The Islamic University–Gaza
Research and Postgraduate Affairs
Faculty of Engineering
Master of Computer Engineering

الجامعـــــــة الإســـلاميــة – غـزة
شئون البحث العلمي والدراسات العليا
كـليـــــة الهندسة
ماجستير هندسة الحاسوب

A new model in Arabic Text Classification Using BPSO/REP-Tree

نموذج جديد في تصنيف النص العربي باستخدام نظرية أمثلية سرب الجزيئات الثنائية مع خوارزمية تقليم الخطأ

Hamza Mohammed Naji

Supervised by

Wesam Ashour
Associate prof. of Computer Engineering

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Computer Engineering

September/2016

إقـــــــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

A new model in Arabic Text Classification Using BPSO/REP-Tree

نموذج جديد في تصنيف النص العربي باستخدام نظرية أمثلية سرب الجزيئات الثنائية مع خوارزمية تقليم الخطأ

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الاخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

I understand the nature of plagiarism, and I am aware of the University’s policy on this.

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted by others elsewhere for any other degree or qualification.

Student's name:	حمزة محمد ناجي	اسم الطالب:
Signature:	حمزة ناجي	التوقيع:
Date:	9/27/2016	التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية-غزة
The Islamic University of Gaza

الجامعة الإسلامية غزة
The Islamic University Gaza
١٣٩٨ - 1978

مكتب نائب الرئيس للبحث العلمي والدراسات العليا   هاتف داخلي1950

الرقم: ج س غ/35/ Ref:

التاريخ: 2016/10/08 Date:

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **حمزه محمد سالم ناجي** لنيل درجة الماجستير في **كلية *الهندسة*** قسم **هندسة الحاسوب** وموضوعها:

نموذج جديد في تصنيف النص العربي باستخدام نظرية أمثلية سرب الجزيئات الثنائية مع خوارزمية تقليم الخطأ

A new Model in Arabic Text Classification Using BPSO/REP-Tree

وبعد المناقشة التي تمت اليوم السبت 07 محرم 1438هـ، الموافق 2016/10/08م الساعة العاشرة صباحاً ، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. وسام محمود عاشور	مشرفاً و رئيساً	....
د. محمد أحمد الحنجوري	مناقشاً داخلياً	....
د. إيهاب صلاح زقوت	مناقشاً خارجياً	....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية *الهندسة*** / قسم وموضوعها: **هندسة الحاسوب.**

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق ،،،

شئون البحث العلمي والدراسات العليا
Research & Graduate Affairs
الجامعة الإسلامية غزة

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبد الرؤوف علي المناعمة

Abstract

Specifying an address or placing a specific classification to a page of text is an easy process somewhat, but what if there were many of these pages needed to reach a huge amount of documents. The process becomes difficult and debilitating to the human mind. Automatic text classification is the perfect solution to this problem by identifying a category for each document automatically. This can be achieved by machine learning; by building a model contains all possible attributes features of the text. But with the increase of attributes features, we had to pick the distinguishing features where a model is created to simulate the large amount of attributes (thousands of attributes). To deal with the high dimension of the original dataset, we use features selection process to reduce it by deleting the irrelevant attributes, words, where the rest of features still contain relevant information needed in the process of classification. In this research, a new approach which is Binary Particle Swarm Optimization (BPSO) with Reduced Error Pruning Tree (REP-Tree) is proposed to select the subset of features for Arabic classification process.

We compare the proposed approach with two existing approaches; Binary Particle Swarm Optimization BPSO with K-Nearest Neighbor (KNN) and Binary Particle Swarm Optimization BPSO with Support Vector Machine (SVM). After we get the subset of attributes that result from features selection process, we use three common classifiers which are Decision Trees J 48, SVM and the prepared algorithm REP-Tree (as a classifier) to build the classification model. We created our own Arabic dataset; the BBC Arabic News dataset that are collected from the BBC Arabic website and another two existing are used datasets in our experiments, Alkhaleej News Dataset and Aljazeera News Dataset. Finally, we present the experimental results and showed that the proposed algorithm is missionary in this area of research. In addition the effect of use Stemming and Normalization on the three datasets is investigated, and the results showed the positive effect on some results (the improvement of the classification accuracy).

Abstract in Arabic

تحديد عنوان مناسب أو وضع تصنيف محدد لصفحة واحدة من النصوص تعتبر عملية سهلة بعض الشيء, ولكن ماذا لو تواجد العديد من هذه الصفحات لتصل إلى كمية هائلة من الوثائق. العملية تصبح صعبة ومعقدة بالنسبة للعقل البشري. تصنيف النصوص الكترونيا هو الحل الأمثل لمثل هذه المشكلة, وذلك من خلال تعريف مسبق لكل صنف أو عنوان بشكل آلي. يمكن تحقيق ذلك بواسطة خاصية تعلم الآلة من خلال بناء نموذج تعلم يحتوي كل الصفات الممكنة للنصوص. لكن مع تزايد هذه الصفات, يلزمنا اختيار الصفات المميزة التي تحاكي مجموعة الصفات الأصلية لتدخل في بناء نموذج التعلم. للتعامل مع مشكلة تضخم مجموعة الصفات الأصلية نستخدم خاصية انتقاء الصفات للتقليل من هذه المجموعة وانتقاء صفات مميزة ذات صلة, واستبعاد الصفات بعيدة الصلة بالبيانات في هذا البحث تم تقديم طريقة جديدة لانتقاء الصفات المميزة, نظرية أمثلية سرب الجزيئات الثنائية مع خوارزمية تقليم الخطأ. قمنا بمقارنة الطريقة المقدمة مع طريقتين سبق تقديمهما في هذا المجال, نظرية أمثلية سرب الجزيئات الثنائية مع خوارزمية الجار القريب ونظرية أمثلية سرب الجزيئات الثنائية مع خوارزمية آلة دعم المتجه. تم استخدام ثلاث خوارزميات تعلم وهي خوارزمية شجرة القرارات, آلة دعم المتجه و خوارزمية تقليم الخطأ التي تم تقديمها مسبقا. تم إنشاء مجموعة بيانات جمعت من موقع الإذاعة البريطانية الناطقة باللغة العربية, وتم استخدام مجموعتي بيانات متاحة الوصول وهي مجموعة بيانات أخبار الجزيرة ومجموعة بيانات أخبار الخليج. وفي النهاية, تم عرض نتائج التجارب التي أجريت للطريقة المقترحة وقد أثبتت أنها طريقة واعدة في هذا المجال مقارنة بغيرها من الطرق والتراكيب. بالإضافة إلى ذلك تمت دراسة أثر إضافة بعض التحسينات, مثل التجذير والتسوية, على مجموعات البيانات الثلاثة, وقد أظهرت النتائج مدى التأثير الإيجابي على بعض قيم الأداء.

Acknowledgment

All Praise and thanks be to Allah! We praise and thank Him, ask Him for His Help and Forgiveness. First of all, I would like to express my gratitude to Almighty Allah, without him I could not finish this thesis.

I would like to thank my parents for allowing me to realize my own potential. All the support they have provided me over the years was the greatest gift anyone has ever given me.

Also I would like to thank my supervisor Dr. Wesam Ashour for his continued encouragement and guidance in every step of my Master of Computer Engineering level. He always finds time to discuss the research and give advice.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Network
ARFF	Attribute Relation File Format
BOW	Bag of Words
BPSO	Binary Particle Swarm Optimization
CSV	Comma-Separated Values
DT	Decision Tree
ED	Euclidean Distance
FN	False Negative
FP	False Positive
FSS	Feature Subset Selection
IDF	Inverse Document Frequency
ISRI	Information Science Research Institute
KE	Knowledge Engineering
KNN	K-Nearest Neighbor
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
ME	Maximum Entropy
NB	Naïve Bayes
OCA	Opinion Corpus for Arabic
OR	Odd Ratio
PSO	Particle Swarm Optimization
REP-Tree	Reduced Error Pruning Tree
RFA	Rocchio Feedback Algorithm
SMO	Sequential Minimal Optimization
SVD	Singular Value Decomposing
SVM	Support Vector Machine
TC	Text Categorization or Classification
TF.IDF	Term Frequency Times Inverse Document Frequency
TN	True Negative
TP	True Positive

Chapter 1

Introduction

Chapter 1

Introduction

1.1 Background and Context

The huge increase of using text in the electronic devices and web sites, in particular, is a motivation for categorizing these texts in automatic manner. That’s because of the insufficiency of human ability to handle them manually. The core task in the categorization is called the Text Categorization or Classification TC. The previous task is the ability of classifying a huge amount of groups of texts; each of them is called a text data-set or Corpora, to some predefined classes. In case of news data-set; for example, the classes can be Sport, Health etc., and other various classes based on their contents.

Text classification process in general consists of two phases. The first one is the preprocessing phase defined as the process that implements on the amount of texts to make some improvements for reducing the unnecessary terms. The preprocessing phase also contains reducing the extra phrases of one term by a process called Stemming. Stemming is the process of eliminating the derived words of one basic word such as the words “making makes” and turning them to their roots as the word "make". Another example of the stemming process are the words (argue, argued, argues, arguing, and argues) turning them to the stem "argu". On the other hands, (argument and arguments) are turned to the stem "argument". The preprocessing phase includes the removing of some prefixes and suffixes from the word instead of extracting the original root.

The second phase of text classification process is the classification step. The process of classifying the preprocessed text in the previous phase and presenting the corpora using a mechanism is called a classifier. To apply such two phases, we need to convert each dataset to a term vector which is the basic of text processing

(Salton and Buckley, 1988). But how many terms we need in each dataset based on what term we need is a question to be answered. The previous question leads us to add a new step in the text classification process, Arabic Text Classification in this thesis.

There is a middle step between preprocessing and classification process called "feature selection" (Li, Xia, Zong and Huang, 2009), it is a complementary process to the preprocessing stage performed after it to reduce the redundant terms (features) and to keep the sufficient terms to continue the classification process (Saeys, Inza and Larranaga, 2007). We demonstrate a combination of Binary Particle Swarm Optimization BPSO and Reduced Error Pruning Tree REP-Tree for the last process of selecting good sets of features for the Arabic TC task. Then we use the second half of the hybridized approach the REP-Tree and use it as a classifier as mentioned above.

The text classification processes can be done easily on the English language due to the smooth environment of it. In contrast, Arabic language is considered a complex language that contains many formations and many different kinds of forms of the word. The aforementioned difficulty in the Arabic language requires greater efforts in dealing with the classification of texts. Thesis focuses on the classification of the Arabic text which is the difficulty of Arabic expressive style when being employed in alternative languages like Persian, Urdu, Iranian language and alternative regional languages of Pakistan, Afghanistan and Persia. The Arabic language contents constitute a 3% of the web text content with the fourth order in languages ordering on-line (Al-Tahrawi and Al-Khatibb, 2015). The previous amount of content needs an accurate and effective classification to help the humans to easily use it .Thus, in the last 10 years the need for the effective and accurate classification has quickly been grown.

There are some classification algorithms that can be done in general text classification and can be proposed in Arabic such as: Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbor(KNN), Maximum Entropy (ME), Artificial Neural Network (ANN), Decision Tree (DT)and the Rocchio Feedback Algorithm. More recently, Reduced Error Pruning tree REP-Tree is investigated in Arabic TC. RET-Tree is a fast decision tree learning machine and it builds a decision tree based on the information gained or reducing the variance. Also, REP-Tree is a fast decision tree learner which builds a decision/regression tree using information gained as the splitting criterion, and prunes it by using reduced error pruning (Aimunandar and Winarko, 2015). REP-Tree was first used in Indian and English text classification in 2015 and 2012 (Kalmegh, 2015), (Patel and Upadhyay, 2012).

