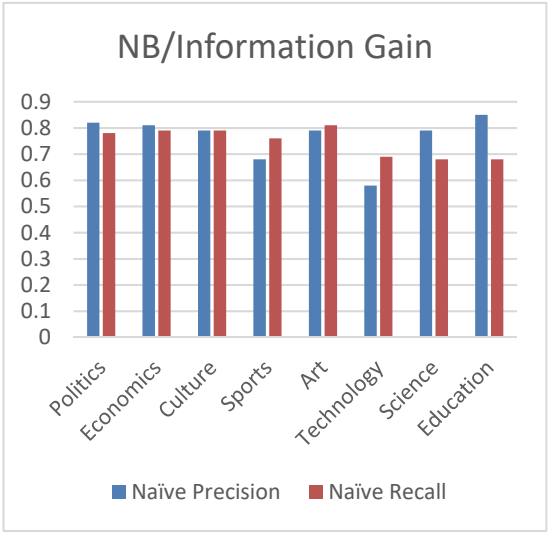

Figure 3. NB precision and recall using information gain

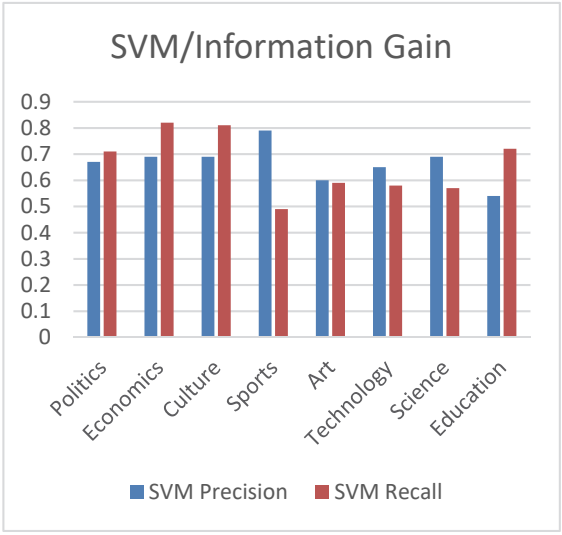

Figure 4. SVM precision and recall using information gain

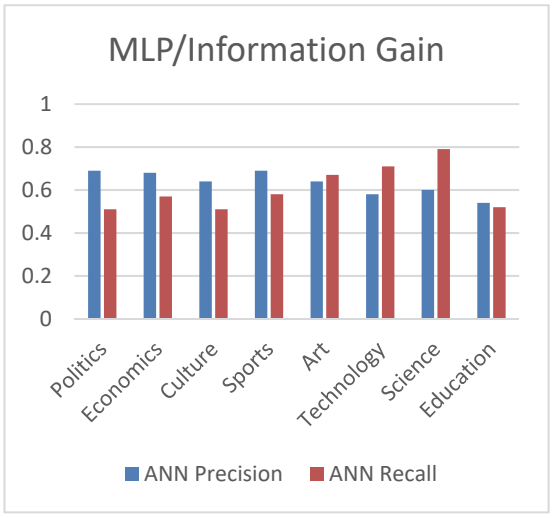

Figure 5. MLP precision and recall using information gain

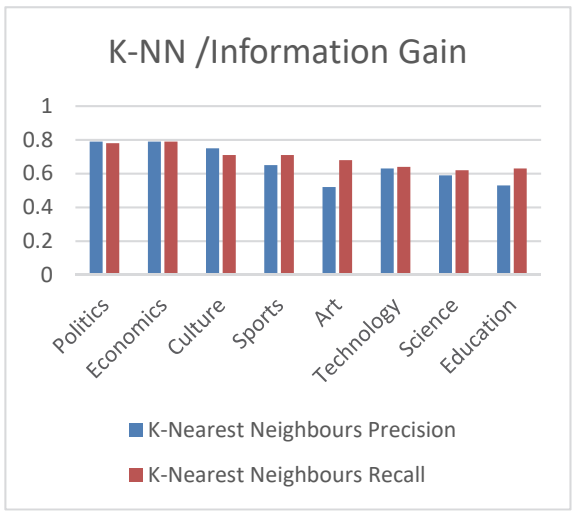

Figure 6. K-Nearest Neighbours precision and recall using information gain