1.2 Research Problem

These are some of the weaknesses that led to the initiation of this thesis. First, the lack of availability of publicly free accessible Arabic Corpora and most of related works in the literature used small collected corpus. Also, some algorithms, classifiers, were neglected in the field of classification of the Arabic text and not addressed, except in the field of classification of English words .Therefore, we chose the Reduced Error Pruning Tree REP-Tree in our study which was not previously used nor proven in the field of Arabic texts classification in this area. The last issue is finding a good subset of features to support the task of Arabic text classification which is still a little bit slow and the scarcity of combinations that presented in this context is slow too.

1.3 Thesis Contributions

- Using REP-Tree was limited in these Literatures (Kalmegh, 2015), (Patel and Upadhyay, 2012) classifying just English and Indian news with no application of Arabic text. In this thesis, the algorithm is used to classify three Arabic Text data-sets, and show the results of this algorithm (classifier).

- Enriching and enhancing the process of finding the good subset of features by propose a new hybrid method (BPSO+REP-Tree).
- Using the REP-tree as a classifier to classify the subset of features that resulted from two existing feature selection combinations (BPSO+ K-Nearest Neighbor KNN and BPSO+ Support Vector Machine SVM).
- Building a new three systems (combinations) for the task of Arabic text-classification. We will explain them in chapter 4.
- Repairing a new Arabic data-set (Corpus) BBC-Arabic News to enable other researchers to deal directly with these datasets and extracting the resulting data.

1.4 Thesis Objectives

(1) To enhance the process of finding the good subset of features.

(2) To show the importance of the preprocessing phase and show the effect of using both Normalization and Stemming in the preprocessing phase.

(3) To show the efficiency of REP-Tree from the comparison results between the REP-Tree classifier and the other classifiers. In addition clarify the differences between them and show the strengths and weaknesses of each one of them. In this context, we selected some common classifiers such as Support Vector Machine SVM and J 48 decision tree.

1.5 Thesis Structure

The rest of the thesis will be as following: In Chapter 2, we present the related work in this context (Arabic Text Classification).

In chapter 3, we present a simplified explanation of the process of classifying texts (classification of Arabic texts) in particular, then we illustrate the importance of classification texts in the electronic forms and websites and present the characteristics of the Arabic language.

Chapter 4 presents the proposed work of the preprocessed corpus with the processes of feature selection by BPSO+REP-Tree, and presents the work of REP-Tree as a classifier then presents the experimental results in the next chapter.

Chapter 5presents the experiments of our proposed work and shows the obtained results of feature selection experiments and machine learning experiments (the classifiers we have selected),in addition gives the comparison results of adding Stemming and Normalization.

Chapter6concludes the thesis with the conclusion and commentary of the experimental results and provides a fast review of the whole thesis to make it easier for the reader to understand the general ideas of the thesis.

1.6 Related Publications

Naji, H., Ashour, W. (2016). Text Classification for Arabic Words Using Rep-Tree. *International Journal of Computer Science and Information Technology IJCSIT*, *8*(2), 101-108. doi:10.5121/ijcsit.2016.8208

Chapter 2

Related Works

Chapter 2

Related Works

In the discussion below, we focus on the works addressing Arabic TC. Since the number and quality of features used to express texts has a direct effect on classification algorithms, the following will discuss the main goal of feature reduction and selection and their impact on TC.

(Duwairi, Al-Refai and Khasawneh, 2009) made a comparison between stemming, light stemming, and word cluster. For training purposes, they chose K-Nearest Neighbor KNN technique, to show that light stemming achieves the highest accuracy and lowest time of model construction.

(Duwairi and El-Orfali, 2014) provided another study of the aspects that affect the performance of classification text representation, feature selection and the choice of the appropriate algorithm. They used two datasets; the first is about the politic issues which includes 300 topics, where the second dataset is the Opinion Corpus for Arabic OCA. The classifiers they used were (SVM, NB, and KNN). The results showed that the performance of the classification methods was dependent on the preprocessing type.

(Rushdi-Saleh, Martín-Valdivia, Ureña-López and Perea-Ortega,2011) compared 3 Feature Subset Selection FSS metrics. They carried out a comparative study to examine the effect of the feature selection metrics in terms of precision. The results in general revealed that Odd Ratio OR worked better than the others. Some studies focused on other techniques like N-gram and different distance measures and proved their effects on Arabic TC.

(Khreisat, 2009) used a statistical method called Maximum Entropy ME for the classification of Arabic words. The author showed that the Dice measures using N-gram outperforms using the Manhattan distance.

(El-Halees, 2007) performed the same classifier of the previous literature but different selection and reduction techniques were applied. The author used normalization and stop words removal to increase the ultimate accuracy.

(Kourdi, Bensaid and Rachidi, 2004) classified Arabic text documents automatically using NB. The average accuracy reported was about 68.78%, and the best accuracy reported was about 92.8%. El-Kourdi used a corpus of 1500 text documents belonging to 5 categories; each category contains 300 text documents. All words in the documents are converted to their roots. The vocabulary size of resultant corpus is 2,000 terms/roots. Cross-validation was used for evaluation. Most of related work in the literature used small datasets, and applied one or two classifiers to classify one corpus which is not enough to evaluate Arabic TC.

(Sawaf, Zaplo and Ney, 2001) used Maximum Entropy ME to make a classification to news articles. The study achieved accuracy about 62.7%.

(Al-Zoghby, Eldin, Ismail and Hamza, 2007) used Association Rules for Arabic text classification, and also they used CHARM algorithm with soft-matching over hard big O exact matching. Data sets consisting of 5524 records. Each record is a snippet of emails having the subject nuclear. The vocabulary size is 103,253 words.

(Harrag, El-Qawasmeh andPichappan, 2009) used the feature selection based on hybrid approach for Arabic text classification. They used direct tree algorithm and the accuracy was of 93% for scientific data set, and 90% for literary data-set. Harrag collected 2 data sets; the first one is from the scientific encyclopedia.

(Al-Shalabi, Kanaan and Gharaibeh, 2006) used KNN classifier with TF.IDF weighting scheme and light stemming to give an accuracy of 95%. Data sets were collected from (Al-Jazeera, An-Nahar, Al-Hayat, Al-Ahram, and Ad-Dostor) and from Arabic Agriculture Organization website. The data set consists of 621 documents belonging to 1of 6 categories (politics 111, economic 179, sport 96, health and medicine 114, health and cancer 27, agriculture 100). They preprocessed the corpus by applying stop words removal and light stemming.

(Brahimi, Touahria and Tari, 2016) addressed sentiment analysis for tweets in the Arabic language using some approaches with two free available datasets of (2000 tweets). They applied the light and root stemmer as a preprocessing phase and investigated the impact of reducing the size of the dataset by selecting the most relevant features on the classification efficiency and accuracy of three well used machine learning algorithms Support Vector Machine (SVM), Naïve Bayes (NB), and K-Nearest Neighbor (KNN).

(Oraby, El-Sonbaty and El-Nasr, 2013) worked on the impact of Stemming by applying the Khoja stemmer (Khoja, 1999), Information Science Research Institute (ISRI) stemmer (Taghva, Elkhoury and Coombs, 2005), and Tashaphyne Light Arabic Stemmer (Tashaphyne, 2010) on two datasets of the opinion classification problem, the results show that the Khoja stemmer is the best one.

(Shoukry andRafea, 2012) performed the classifiers Support Vector Machine SVM and Naïve Bayes NB on a dataset collected from twitter website. They applied the experiments on 2 documents of Arabic tweets and the results showed that the Support Vector Machine SVM was better than Naïve Bayes NB.

(Al-Thwaib, 2014) used the Sakhr summarizer Sakhr company website 2016 as a feature selector to choose the best words of documents instead of using all words and they used the TF feature. Documents, after using TF for feature selection, are classified using SVM classifier; the data set they used consists of 800 Arabic text documents. It is a subset of 60913-document corpus collected from many newspapers and other web sites. He succeeded to increase the accuracy by using the summarized corpus as input for Support Vector Machine SVM classifier.

(Al-Hindi and Al-Thwaib, 2013) made a comparison between two data-sets, each one contained 1000 Arabic documents. Text summarization was applied on one without the other. Accuracy has not improved much, but there was a difference in the time. When they used summarized documents, less time was needed to build the learning model.

(Sallam, Mousa and Hussein, 2016) An improved features selection technique used the Frequency Ratio Accumulation Method classifier with normalization and two stemming mechanisms: ISRI and Tashaphyne stemmers to improve the accuracy of Arabic text categorization. Three different known data sets predefined collected from the website www.aljazeera.net. They recorded an accuracy of 95.83% for Tashaphyne stemmer and 93.06% for ISRI stemmer.

(Abu-Errub, 2014) proposed a method to classify Arabic text by comparing a document with predefined documents categories based on its contents using the Term Frequency Times Inverse Document Frequency TF.IDF method measure. After that the document is classified into the appropriate sub-category using Chi Square measure. The dataset used in this study contained 1090 documents for training and 500 documents for testing, categorized into ten main categories. The results show that the proposed algorithm can classify the Arabic text datasets into predefined category.

(Goweder, Elboashi and Elbekai, 2013) used their developed technique, Centroid-based, to classify Arabic text. The proposed algorithm is evaluated using a dataset containing a 1400 Arabic documents collecting from 7 different classes. The results show that the adapted Centroid-based algorithm can classify Arabic documents without problems. They used some measurements Micro-averaging recall, precision, F-measure, accuracy, and error rates respectively. The measurements factors record a performance percentage of 90.7%, 87.1%, 88.9%, 94.8%, and 5.2% according to the previous order of measurements.

(Abidi and Elberrichi, 2012), in this paper, they presented a comparative study to assess the effect of a conceptual representation of the text. The K-Nearest Neighbor used and feature extraction was achieved via three preprocessing schemes Bag of Words, N grams, and a conceptual representation. The F-measure of Bag of Words is 64%, 68% for N gram's F-measure, and 74% for F-measure conceptual representation. Finally, the conceptual representation was the best one as the results shown.

(Raho,Al-Shalabi, Kanaan and Nassar,2015) investigated the importance of feature selection in Arabic corpus classification by making a comparison of the performance between different classifiers in different situations using feature selection with stemming, and without using stemming. The dataset collected from BBC Arabic website and the classifiers they used are DT, K nearest neighbors KNN, Naïve Bayesian Model NBM method and Naïve Bayes NB; also they used factors Measurements such as precision, recall, F-Measures, accuracy and time. The results showed the Accuracy of each classifier as the following: (D.T 99.4%, KNN 66.3%, NBM 92%, and NB 91.9%).

(Mohammad, Al-Momani and Alwada, 2016) provided a comparative study of Arabic text classification between three types of classifiers (k-Nearest Neighbor, Decision Trees C4.5, and Rocchio Classifier). These well-known algorithms are applied on a collected Arabic data set. Data set used consists from 1400 documents belongs to 8 categories, the same number of documents was used in the study experiments. They used two types of Measurements precision and recall, and the results of the experiments showed that the K-Nearest Neighbor records an average of 80% for Recall and 83% for precision, While Rocchio Classifier records an average of 88% for Recall and 82% for precision. Both of the previous Classifiers are better than C4.5 with average of 64% for Recall and 67% for precision.