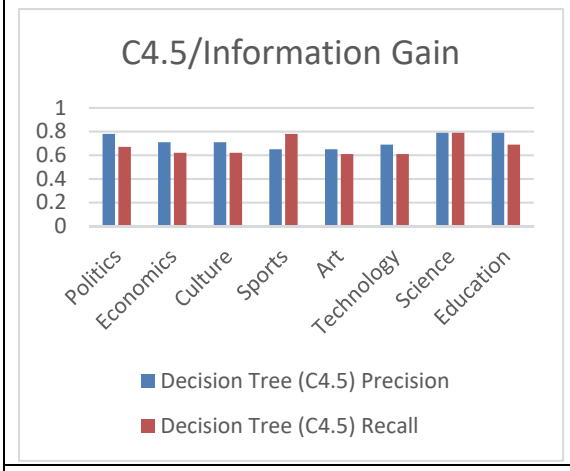

Figure 7. C4.5 precision and recall using information gain

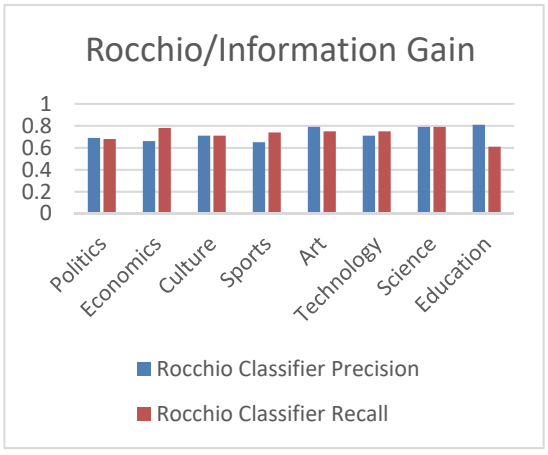

Figure 8. Rocchio precision and recall using information gain

Figures 3,4,5,6,7,8 and tables 3,4 show precision and recall when experiments done using information gain as feature selection. experiments show that NB gives the best results with 0.76 precision and 0.74 recall.