(Hussien, Olayah, AL-dwan and Shamsan,2011) compared the Sequential Minimal Optimization SMO, Naïve Bayesian NB and J48 (C4.5) Algorithms using Weka tool for data mining, these algorithms implemented on Arabic data set. They used two approaches for preprocessing phase: elimination stop word and the normalization approach. The results showed that the SMO classifier achieved the highest accuracy (SMO accuracy is 94.8%) and the (error rate is 5.2%), followed by J48 (C4.5) (J48 (C4.5) accuracy is 89.4%) and the (error rate is 10.85%), then the NB classifier with (NB accuracy is 85.07%) and the (error rate is 14.93%).

(Kanan and Fox, 2015). This study talks about a new approach in Arabic text classification stemming; they developed a new model called tailored stemming, a new Arabic light Stemmer, with the usage of Support Vector Machine SVM classifier. The experiments were performed under 10-fold cross-validation training type, and gave these results for the predefined classes after using SVM as the following: Art and Culture 91.8%, Economics 93.5%, Politics 91.5% and Society 99.1%.

(Gharib, Habib and Fayed, 2009), applied the Support Vector Machine SVM classifier on Arabic texts corpus and compared it with these classifiers: Bayes classifier, K-Nearest Neighbor classifier and Rocchio classifier. They collected the corpus from three main Egyptian newspapers ElAhram [.ahram.org.eg/], ElAkhbar [. Akhbarelyom .Org .eg/], and ElGomhoria [. Algomhuria. Net. eg/] during the period from August 1998 to September 2004. 1,132 documents that contain 95138 words (22347 unique words). The results showed that the Rocchio classifier did well when the corpora is small or it has a small feature, on the other hands, the SVM is better than other classifiers when the corpora is large.

(Al-Harbi, Almuhareb, Al-Thubaity, Khorsheed and Al-Rajeh, 2008) used a common used classifiers Support Vector Machine SVM, and C5.0 classifier. C5.0 classifier, in general, gives better accuracy of 92.12% and accuracy of 86.42% using SVM. The tools used in the experiments are Rapid Miner and Clementine tools. The

Rapid Miner open source software was used to provide an implementation for the SVM algorithm and Clementine for the C5.0 decision tree algorithm.

(Al-Anzi and Abuzeina, 2016) grouped the similar unlabeled document into pre-specified number of topics using Latent Semantic Indexing LSI and Singular Value Decomposing SVD methods. The corpus they used contains 1000 documents of 10 topics, 100 documents for each topic. The results showed that EM method is the best of other methods with an average categorization accuracy of 89%.

(Zubi, 2009). This study is about using the web contents and applies some Arabic classification techniques on it. The general purpose of this study is to compare between two classifiers. The author used the K-Nearest Neighbor KNN Classifier and Naïve Bayes NB Classifier to apply the experiment. As mentioned by the author in his study. A corpus of Arabic text documents was collected from online Arabic newspapers archives, including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, and AlDostor as well as few other specialized websites. He collects 1562 documents classifying it into 6 different categories. After the comparison experiment finished, the results showed that the K Nearest Neighbors KNN with an average of (86.02%) was better than Classifier Naïve Bayesian with accuracy of (77.03%).

(Zrigui, Ayadi, Mars and Maraoui, 2012).They developed a new model based on the Latent Dirichlet Allocation (LDA) and the Support Vector Machine SVM; they used the LDA to sample “topics” of groups of texts. The results showed that the proposed LDA-SVM algorithm is able to achieve high effectiveness for Arabic text classification task (Macro-averaged F 1 88.1% and Micro-averaged F – 91.4%).

Chapter 3

Data Mining Concepts

Chapter 3

Data Mining Concepts

3.1 Data Mining Concepts

Data mining is considered as the process of extracting information patterns from large sets of data. One application of the data mining is the text mining expression, which defined as the process of detecting meaningful and interesting linguistic patterns from natural language texts (Hotho, Nurnberger and Paab, 2005), (Witten and Frank, 2005). In comparison with data saved in databases format, information stored in text documents is unstructured and is difficult to be dealt with. To deal with such data, a preprocessing is required to transform textual data into a suitable format for automatic processing (Hotho, et.al, 2005), (Singh, 2005). When we deal with data mining the information of it, which is usually unclear or undefined, we use some electronic or automatic approaches to make the process of using these information easy. When we apply the previous talking about data mining on text data mining, we found that the information is clear in texts, but we need to know the suitable way that represented the texts in the machine, Computer.

Now we can define the field of Text Mining as a field of data mining concerned with the representing data stored in texts in a suitable shape for automatic processing and to extract the clear information of the text datasets by applying the algorithms and methods from machine learning and statistics (Singh, 2005). The common problems in this field(as the researchers mentioned) are the text representation, information retrieval, text summarization, document clustering, and text classification. In each problem of them, we need to use the algorithms and methods from machine learning and statistics. Now we want to present each problem with a brief definition as the following:

Information Retrieval Problem: the information needed to be retrieved is started when a user enter a query in a system. The query does not get a single object but maybe many objects match the single query based on different ranks of relevancy. A task of the information retrieval systems returns documents that contain the most relevant information to the given query. In order to achieve this purpose, text mining techniques are used to analyze text data and to make a comparison between the extracted information and the given queries to find out documents that include answers (Witten and Frank, 2005), (Singh, 2005).

Text Representation Problem: is concerned with the problem of how to represent text data in an appropriate format for automatic processing. In general, documents can be represented in two approaches, the first one as a bag of words where the context and the word order are discarded and the second approach is to find common phrases in text and deal with them as single terms (Witten and Frank, 2005), (Singh, 2005).

Text Summarization problem: is an automatic detection of the most important phrases in a given text document and to create a condensed version of the input text for human use (Singh, 2005). Text summarization can be done for a single document or a document collection, multi-document summarization. Most approaches in this area focus on extracting informative sentences from texts and building summaries based on the extracted information. Recently, many approaches have been tried to create summaries based on semantic information extracted from given text documents (Witten and Frank, 2005), (Singh, 2005).

Text Classification Problem: is the assignment of text documents into one or more predefined categories based on their content (Singh, 2005), (Sebastiani, 2002). It is a supervised learning problem where the categories are known in advance (Singh, 2005). For the text classification problem, many machine learning techniques including decision trees, K-Nearest Neighbor, Support Vector Machine and Naive Bayes algorithm have been used to build text classification models.

The Document Clustering Problem: is a machine learning technique that is used to identify the similarity between text documents based on their content. Unlike text classification, document clustering is an unsupervised method in which there are no predefined categories. The idea of document clustering is to create links between similar documents in a document collection to allow them to be retrieved together (Witten and Frank, 2005), (Singh, 2005), (Wajeed and Adilakshmi, 2009).

3.2 The Text Classification Problem

In this subsection, we will explain the concept of the text classification problem aforementioned above. We can observe the ever-increasing spread of texts in electronic form or specifically in the computer. These texts cannot be dealt with except when being classified into compilations based on their content. This process can manually be applied to a simple set of texts, but what if we started to deal with a larger number of documents that are difficult to be handled manually. The foregoing can be resolved by automatic classification of texts. Text classification can be defined as the task of assigning natural language texts into one or more predefined categories based on their content (Singh, 2005), (Sebastiani, 2002). It can be considered as a natural language problem solution that aims to discard the manually classification and replace them with automated machine learning techniques of text classification. The first approach that preceded machine learning is the Knowledge Engineering KE, which is defined as a set of rules that manually encodes expert knowledge to specify how to classify text documents based on given categories. Next, the machine learning techniques became ranking Top for the text classification problem after 1990. It’s defined as the process of building a text classifier by the machine learning of the text feature inside a set of pre-classified text documents. So, here we find the difference between automatic machine learning classifiers and Knowledge Engineering technique, machine learning classifier for Text Classification task is built automatically and does not need manual definition by domain experts (Sebastiani, 2002). We will adopt in the coming study on the concept of Text

Classification TC Process, and now we will enumerate the elements of this process as it is shown in Figure (3.1) (Sebastiani, 2002).

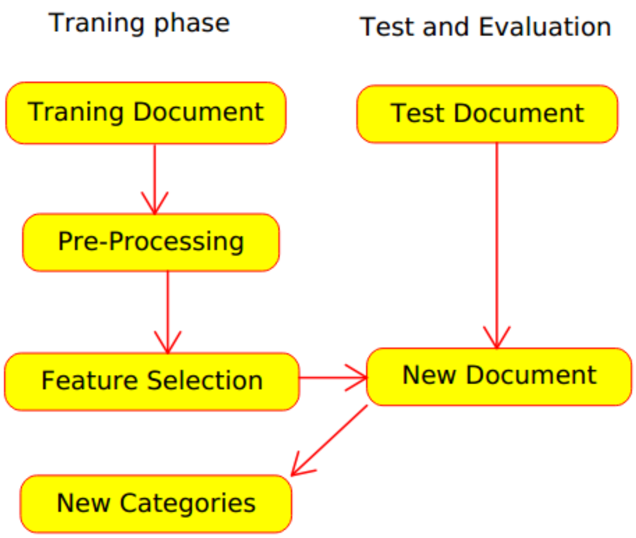


Figure (3.1): Text Classification Phases

The main three stages of text classification are text pre-processing, features selection, and then the final step is choosing the classifier to build the learning model of the training data to be ready for evaluation phase. The first phase is the pre-processing, in this phase we remove non-informative words such as punctuation marks and spaces by Tokenization process. In Arabic text-classification case, removing the non-Arabic letters is one of the pre-processing tasks, and removing the numbers (diacritics) special characters and punctuations. Also, we remove the stop words, pronouns, conjunctions, and prepositions. Another thing can be done in this phase which is the stemming process represented in reducing an inflected or derived word to its stem. The stem needs not to be a valid morphological root of the word as far as related words map to the same stem. The main advantage of this preprocessing step is to reduce the number of terms in the corpus so as to reduce the computational and storage requirements of TC algorithms. The second step we will improve the classification accuracy by the feature selection phase, it ignores the irrelevant and noisy terms or words, features, that do not have enough information for text classification process. The final step is to choose the classifier that will be applied on the training dataset using the best subset of features selected by the feature selection step then test the data to get the actual results.

3.2.1 Types of Text Classification

As mentioned in the literature (Sebastiani, 2006) the author classifies the text-classification problem into some types including soft and hard text classification, flat and hierarchal text-classification, and single label and multi label text classification. We explained all types as the following:

1. **Hard or Binary Classification.** Is one type of classifying texts depending on whether a specific document D is closed to category C or not, where the soft or decision classification type is a numeric score in a specific range used to check the rank of the classification accuracy (if that document D closed to class C) (Sebastiani, 2006).
2. **The Flat and the Hierarchal Text-Classification.** The author here divides the text-classification process into two types; the first type is called flat, which deals with the huge number of documents through one category with discarding the subtitle, so the search process in this type becomes more difficult. The hierarchal type solved the problem by dividing the main category into sub categories, for example, the main category is "News" can split into four sub classes like "polices, economics, sport,etc." (Wajeed and Adilakshmi, 2009).
3. **The Single-label and Multi-label Text-Classification.** We can define the process of assigning document D in a predefined category C by assigning documents in just one predefined category single-label, but the multi-label type can assign the documents in more than one category (Sebastiani, 2006).

Now, we will present the steps or the main phases of the text-classification process in the following subsection.

3.3 Text Classification Main Phases.

As mentioned previously, the classification process contains these three main tasks:

3.3.1 The Preprocessing Task.

It is the first phase in the text-classification process. Its role is to convert a text documents into a useful data-set to facilitate the classification process (Chen and Ho, 1999), (Witten and Frank, 2005). Here we will explain this process for especially Arabic text-classification and we will present the common steps as the following:

Tokenization*:* is the process of converting a sentence into tokens. The tokens (terms) here are the remaining characters after removing the spaces and punctuations in each document (Hotho, et.al, 2005), (Witten and Frank, 2005).

Stop words discarding: this is another task of the pre-processing tasks which concerned with removing the stop words from the documents. These words are the terms that occur frequently like prepositions e.g. for English language (in, the, on …), and for Arabic words (non-Arabic letters, من, إلى, إن …) (Chen and Ho, 1999), (Forman, 2002).

Stemming: is a pre-processing task to return the word or term to its root form where morphological information is used to match various shapes or patterns of words (Duwairi, et.al, 2007).The English language e.g. like “make, making, and maker ”can be returned to their stem (root) make”. But, what is the advantage of the word stemming in the text-classification process?. The answer is the dimension reducing to minimize the high dimension space problem to increase the accuracy of the classification process in systems (Larkey, Ballesteros and Connell, 2007). In our case, the common stemmers to be used in the experiments are the (Khoja stemmer, and light stemmer). Some extensions were added to the stemming process like: normalization as mentioned in the literature (Larkey, et.al, 2007).

Normalization process contains the following steps:

- Replace “ة” with “ه”.
- Replace “ي” with “ى”.
- Replace the aleph shapes with rooted shape. “أ, آ, إ” to “ا”.
- Remove the words from the prefixes and suffixes letters. Prefixes letters are a group of redundant letters located before the words like: (بـ, كـ, لل, لـ). The suffixes letters are a group of redundant letters located at the end of the words like: (ـه, ـة, ا, به, ان, يه, ون) (Chantar and Corne, 2011).
- Remove leading whitespace, extra spaces, non-letter, and non-space characters.
- Remove leading double lam at start of string.
- Bag of words BOW: is an approach in text pre-processing for text representation to make the whole documents to be represented as a vector of weights. According to convert texts into vectors, the BOW represents the terms as a single word (Silva and Ribeiro, 2003), (Song, Liuand Yang, 2005).

There are many ways of term weighting explained as the following:

- **Term Frequency:** A simple way to start with is by eliminating documents that do not contain all three words "the", "brown", and "cow", but this still leaves many documents. To characterize them, we can count the amount of times every term happens in each document and total them all together; the number of times a term happens through document is named its term frequency (Salton, Wong and Yang, 1975).
- **Inverse Document Frequency:** considered as the measure of the rare terms in a document the opposite of the Term Frequency which is measure the common terms. Using the term "the" ,which so popular , may tend to incorrectly emphasize documents that happen to use the word "the" a lot of oftentimes, while not giving enough weight to the additional significant terms "brown" and "cow”. The term "the" isn't a decent keyword to characterize relevant and non-relevant documents and terms, in contrast to the less common words "brown" and "cow". Therefore, associate inverse document frequency factor is included that

diminishes the terms weighting that occur terribly oft within the document set and will increase the burden of terms that seldom occur (Silva and Ribeiro, 2003).

- **Term Frequency-Inverse Document Frequency**: A high weight in TF.IDF is reached by a high term frequency in the given document and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Since the quantitative relation within the IDF's log function is usually larger than or adequate to one, the value of IDF and TF.IDF is bigger than or adequate to zero (Silva and Ribeiro, 2003). As a term appears in additional documents, the quantitative relation within the log approaches one transfer the IDF and TF.IDF nearer to zero. In several things, short documents tend to be diagrammatic by short vectors. Whereas a lot of larger-term sets are assigned to the longer documents. Normally, all text documents must have an equivalent importance for text mining aims. This means that a standardization factor can be included into the term-weighting to equalize the length of the document vectors (Ramos, 2003). In most cases, it's so difficult of the morphological variants to recognize by matching only. The text recognition process need additional algorithm called stemming that we mentioned above.

3.3.2 The Feature Selection Task

The reason of this step is the problem of data with high dimension terms space. Many of the classification techniques are not able to interact with problems with large amount of features, terms, space. So, with feature selection we aim to reduce the impact of this problem by eliminating the irrelevant and non-informative terms to reduce the high dimension of feature space (Mesleh and Kanaan, 2008). This reduction of the dimension can be done using the features selection by selecting a subset or a small group of the original group of features. It is also used to select features that contain sufficient and important information of the corpora (dataset) (Torkkola, 2004), (Martinez and Kak, 2001). We conclude that the feature selection process enhances the performance of the classification process and also presents the relation between the dimension of the corpora and the performance. So, when the number of the features in one corpus exceeding the accurate number of features the performance starts degradation gradually. This middle phase of the text-classification process requires the following steps (Liu, Zhang, Ma and Wu, 2004), (Chen, Tokuda and Nagai, 2003).

3.3.3 The Feature Selection Steps

1- Search Strategies.

Feature selection search strategies can be defined as the process of generating subsets of features from the original dataset for evaluation by an objective function which is the second step in feature selection process**.** Before talking about the search strategy in the features selection process. Let us list the general classification of search strategies approaches which are three classes, random, sequential, and complete search approaches. The random search approach tends to produce subsets of features and evaluates them to gain the best one of them (Mesleh and Kanaan, 2008). Where in the sequential search approach, features reduced or increased sequentially. The last one is the complete search approach which the whole subsets of terms are generated and tested to find the most accurately. Now, according to the general classification, we can list three types of feature selection search strategies as the following classes; the first one is the forward selection, the second is the backward eliminating selection, and the last one is the random selection approach (Torkkola, 2004). In the forward selection, the search starts with an empty set with no features, according to the goodness between features the set increases gradually. The forward selection also is considered as a type of sequential selection (Liu and Yu, 2005). In the backward eliminating, the data-set invoked with full amount of features, then the unrelated or unwanted features are eliminated. In the third type (which is the search methodology in this thesis), the random search selects the best features from the full amount of data-set features, depending on an evaluation function or objective function (Liu and Yu, 2005).

2- Objective Function

The main goal of this function is to check the goodness of one subset from the others subsets. It's invoked by the selection algorithm. The filter approach is one type of the objective function used to choose the best rank from other ranks of features from a data-set and remove the low ranks (Doan and Horiguchi, 2004), (Yang and Pedersen, 1997). Another type of objective function is the wrapper function, the chosen objective function in our work, which is used to select features with the best classification accuracy measured by the chosen classifier to be ready for the representation (Ng, Goh and Low, 1997). We choose the wrapper approach because it concerns about how much better a subset of features work together and, thus, can discover the nonlinear interactions throw big set of attributes. The disadvantages of filter approach are the concerning of missing the previous interactions between attributes and neglect some relevant and important attributes. More about feature selection process will explained in the next chapter.

3.3.4 PSO and BPSO

Before talking about BPSO as a feature selection algorithm, we will first describe the intended of the word “Swarm” in full definition of PSO “Particle Swarm Optimization” algorithm. What is the swarm and where this name came? That’s what we got from the final meaning of the definition. Many forms of life in some organisms affected the aspirations of some researchers and invited them to develop some successful theories for solving problems based on this random life. There is a group of successful theories based on this mode of thinking, including the DNA counting, membrane algorithm, Particle Swarm Optimization algorithm, artificial immune systems algorithm, and Ant Colony Optimization algorithm. One of the algorithms is the Particle Swarm Optimization algorithm that was developed in the 1995 by Eberhart and Kennedy (Kennedy and Eberhart, 1995). This idea has been built on the basis of the collective behavior of flocks of birds. PSO creates a random optimization algorithm to give solutions, particles, for some positions in the search space. Each of those particles holds an initial random velocity within the search space symbolized by V i = (V $_{i\,1}$; V $_{i2}$; ...V $_{iN}$), and each particle is symbolized by P $_{i}$ = (P $_{i\,1}$; P$_{i2}$; ...; P $_{iN}$). Update its velocity according to its experience or other particles experiences. For the best particle in the search space, swarm, we called it the best global symbolized by g, and when the velocity has been updated, the particle it finds the new position with the latest velocity according to the following equations (Yang, Chen and Zhao, 2007):

The main equation is:

$$Xid = Xid + Vid \qquad (3.1)$$

New position = Current position + New velocity.

$$Vid = \omega * Vid + C1 * rand(\quad) * (Pid - Xid) + C2 * rand(\quad) * (Pgd - Xid) \qquad (3.2)$$

Where

Where rand () is a random number between (0, 1) (Chantar and Corne, 2011). c1, c2 are acceleration factors. Usually c1 = c2 = 2. P$_{gd}$ = global best. Vid = velocity of particle (Tsai, Su, Chen and Lin, 2012).

X$_{i}$ is the current position of the particle initialized with random binary values. Where 0 means that the corresponding feature is not selected and by 1 means that the feature is selected. P$_{i}$ is the best previous position of the particle and initialized by the same value of X$_{i}$. V$_{i}$ is the velocity of P$_{i.}$

What if there was no previous velocity, then particles will navigate to the same position (current position), and that is the (local search). But if we get a new velocity, then particle will extend its search (the global search). Some problems resulted from the previous questions. Inertia weight ω solve these problems by balancing the local and global search. (Shi and Eberhart, 1998) perform a sequence of experiments to give the best value of ω which is 1.2.In Binary Particle Swarm Optimization Binary PSO, particle position is considered as a binary vector, but how binary vectors deal with velocities. (Kennedy and Eberhart, 1997) provided some equation to deal with velocity, a vector, (with real value in which this value is kept between (0, 1)), provides a group of probabilities. According to the previous we can use the BPSO to select the relative features in the Arabic Text Classification. As mentioned in the literature (Kennedy and Eberhart, 1997), the probability of bit changing is determined by the following:

$$S(Vid) + \frac{1}{1 + e^{-Vid}} \tag{3.3}$$

$$If\ (rand(\quad) < S(Vid))then\ Xid = 1;$$

Else = 0

Where rand () is a random number between (0, 1) (Chantar and Corne, 2011). c1, c2 are acceleration factors. Usually c1 = c2 = 2. P_{gd} = global best. Vid = velocity of particle (Tsai, Su, Chen and Lin, 2012).

3.3.5 Machine Learning Algorithm (The Classifier).

The phase of choosing the appropriate classifier can be applied on the subset that resulted from the feature selection phase. There are some popular classifiers in this context including decision trees, K nearest neighbor, Support Vector Machine and finally the Reduced Error Pruning Tree REP-Tree which will be used in this thesis.

Decision Trees J 48: is a machine learning tool that contains leaves, nodes, and roots like a real tree. It takes the form of a tree which has a collection of testing features built on the tree after constructed it by the training set. There are some popular examples of the decision trees like ID3 tree, and the upgraded version of ID3 tree (J 48). Decision trees are used to classify the unseen instances by the mechanism of check some features at each node to get the appropriate class of an unseen instance (Mitchell, 1997).