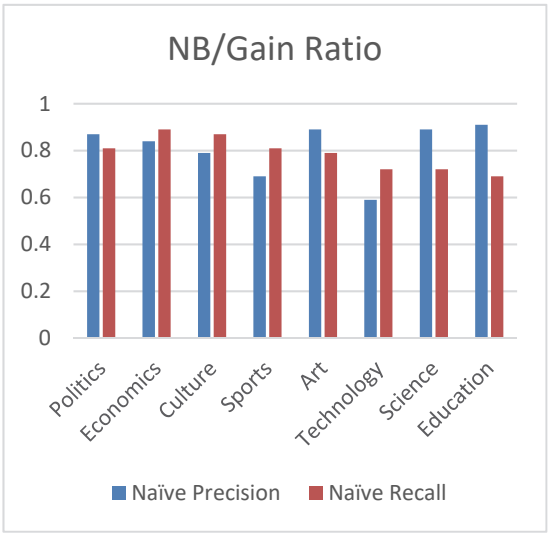

Figure 9. NB precision and recall using Gain Ratio

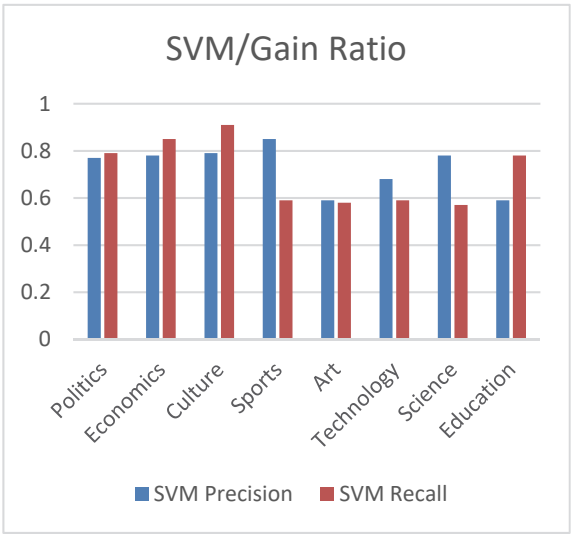

Figure 10. SVM precision and recall using Gain Ratio

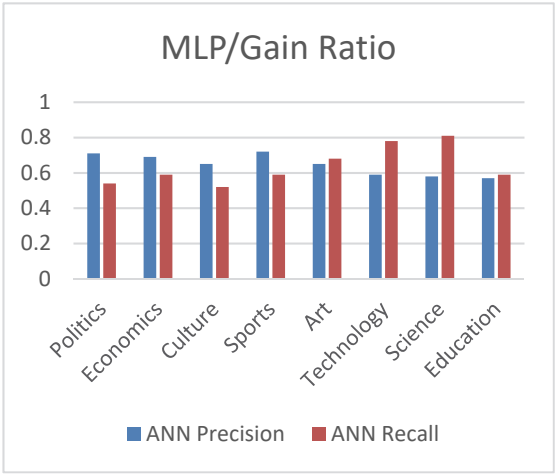

Figure 11. MLP precision and recall using Gain Ratio

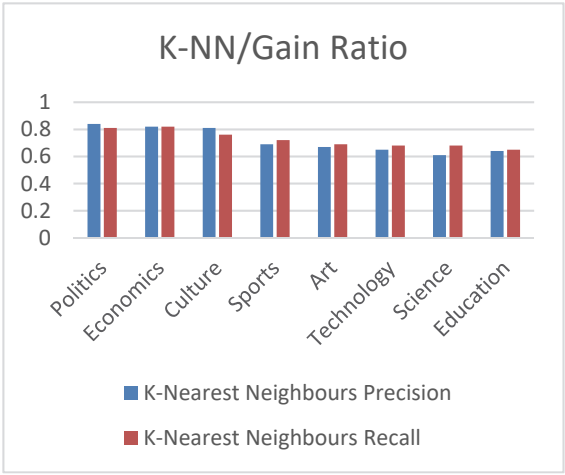

Figure 12. K-NN precision and recall using Gain Ratio

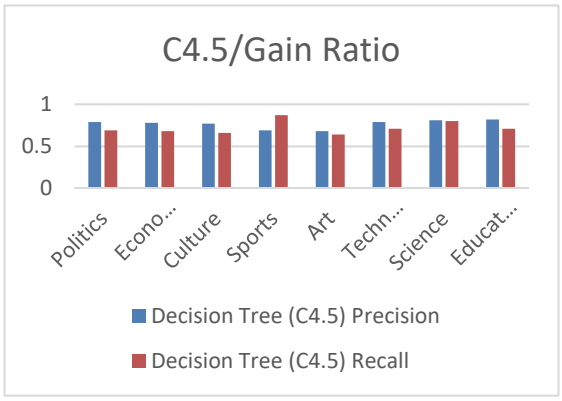

Figure 13. C4.5 precision and recall using Gain Ratio

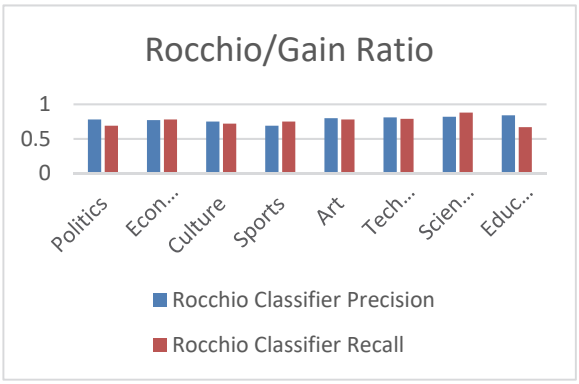

Figure 14. Rocchio precision and recall using Gain Ratio

Figures 9,10,11,12,13,14 and tables 5,6 show precision and recall when experiments done using gain ratio as feature selection. experiments demonstrate that NB gives the best results with 0.80 precision and 0.78 recall.