K-Nearest Neighbor KNN: this classifier is considered as a popular tool in text classification process by many researchers (Mitchell, 1997). The main idea of it is to store the training set in N-dimension space. When we add new instances and want to classify it, a set of similar training set of instances is invoked and is used to determine the category of the new instance. This can be done by measuring the similarity between two instances, neighbors, by applying the popular algorithm Euclidean Distance ED using the following equation (Mitchell, 1997):

$$D(x,y) = \sqrt{\sum_{i=1}^{N}\left(ai(x) - ai(y)\right)^2} \qquad (3.4)$$

x : Point 1.

y : Point 2.

a_i : The Value of each point.

In Figure (3.2) below we present the classification process of an instance depending of its K-Nearest Neighbor.

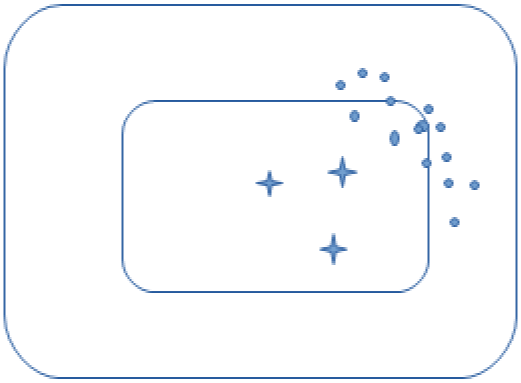


Figure (3.2): The classification process of the instances according of its K-Nearest Neighbor (Mitchell, 1997).

If K parameter of KNN is set to one the unseen or hidden instances will be classified as a positive instance. The dot (.) sign refers to instances and the (+) sign refers to class or category.

Support Vector Machine SVM: is a popular machine learning technique. The main idea of this classifier is mapping the input points in N-dimension space into another higher dimensional space and then a maximal separating hyper plane is found. It aims to separate amounts of data based on the optimal hyper plane between vectors which are linearity separable (Burges, 1988). The separation process depends on the maximum distance between the two sides of hyper plane and the nearest vectors in the training data sample. The linear algorithm of this classification process SVM can be defined as:

$$f(x) = W.P+ b. \tag{3.5}$$

P: is the vector of the training data-set.

b: is the bias, to manipulate the decision boundary of the linear hyper plane.

The W and b that solved the problem to determine the classifier (Cortes and Vapnik, 1995):

$$a= sign\big(f(W.P + b)\big) \tag{3.6}$$

W: is the weight vector for the best hyper-plane and b is known as the bias as mentioned above. The class of P after training (test instance) can be found using the following linear decision or activation function (a) as the following equation (Singh, 2005), (Wajeed and Adilakshmi, 2009):

$$a= sign\ f(x) \tag{3.7}$$

Where sign is the Continuous Log-Sigmoid Function, sig (n).

Figure (3.3) below shows the linear separable of a huge concentrate of data.

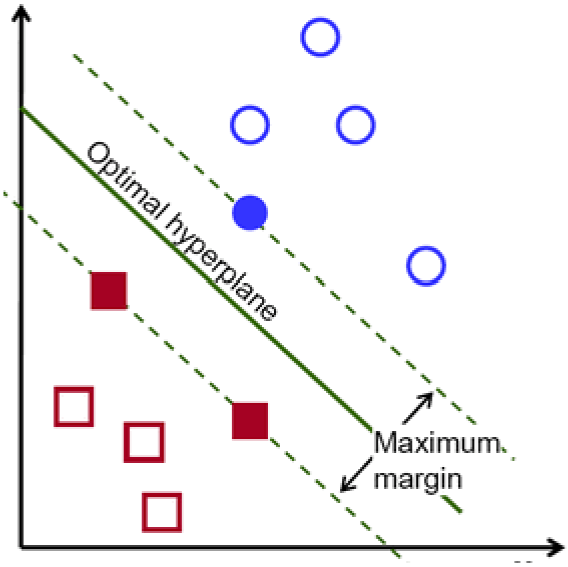


Figure (3.3): The best hyper-plane (best linear separable) (Cortes and Vapnik, 1995)

Where the closed (two squares and the single circle) are the support vectors.

Reduced Error Pruning Tree REP-Tree

More recently, Reduced Error Pruning tree REP-Tree is investigated in Arabic TC (Naji and Ashour, 2016). REP-Tree is a fast decision tree learning algorithm and it builds a decision tree based on the information gained or reducing the variance. REP-Tree is a fast decision tree learner which builds a decision/regression tree using information gained as the splitting criterion, and prunes it using reduced error pruning (Mitchell, 1997). REP-Tree was first used in Indian and English text classification in 2015 and 2012 (Kalmegh, 2015), (Patel and Upadhyay, 2012). The REP-Tree first starts the training process on the existing dataset, then build the training model by decisions, then get a mix results of some instances from the first learning step and from the pruned dataset which is a part of the dataset for post-pruning of the tree, then performing the test process. For a sub-tree of the tree, if replacing it by a node or leaf, which doesn't take more prediction errors on the pruning set than the original set, the tree replaces by a leaf. That means that the REP-Tree prunes each node after the natural classification. If the misclassification error determined for the instances from the pruned data set is not larger than the misclassification error rate computed on the original training data, the misclassification error can be presented in the Figure (3.4) below.

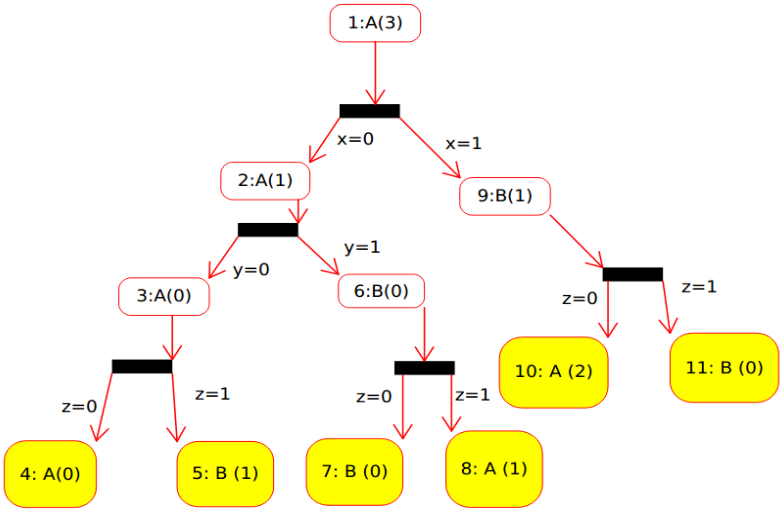


Figure (3.4): The misclassified detection in the pruning set of REP-Tree (binary sample), (Mitchell, 1997)

By using a pruning set shown in the following table, Table (3.1), contains some samples:

Table (3.1): A pruning set in REP-Tree classification (binary sample) (Witten and Frank, 2005)

Category	X	Y	Z
A	0	0	1
B	0	1	1
B	1	1	0
B	1	0	0
A	1	1	1
B	0	0	0

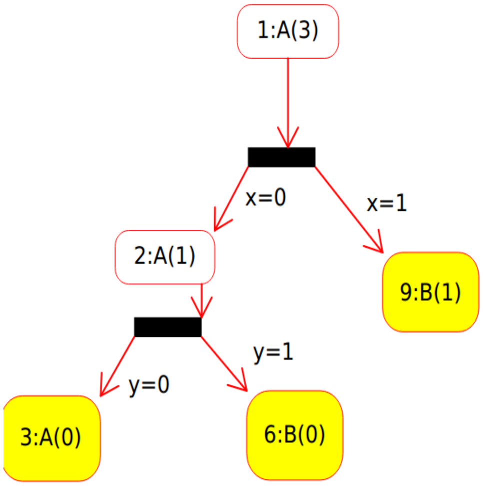


Figure (3.5): The final REP-tree (Witten and Frank, 2005)

The REP-tree begins from bottom from node three. We show that node three can be produced into a leaf which makes the minimum errors, on the pruning set, than as a sub tree. As a su-btree (the pruned tree) the classification occurs at nodes four and five. One error happened in node five; but no errors happened in node three. The same matter happened in node six and node nine. However, node number two cannot be made into a leaf since it makes one error while as a sub tree, with the newly-created leaves three and six. It makes no errors as shown in Figure (3.5). The pruning comes as a solution to the sub-tree replication problem that happened with the decision tree starts splitting. The definition of this case as *"When sub tree replication occurs, identical sub trees can be found at several different places in the same tree structure"* (Witten and Frank, 2005).

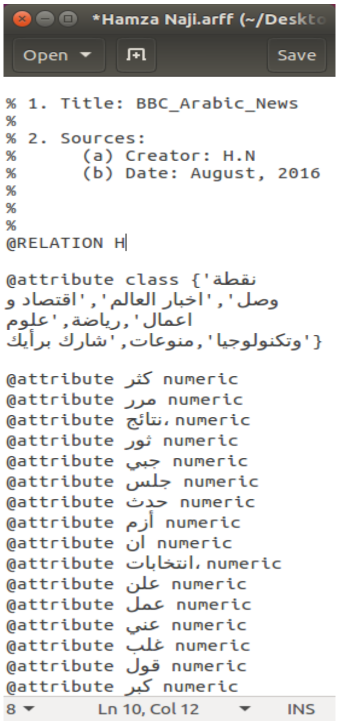
3.4 The Machine Learning Tool (Weka-Tool)

Waikato Environment for Knowledge Analysis Weka is a popular free and open source software tool for machine learning purposes written by the Java programming language based on the General Public License(GPL) of GNU (Weka 3: Data Mining Software in Java). The Name of the tool doesn't indicate the content, but letters taken from the phrase "Weka" do so, also is known as wooden which an endemic bird of a

New Zealand is. Weka tool contains many features and a huge package of classification and data mining algorithms which facilitate research process in front of researchers interested in this area by providing the easiest and simplest ways and presenting results in graphical interfaces that are easy to understand. There are two ways of sorting and representing data in Weka as the following (jmlr: Machine Learning Open Source Software):

- Comma-Separated Values CSV format: CSV is a comma separated values file with the .csv extension that allows text data to be stored in a table shape structure. CSV can be used in Weka tool by taking the form of CSV of information stored inside this file and separated by commas.

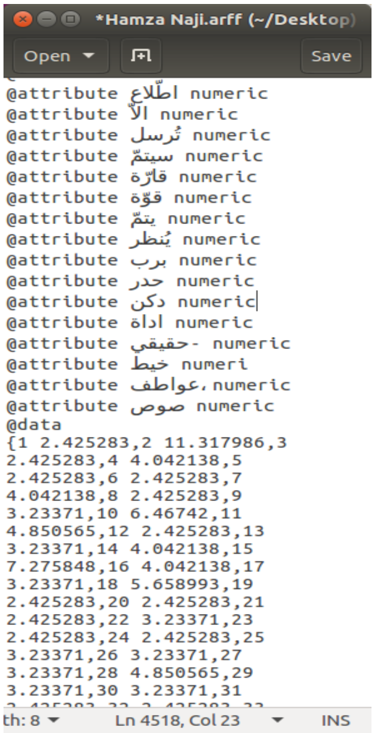
- Attribute Relation File Format ARFF: is another type of the data formatting which Weka can deal with, and its considered as the standard file format for Weka tool because it is developed by the Weka Project at the Department of Computer Science of The Waikato University in Zealand. The following graphs show the ARFF files components and the way data represented in (Witten and Frank, 2005).



```
% 1. Title: BBC_Arabic_News
%
% 2. Sources:
%       (a) Creator: H.N
%       (b) Date: August, 2016
%
%
%
@RELATION H

@attribute class {'نقطة
وصل','اخبار العالم','اقتصاد و
اعمال','رياضة','علوم
وتكنولوجيا','منوعات','شارك برأيك'}

@attribute كثر numeric
@attribute مرر numeric
@attribute نتائج، numeric
@attribute ثور numeric
@attribute جبي numeric
@attribute جلس numeric
@attribute حدث numeric
@attribute أزم numeric
@attribute ان numeric
@attribute انتخابات، numeric
@attribute علن numeric
@attribute عمل numeric
@attribute عني numeric
@attribute غلب numeric
@attribute قول numeric
@attribute كبر numeric
```



```
@attribute اطّلاع numeric
@attribute الآّ numeric
@attribute تُرسل numeric
@attribute سيتمّ numeric
@attribute قارّة numeric
@attribute قوّة numeric
@attribute يتمّ numeric
@attribute يُنظر numeric
@attribute برب numeric
@attribute حدر numeric
@attribute دكن numeric
@attribute اداة numeric
@attribute حقيقي- numeric
@attribute خيط numeri
@attribute عواطف، numeric
@attribute صوص numeric
@data
{1 2.425283,2 11.317986,3
2.425283,4 4.042138,5
2.425283,6 2.425283,7
4.042138,8 2.425283,9
3.23371,10 6.46742,11
4.850565,12 2.425283,13
3.23371,14 4.042138,15
7.275848,16 4.042138,17
3.23371,18 5.658993,19
2.425283,20 2.425283,21
2.425283,22 3.23371,23
2.425283,24 2.425283,25
3.23371,26 3.23371,27
3.23371,28 4.850565,29
3.23371,30 3.23371,31
```

Figure (3.6): The ARFF file

Figure (3.6) shows ARFF file representing the data-set of BBC_Arabic of the used corpora which represents the components of ARFF file and the Arabic words after pre-processing phase; for example, Stemming and Normalization processes.

3.5 Accuracy Evaluation

When we get the results of an experiment in the field of data mining, we need to verify the accuracy of the used machine learning algorithm. In text classification, we take a part of the data to be classified to be used to build a model of learning and the rest of data is used to evaluate the classification process (Mitchell, 1997). This can be done by two approaches: hold out and cross validation. The first one splits the available dataset into three third, two thirds as training set and the rest third is for testing. An enhanced approach to the previous approach is the stratified hold out, this enhance came to solve the problem of some classes didn't appear in the training port. So each class will be existing by collecting samples from all classes of the original dataset. The second approach is the cross validation, here the dataset splits into k folds and the whole folds split to training data and test data (Mitchell, 1997). Also by the last approach we solve the previous problem, and we guarantee that all of classes will appear in the training set.

3.5.1 F1- Measure

The accuracy measurement of the classification adopted in the coming experiments is the F1-Measure which is commonly used in the field of information retrieval. We can get the value of F1-Measure by two factors, precision and recall generated by the classifier which predicts the accurate class. For more clarification, suppose that there is a test data set and a document D related to the class C. There are four types of prediction for the case C class (Witten and Frank, 2005).

- **True Positive:** D is in class C in the training set, and also correctly predicted to be in the class C in the test set evaluation.
- **True Negative:** D is not in class C in the training set, and also correctly predicted not to be in the class C in the test set evaluation.
- **False Positive:** D is in class C in the training set, and also correctly predicted not to be in the class C in the test set evaluation.
- **False Negative:** D is not in class C in the training set, and also correctly predicted to be in the class C in the test set evaluation.

Now we will give a definition for both Precision and Recall as the following:

- **Precision** is the probability that a (randomly selected) retrieved document D is relevant to the predicted class C.

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{2.7}$$

Where TP is the True Positive and FP is the False Positive.

- **Recall** is the probability that a randomly selected relevant document D is retrieved in a search.

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{2.8}$$

Where TP is the True Positive and FN is the False Negative.

- **F1- Measure:** we can determine it from Precision and Recall which is the harmonic mean of precision and recall.

$$F1 - Measure = \frac{2.\,|TP|}{2.\,|TP| + |FP| + |FN|} \tag{2.8}$$

Where TP is the True Positive, FP is the False Positive and FN is False Negative.

3.6 Summary

This chapter presented the related works of the Text-Classification process in general and the Arabic Text-Classification literatures. The classification process is explained; also the phases of this process are investigated individually. This chapter, also, illustrates the features selection process as a middle phase in the classification process to reduce the huge amount of irrelevant features and choose the best subset of features to enhance text-classification field. In this chapter the machines learning Weka tool is presented and also the F-measure is explained.

Chapter 4

Proposed Work

Chapter 4

Proposed Work

In this chapter, the whole Arabic text classification process will be explained then it will divide the work into a collection of systems, each system has special combinations to produce the final process of classification after preparing the dataset. These combinations are taken from what has already been explained in previous chapters.

4.1 Arabic Text Datasets

In this subsection, we will present the datasets used in the experiments of our thesis. The used datasets are as the following:

4.1.1 BBC-Arabic News Dataset

The first data set contains the number of 4680 documents of BBC-Arabic news, classified into the following predefined categories {'Middle East', 'World News', 'business', 'sport', 'newspapers', 'Science', 'Misc.'}. We choose a random set of existing documents 3000 documents manually, with the knowledge that classify types in all documents as “single label” classification as mentioned in chapter (3.2.1 section) “types of text classification”. The following table, Table (4.1), shows the division of the documents into seven preset categories.

Table (4.1): The division of BBC-Arabic news Dataset based on 60% training set.

#	Class	Training Set	Testing Set	Full Dataset
1	Middle East	630	420	1050
2	World News	222	148	370
3	Business	124	82	206
4	Sport	348	232	580
5	Newspapers	234	155	389
6	Science	141	94	235
7	Misc.	102	68	170
Total		**1801**	**1199**	**3000**

- Note that BBC-Arabic data-set is collected during our work, and other two datasets are already existing in the literatures (Arabic Corpora - Mourad Abbas.) and (Arabic Corpora - Alj-News.).

4.1.2 Alkhaleej News Dataset

We present the second data set which contains a number of 5690 documents for Alkhaleej News Dataset (Arabic Corpora - Mourad Abbas.), (Arabic Corpora - Alj-News.) that classified into the following predefined categories {'International News', 'Local News', 'Sport', 'Economy'}. We choose a random set (2770 documents) with the knowledge that classify types in all the documents as a single label classification (Abbas, Smaili 2005). The following table, Table (4.2), shows the division of the documents into four preset Categories.

Table (4.2): The division of Alkhaleej News Dataset based on 60% training set.

#	Class	Training Set	Testing Set	Full Dataset
1	Local News	630	400	1030
2	International News	480	320	800
3	Economy	264	176	440
4	Sport	300	200	500
Total		**1674**	**1096**	**2770**

4.1.3 Aljazeera News Dataset

We present the second data set which contains a number of 1500 documents for Aljazeera dataset (Arabic Corpora - Alj-News.) classified into the following predefined categories {'Politics', 'Science', 'Sport', 'Economy', 'Art'}. We apply all 1500 documents without choosing a random set. The following table, Table (4.3), shows the division of the documents into five preset Categories with the same number of documents for each one.

Table (4.3): The division of Aljazeera News Dataset based on 60% training set.

#	Class	Training Set	Testing Set	Full Dataset
1	Politics	180	120	300
2	Science	180	120	300
3	Sport	180	120	300
4	Economy	180	120	300
5	Art	180	120	300
Total		**900**	**600**	**1500**

The tables above show that data is splittedinto two parts data for learning and data for testing based on 60% of learning; this style existed in Weka tool with many options for this purpose. Now, after we presented the data sets and ready for the classification process, we will list the list of classification processes.

4.2 Text Classification Processes

This subsection presents all the processes that we need in our experiments and we can present it as the following:

1- Text Pre-processing Process.

(1) *Tokenization.*

(2) *Stop words discarding.*

(3) *Stemming (Khoja stemmer, and light stemmer).*

(4) Normalization process contains the following steps:

Replace “ة” with “ه”.

Replace “ي” with “ى”.

Replace the aleph shapes with rooted shape. “إ ,آ ,أ” to “ا”.

Remove the words from the prefixes and suffixes letters.

2- Features Selection Process

Features selection using Binary Particle Swarm Optimization BPSO with:

a) K-Nearest Neighbor KNN.

b) Support Vector Machine SVM.

c) Reduced Error Pruning-Tree Rep-Tree. (This classifier is recently used as mentioned in chapter 3 and wasn't used for Arabic text classification. We will use it for two purposes to classify the Arabic texts ; the first is as an evaluator to evaluate the resultant features where selected by BPSO algorithm by measuring the accuracy of selecting process, and the second purpose as a classifier to test the rest data-set (40% of dataset).

In general, we can say that the three algorithms will be used with BPSO to check the fitness of a particle in an appropriate position. We will mention the steps of the previous three algorithms separately when using them in each proposed system.

3- Machine Learning Process

- Decision Tree J 48.
- Support Vector Machine SVM.
- Reduced Error Pruning-Tree Rep-Tree.

J 48 (C 4.5) and SVM are two common used classifiers in this area as mentioned in the previous chapter, in addition we use them to compare the proposed systems "combinations" with others such: (Tahrawi and AlKhatibb, 2015) and (Chantar, 2013). Also they used the BPSO with these two classifiers to produce combinations in Arabic text feature selection area. So we should use these classifiers to prove our feature selection approach (BPSO+REP-Tree), of course in addition to use it as a classifier.

4.3 The Proposed Systems

In this section, we will give a set of regulations contain some processes that listed in the previous section, and then a comparison will be performed between all the existing combinations in the form of independent systems and extract the results in the next chapter.

4.3.1 System A: Binary Particle Swarm Optimization and K-Nearest Neighbor.

System A is the first proposed system. It works on the classification of Arabic documents using the three main processes preprocessing, feature selection, and classifications as mentioned. This system contains three processes shown in Figure (4.1):

(1- *Tokenization, Stop words discarding 2- BPSO/KNN, 3*- J 48).

(1- *Tokenization, Stop words discarding 2- BPSO/KNN, 3*- SVM).

(1- *Tokenization, Stop words discarding 2- BPSO/KNN, 3*- Rep-Tree).