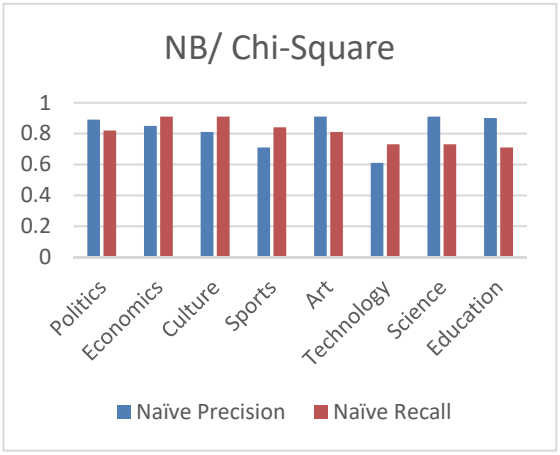

Figure 15. NB precision and recall using Chi-Square

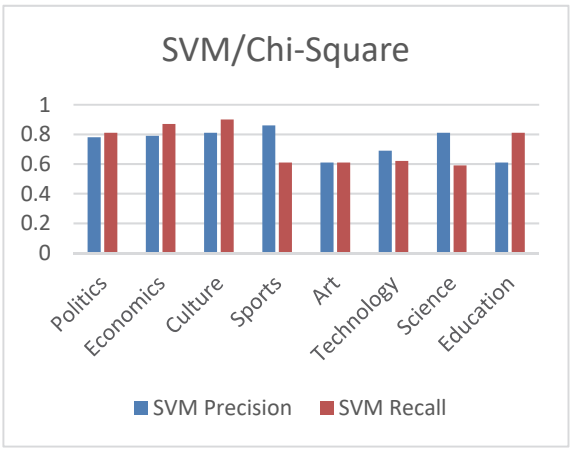

Figure 15. SVM precision and recall using Chi-Square

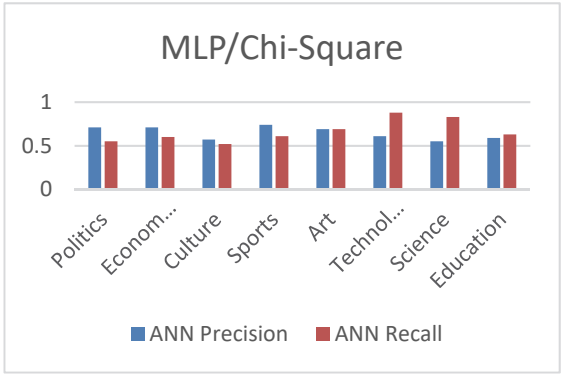

Figure 16. MLP precision and recall using Chi-Square

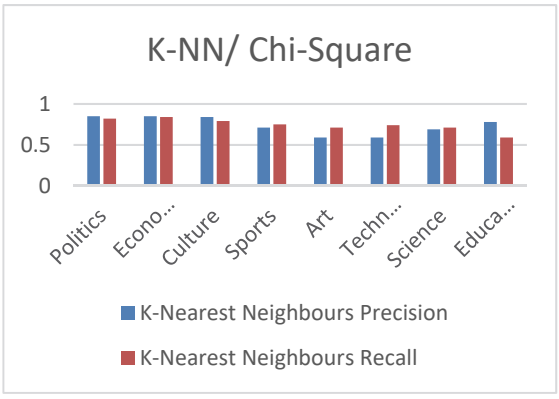

Figure 17. KNN precision and recall using Chi-Square

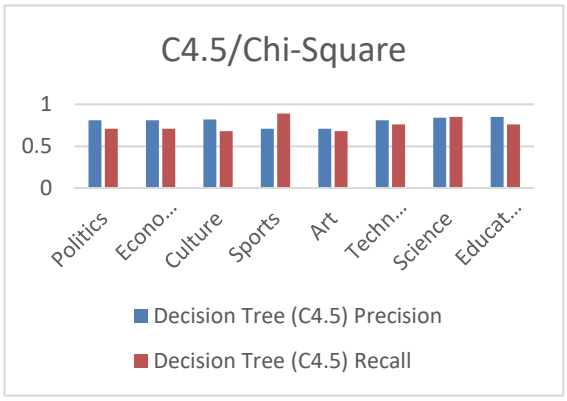

Figure 18. C4.5 precision and recall using Chi-Square

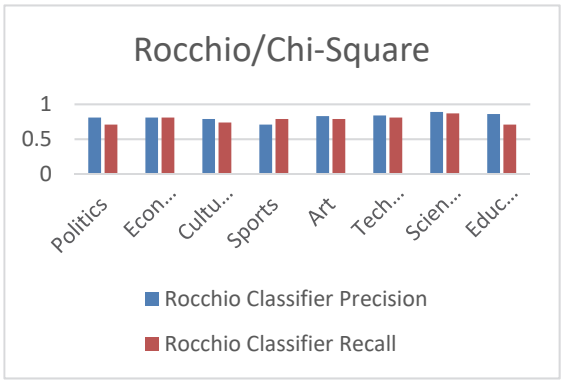

Figure 19. Rocchio precision and recall using Chi-Square

Figures 15,16,17,18,19,20 and tables 7,8 show precision and recall when experiments done using Chi-square as feature selection. experiments demonstrate that NB gives the best results with 0.82 precision and 0.80 recall.

Tables 3,4,5,6,7 and 8 shows that when using chi-square as feature selection results are little bit better than gain ration and information gain.

13. Conclusion and Future Work

Text classification is one of the most important topics. Readers can find a lot of researches talk about text classification using English dataset. Unfortunately, the number of researches conducted with Arabic dataset is not enough. Author thinks that this research could be considered as a reference for almost all researchers who are interested in text classification using Arabic dataset. This research talks about the most famous methods used in classification such as NB, SVM, MLP, K-NN, C4.5 and rocchio classifiers. Besides that, this research uses different feature selection methods such information gain, gain ratio and chi-square. Experiments in this research are done with a relatively huge dataset. Experiments show that NB is the best with small difference if compared with SVM. Also, experiments show that chi-square is little bit better than gain ratio and information gain.

References

Abdel-Salam, O., Sameh, G., Ali, Al-ibrahim, Nidhal, El-Omari, & Adel, H. (2013). Performance and Effectiveness Examination of the IQE and AQE with Application on Arabic Content. *International Journal of Current Engineering and Technology, IJCET, June-2013, 3*(3), 795-797.

Abdullah, H., & Wahbeh, M. Al-K. (2012). Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text, ABHATH AL-YARMOUK. *Basic Sci. & Eng., 21*(1), 15- 28.

Abu-Errub, A. (2014). Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements. *International Journal of Computer Applications, 93*(6), 40-45.

Adel, H. (2018). Apply Two Feature Selections (Chi-square and Symmetric Uncertainty) Using C4.5 Classification Algorithm Based on Arabic dataset._ISER-431st International Conference on Advances in Business Management and Information Science (ICABMIS), Istanbul, Turkey on 6th - 7th September.

Adel, H. (2018). Comparing Two Feature Selections Methods (Information Gain and Gain Ratio) On Three Different Classification Algorithms Using Arabic dataset. *Journal of Theoretical and Applied Information Technology, 96*(6).

Adel, H. (2019). Using Polynomial Neural Networks for Arabic Text Categorization. *European Journal of Scientific Research, 152*(3).

Adel, H., & Raed, A. (2011). Spam Detection Using Assisted Artificial Immune System. *International Journal of Pattern Recognition and Artificial Intelligence, 25*(8), 1275-1295.

Adel, H., & Raed, A. (2013). Genetic optimized artificial immune system in spam detection: A review and a model. *Artificial Intelligence Review, 40*(30), 305-377.

Adel, H., Omar, Al-M., & Tariq, A. (2016). Arabic Text Categorization Using k-nearest neighbour, Decision Trees

(C4.5) and Rocchio Classifier: A Comparative Study. *International Journal of Current Engineering and Technology, 6*(2).

Adel, H., Tariq, A., & Omar, A. (2016). Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network. *GSTF Journal of Computing, 5*(1), 108-115.