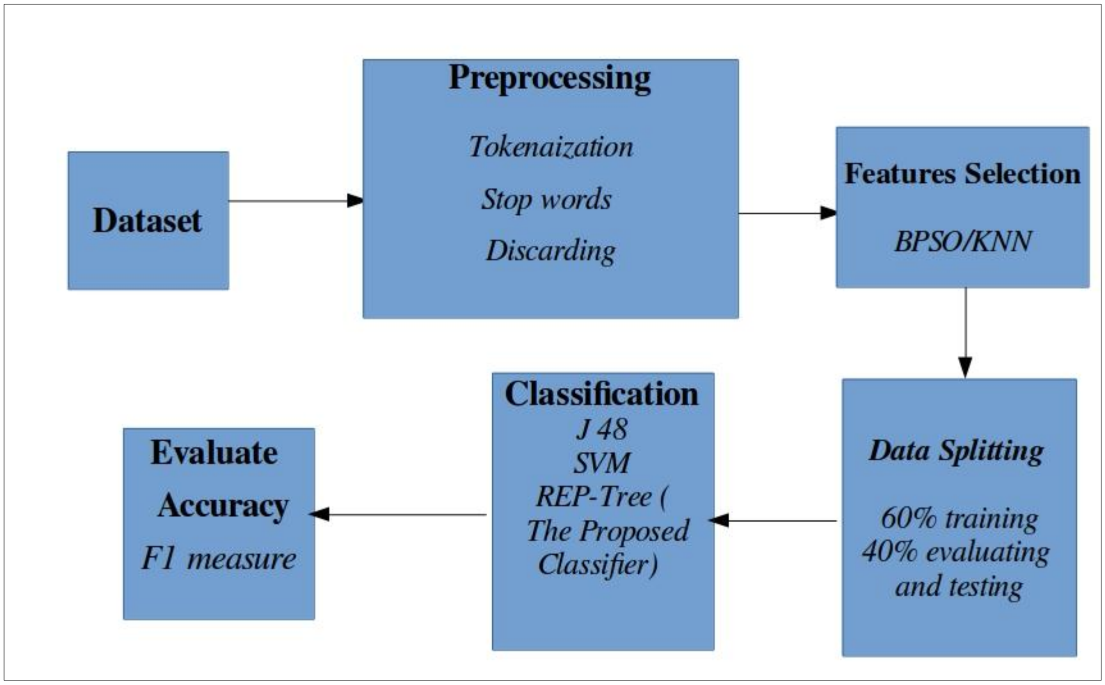


Figure (4.1): System **A**.

Figure (4.1) shows the processes of system **A** using the BBC-Arabic dataset with the previous processes without using Stemming or Normalization. Next, we repeat the same experiment after the addition of Stemming or Normalization.

4.3.1.1 BPSO+KNN Experiment Steps

Step 1. We need to prepare a population of particles in the features space and spread particles randomly. X_i is the current position of the particle initialized with random binary values. Where0 means that the corresponding feature is not selected and by 1 means that the feature is selected. P_i is the best previous position of the particle and initialized by the same value of X_i. V_i is the velocity of P_i.

- According to the evaluation of each particle in the swarm gbest (global best) initializes by the best fitness value of a particle.

Step2.(Determining the fitness).Fitness of subset resulted by particle with the evaluation process occurs after each feature selection iteration. The best fitness is the best accuracy in the evaluation process of the selected subset of features measured by classifiers algorithms (KNN) according to the following equation (Chantar and Corne, 2011).

$$Fitness = (\alpha * Acc) + \left(\beta * (N - T/N)\right) \tag{4.1}$$

Where

- Acc refers to the classification accuracy of the particle using chosen classifier.
- To make a balance between classification accuracy and the dimension of the feature sub set that selected by particles, we use the β and α parameters to do this purpose, with range of [0, 1] for α, and 1- α for β.
    - N refers to the all features.
    - T refers to the selected features using particle P.
- The fitness now is updated and then the private best of each particle is updated for each particle.

Step 3. (Updating gbest).The gbest is now updated.

Step 4.(Updating position).According to the BPSO velocity equation from chapter three, we can alter and update both velocity and position for all particles (Mendes, Kennedy and Neves, 2004).

$$Vid = \omega * Vid + C1 * rand(\quad) * (Pid - Xid) + C2 * rand(\quad) * (Pgd - Xid) \tag{4.2}$$

$$Xid = Xid + Vid \tag{4.3}$$

As mentioned in the literature (Kennedy and Eberhart, 1995), the probability of bit changing is determined by the following:

$$S(Vid) + \frac{1}{1 + e^{-Vid}} \tag{4.4}$$

$$If\ \left(rand(\quad) < S(Vid)\right) then\ Xid = 1;$$

Else = 0

Where rand () is a random number between (0, 1) (Chantar and Corne, 2011). c1, c2 are acceleration factors. Usually c1 = c2 = 2. P_{gd} = global best. Vid = velocity of particle (Tsai,Su, Chen and Lin, 2012).

Step 5. If the fitness value is better than the best fitness value (gbest) in history then set current value as the new gbest.

Step 6. Now for evaluation in our case KNN, we use the Euclidean Distance ED to measure the relevancy between current instance and the other instances in the data-set.

Step 7. Define the repository R.

- If the predicted classifications of instances were similar to the predefined classification, increase repository R by 1.

Step 8. Now, we can measure the classification accuracy of particle P by (Chantar and Corne, 2011).

$$ClassificationAccuracy = \frac{R}{N} \tag{4.5}$$

Where R is the group of results after testing the features from all training set N.

4.3.1.2 The Experiment Parameters (BPSO+KNN).

(1) Inertia weight (ω): in the previous equation (3.2) is to balance the local search and the global search (Chantar and Corne, 2011), and from the literature the best value of ω is 1.2.

(2) The swarm dimension is 50 units.

(3) Iterations are 200 iterations.

(4) [0, 1] for α, and 1- α for β. If we use the 1 for α then β = 0 and this mean that the dimension of the features subset is neglected, so we choose a random number between [0, 1] for α (0.70); and β is 1 – 0.70 = 0.30.

Now we can add some enhancements to System **A** such as (Stemming or Normalization), then we will study the effects of adding these enhancements before using J 48, SVM, and Rep-Tree with BPSO+KNN feature selection algorithm. That can be clarified by Figure (4.2) bellow.

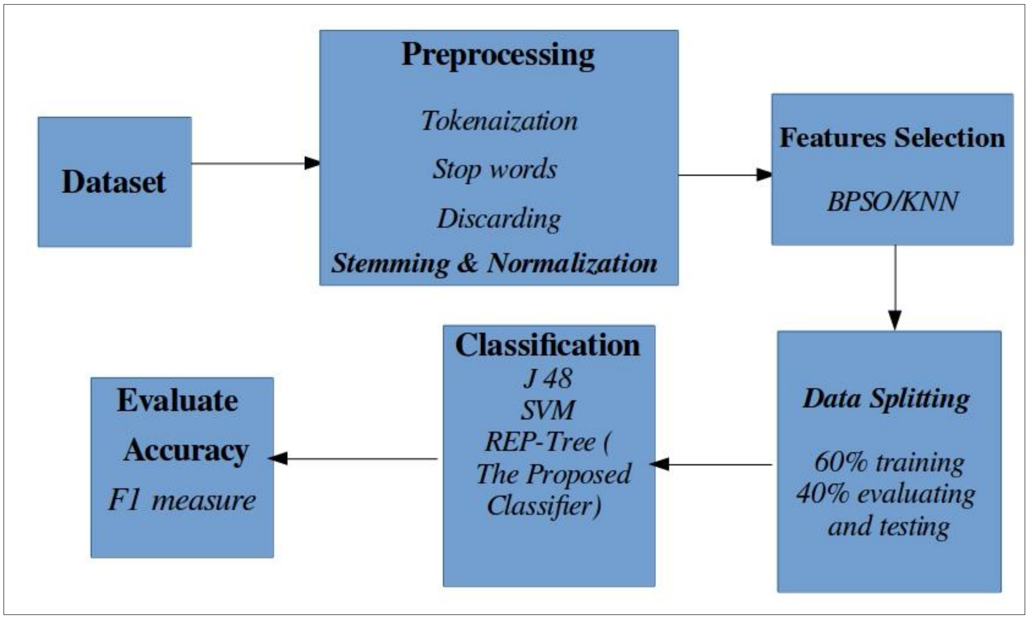


Figure (4.2): System A with adding Stemming and Normalization

Figure (4.2) shows that System **A** has been changed after the additional of such enhancements such as Stemming and Normalization before the classification of data into predefined Categories; we will present the results of the effectiveness of this process in the next chapter.

4.3.2 System B: Binary Particle Swarm Optimization and Support Vector Machine.

The second system in this also studies inserting the second middle phase (Feature Selection). In this system we will use the BPSO with SVM, and then classify the resultant features by (Decision Trees J 48, Support Vector Machine SVM, and Reduced Error Pruning-Tree Rep-Tree) as shown in Figure (4.3).

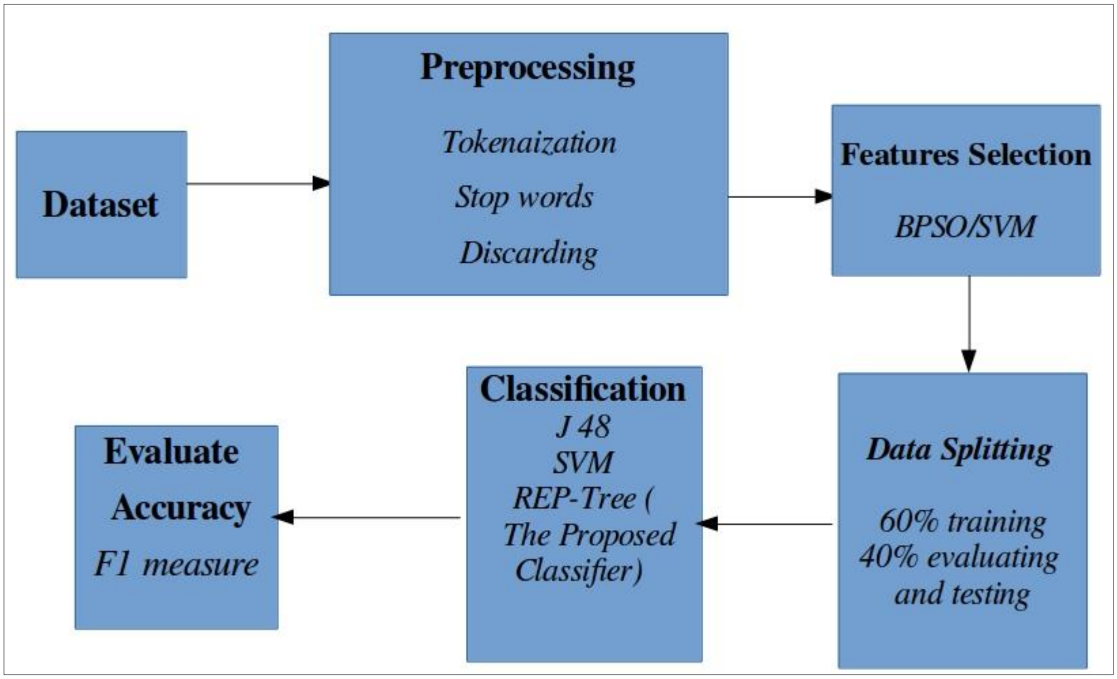


Figure (4.3): System **B**.

Figure (4.3) shows the processes of system **B** using the BBC-Arabic dataset with the previous processes by adding the BPSO+SVM as a feature selection, and the resultant features will be classified using the three classifiers SVM as classifier, J48, and REP-Tree for Arabic words, without using additional enhancements such as stemming or/and normalization. Next, we repeat the same experiment after the addition of Stemming or Normalization.

BPSO+SVM Experiment Steps

Step 1. The same in system A.

Step2.(Determining the fitness). Here we use the previous equation in system A, (3.1). Here we use SVM to measure the classification instead of KNN in the previous system A.

Step 3. (Updating gbest) the same in A.

Step 4.(Updating position) the same in A using the equations (3.2), (3.3), and (3.4).

Step 5. The same in A.

Step 6. Now for evaluation in our case SVM, we use the SVM classifier in Weka tool to measure the relevancy between current instance and the other instances in the data-set.