Aisha, A., Nazlia, O., & Adel, A. (2014). A COMPARATIVE STUDY OF COMBINED FEATURE SELECTION METHODS FOR ARABIC TEXT CLASSIFICATION. *Journal of Computer Science, 10*(11).

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., & Al-Rajeh A. (2008). Automatic Arabic Text Classification. 9es Journées internationals, France, pp. 77-83.

Al-Zaghoul F., & Al-Dhaheri S. (2013). Arabic Text Classification Based on Features Reduction Using Artificial Neural Networks", UKSim, pp. 485-490.

Anping, Z., & Yongping, H. (2011). A Text Classification Algorithm Based on Rocchio and Hierarchical Clustering", D.-S. Huang et al. (Eds.): ICIC 2011, LNCS 6838, pp. 432–439, 2011. c Springer-Verlag Berlin Heidelberg.

Aymen Abu-Errub. (2014). Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements. *International Journal of Computer Applications*, *93*(6).

Badr, H., Abdelkarim, M., Hanane, E., & Mohammed, E. (2014). A comparative study of decision tree ID3 and C4.5, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications.

Borrajo A. L., Seara, V. E., & Iglesias, L. (2015). TCBR-HMM: An HMM-based text classifier with a CBR system. *Applied Soft Computing, 26*, 463-473.

Caudill, M., & Butler, C. (1992). Understanding Neural Networks. *Computer Explorations, 1*(2). MIT Press, Cambridge MA, USA.

Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines (and other kernel-based learning methods). Cambridge University Press.

Edy, B., Haviluddin, N. D., Awang, H. K., & Masna, W. P. (2018). *Performance of Decision Tree C4.5 Algorithmin*, Student Academic Evaluation, Springer Nature Singapore Pte Ltd. 2018 R. Alfred et al. (Eds.): ICCST 2017, LNEE 488, pp. 380–389, 2018. https://doi.org/10.1007/978-981-10-8276-4_36

Eldos, T. (2003). Arabic Text Data Mining" A root Based Hierarchical Indexing Model. *International Journal of Modeling and Simulation, 23*(3), 158-166.

Fouzi, H., & Eyas, Al-Qawasmah. (2010). Improving Arabic Text Categorization Using Neural Network with SVD. *Journal of Digital Information Management, 8*(4).

Franca, Debole et al. (2003). Supervised Term Weighting for Automated Text Categorization", proceedings of SAC-03, 18th ACM Symposium on Applied Computing, Melbourne, 2003, USA,

Gongde, G., Hui, W., David, B., Yaxin, B., & Kieran, G. (2006). Using kNN Model-based Approach for Automatic Text Categorization. *Soft Computing, 10*(5), 423-430.

Johannes, F. (1998). A Study Using n Gram Features For Text Categorization. Technical Report OEFAI-TR-1998-30.

Ladha, L., & Deepa, T. (2011). FEATURE SELECTION METHODS AND ALGORITHMS. *International Journal on Computer Science and Engineering (IJCSE), 3*(5).

Li, B. L., Yu, S., & Lu, Q. (2003). An Improved k-Nearest Neighbor Algorithm for Text Categorization", Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China.

Liu, H., & Motoda. (1998). Feature Extraction, construction and selection: A Data Mining Perspective.", Boston, Massachusetts (MA): Kluwer Academic Publishers.

Majed, I. H., Fekry, O., Minwer, A. L., & Ahlam, S. (2011). ARABIC TEXT CLASSIFICATION USING SMO, NAÏVE BAYESIAN, J48 ALGORITHMS. *International Journal of Research and Reviews in Applied Sciences, 9*(2).

Mofleh, A. (2012). Arabic Text Categorization Using Classification Rule Mining. *Applied Mathematical Sciences, 6*(81), 4033-4046.

Mohammad, Ali H. Eljinini, Wa'el, M. H., Adel, H., & Mohammad, G. (2010). Performance of NB and SVM Classifiers in Arabic Text Data, proceedings of the 14th International Business Information Management Association, Conference on Global Business Transformation through Innovation and Knowledge Management, Istanbul, Turkey. IBIMA.

Montanes, E., Ferandez, J., Diaz, I., Combarro, E. F., & Ranilla, J. (2003). Measures of Rule Quality for Feature Selection in Text Categorization. 5th international Symposium on Intelligent data analysis, Germeny-2003, Springer- Verlag 2003, Vol2810, pp.589-598.

Motaz, K. S., & Wesam, A. (2010). Arabic Text Classification Using Decision Trees. proceedings of the 12th international workshop on computer science and information technologies CSIT'2010, Moscow – Saint-Petersburg, Russia.

Nidhi, V. G. (2011). Recent Trends in Text Classification Techniques. *International Journal of Computer Applications (0975 – 8887), 35*(6).

Raed, Abu-Z., & Adel, H. (2011). Application of Genetic Optimized Artificial Immune System and Neural Networks in Spam Detection. *Applied Soft Computing, 11*(4), 3827-3845.

Rasha, E., & Mahmoud, A. (2015). Arabic Text Classification review. *International Journal of Computer Science and Software Engineering (IJCSSE), 4*(1).

Rehab, D. (2005). Machine learning for Arabic text categorization. *Journal of American society for information science and technology (JASIST), 57*(8),1005-1010.

Rehab, D. (2007). Arabic c text categorization. *The international Arab journal of information technology, 4*(2).

Rish, I. (2001). An empirical study of the naive Bayes classifier. IJCAI Workshop on Empirical Methods in AI.

Riyad, A., Ghassan, K., & Manaf, H. G. (2006). Arabic Text Categorization Using kNN Algorithm. *Proceedings of the 4th International Multiconference on Computer Science and Information Technology, 4*, 5-7.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychological Review, 65*(6), 386–408.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning Internal Representations by Error Propagation", Parallel Distributed Processing, MIT Press, Cambridge MA, USA.

Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.). Prentice Hall. ISBN 978-0137903955..

Saleh, A. (2011). Automated Arabic Text Categorization Using SVM and NB. *International Arab Journal of e-Technology, 2*(2).

Sameh, G., Adel, H., & Ali, A. (2013). Innovative Artificial Neural Networks-Based Decision Support System for Heart Diseases Diagnosis. *Journal of Intelligent Learning Systems and Applications, 5*(3), 176-183.

Sebastiani, F. (1999). A Tutorial on Automated Text Categorization". Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, pp7-35.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47.

Sebastiani, F. (2002). Machine learning in automated text categorization' ACM Publication. *ACM Computing Surveys, 34*(1), 1–47.

Syiam, M. M., Fayed, Z. T., & Habib, M. B. (2006). AN INTELLIGENT SYSTEM FOR ARABIC TEXT CATEGORIZATION. *International Journal of Intelligent Computing and Information Sciences IJICIS, 6*(1).

Tarek, F. G., Mena, B. H., & Zaki, T. F. (2009). Arabic Text Classification Using Support Vector Machines. *International Journal of Computers and Their Applications, 16*(4).

Thorsten, J. (1998). Text categorization with support vector machines: learning with many relevant features". In Proceedings of the 10th European Conference on Machine Learning ECML-98, Chemnitz, Germany. Pages 137–142.

Vladimir, N. V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag Berlin.

Wail, H. K., Haytham, Saleem AL-SARRAYRIH & Lars, K. (2014). Arabic Text Categorization Using Improved k-Nearest neighbour Algorithm. *Journal of Applied Computer Science & Mathematics, 18*(8).

Wang, Y., & Wang, X. J. (2005). A New Approach to feature selection in Text Classification. *Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, 6*, 3814-3819.

Xindong, W., Vipin, K., J. Ross Quinlan, Joydeep, G., Qiang, Y., Hiroshi, M., … Dan, S. (2007). Top 10 algorithms in data mining, Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007 Published online: 4 December 2007, © Springer-Verlag London Limited 2007, Knowl Inf Syst (2008), *14,* 1–37. https://doi.org/10.1007/s10115-007-0114-2

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 42-49.

Zi-Qiang, W., Xia, S., De-Xian, Z., & Xin, L. (2005). An Optimal Svm-Based Text Classification Algorithm" Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp. 13-16, 2006. Barizal, pp.122-129.