Then repeat both step 7 and 8 as mentioned in system A. also the same previous parameters in system A. experiments.

Now we can add some enhancements to System **B** such as Stemming or Normalization, then we will study the effects of adding these enhancements before using J 48, SVM, and Rep-Tree with BPSO+*SVM feature selection algorithm.* That can be clarified by Figure (4.4) bellow.

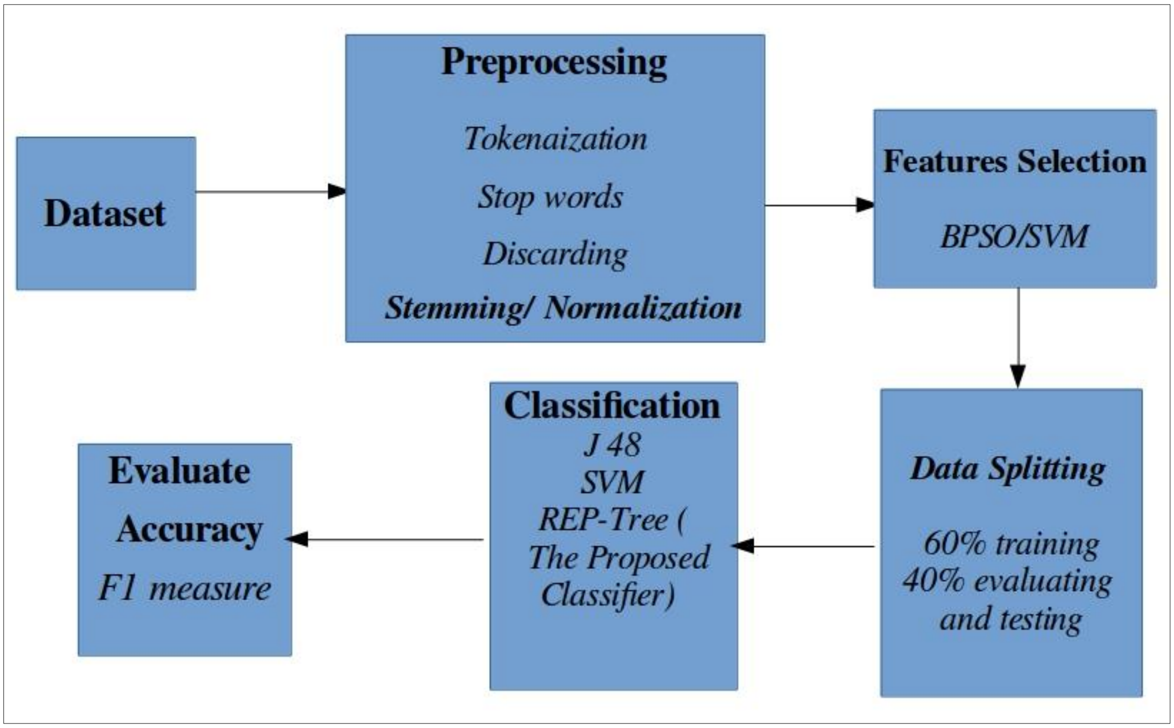


Figure (4.4): System **B** with adding Stemming and Normalization

Figure (4.4), shows that System **B** has been changed after the additional of such enhancements such as Stemming and Normalization before the classification of data into predefined categories; we will present the results of the effectiveness of this process in the next chapter.

4.3.3 System C: Binary Particle Swarm Optimization and Reduced Error Pruning Tree.

The last system in this study also involves inserting the middle phase Feature Selection including the previous processes and contents in system A, and B. In this system we will use the BPSO with Reduced Error Pruning-Tree Rep-Tree where it was not used in Arabic text classification field yet and it was recently used in English news classification. Finally, we will classify the resultant features by Decision Trees (J 48), Support Vector Machine SVM, and Reduced Error Pruning-Tree. Rep-Tree) (As a classifier) as shown in Figure (4.5).

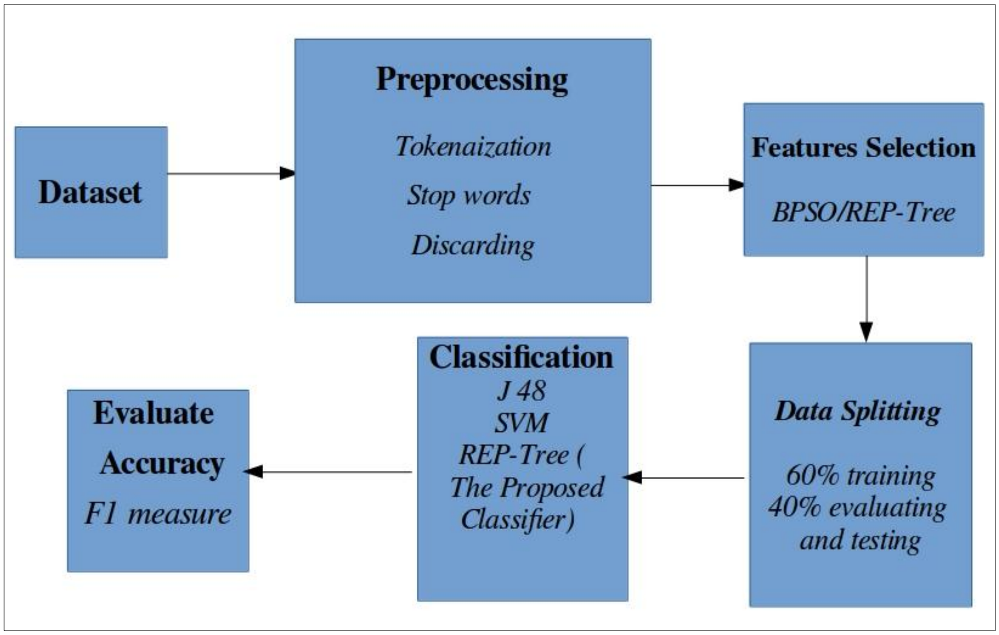


Figure (4.5): System **C**.

Figure (4.5) shows system **C** with adding the BPSO+REP-Tree as a features selection REP-Tree here (evaluator), and the resultant features will be classified using the three classifiers (SVM, J48, and REP-Tree (as classifier)) for Arabic words, without using additional enhancements such as stemming or / and normalization. Next, we will make comparisons between the three classifiers and then show the results in the next chapter.

3.3.3.1 BPSO+REP-Tree Experiment Steps

Step 1. The same in system A.

Step2.(Determining the fitness).Here we use Reduced Error Pruning-Tree REP-Tree as a feature evaluator to measure the classification accuracy of the particle in a training set, instead of KNN in system A.

Step 3. (Updating gbest) the same in A.

Step 4. (Updating position) the same in A. using the equations (3.2), (3.3), and (3.4).

Step 5. The same in A.

Step 6. Now for evaluation in our case REP-Tree we use REP-Tree classifier in Weka tool to measure the relevancy between current instance and the other instances in the dataset.

Then repeat both step 7 and 8 as mentioned in system A. also the same previous parameters in system A. experiments.

We can alternate the last three steps by measuring the F measure factor to estimate the classification accuracy.

We can list the previous steps in short and general points as the following:

(1) First and after preparing the features ,terms, space and spread particles randomly, we determine the accuracy of the classification (Acc) of a particle P in training data-set by using Reduced Error Pruning-Tree REP-Tree.

(2) Start extracting and filtering the features subset of the training set that selected by particle.

(3) Evaluate the previous extracted features data-set by the REP-Tree by 60 % training set validation.

(4) Determine the F measure factor that result from the REP-Tree experiment to determine the fitness of the particle.

Like the previous systems, we will compare between System **C** with extra improvements after the experiments of the system. The effective of addition of stemming or/and normalization will be shown. When adding the previous improvements, Figure (4.6) shows system **C.**

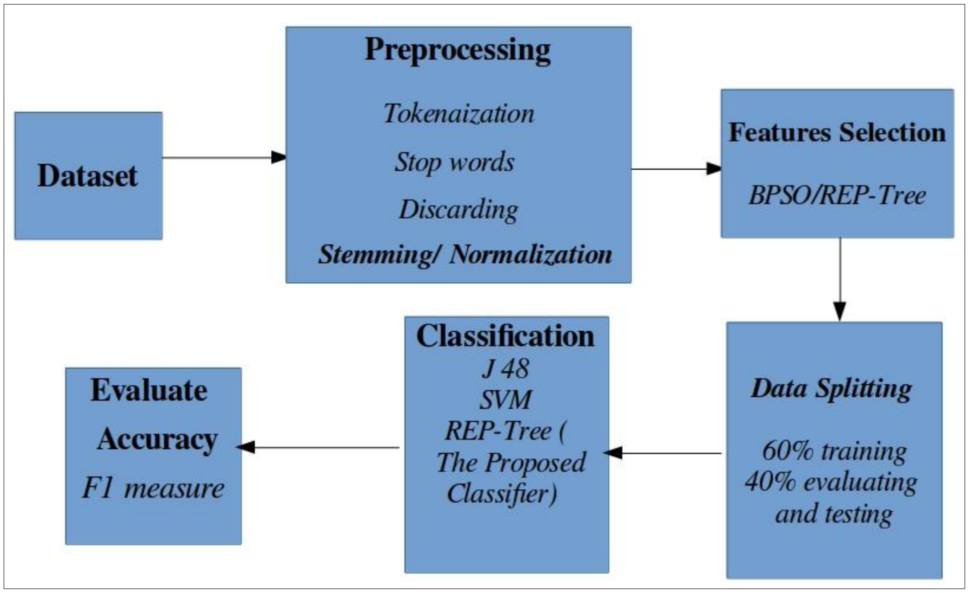


Figure (4.6): System **C** with adding Stemming and Normalization

The previous figure investigates the additional enhancements such as stemming or/and normalization had been added to the System **C** before the classification of data into predefined Categories. We will clarify the results of this system in the next chapter.

4.4 Summary

In this chapter we explained each of the three systems by listing the concept of each equation and its symbols, then we present some related figures that explained each system and its improvements such as normalization and word stemming. Finally, the experiments steps of each system are listed and the appropriate parameters are assigned.

Chapter 5

Experimental Results

Chapter 5

Experimental Results

In this chapter, the experimental results of the previous systems are described in last chapter. We have executed our experiments on three data-sets, the BBC-Arabic news dataset, Alkhaleej News dataset, and Aljazeera News dataset. As mentioned in the previous chapter, we split the data into 60% for training and 40% for testing, and then display the results in Tables and Figures. After that, we will compare every system with the other in specific graph. We will start presenting the results of system **A** using the three classifiers which have been previously described in Chapter 4. Then gradually we will review the results of system **B**, and finally we end with system **C.** The effect of adding both Stemming and Normalization is examined, and record the results for accuracy of the classification process for each system.

5.1 System A.A (“BPSO+KNN”/J 48)

The experimental results of system **A** with J 48 tree are shown by Table (5.1), (5.2), and (5.3) using the previous three datasets:

Table (5.1): System **A** with J 48 tree applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	67.3	69.7	68.4
World News	81.5	85.4	83.4
Business	72.4	73.4	72.8
Sport	84.2	79.7	81.8
Newspapers	87.3	88.9	88.0
Science	62.7	86.1	72.5
Misc.	83.9	89.6	86.6
Average	**77**	**81.8**	**79**

Table (5.1) shows the classification of BBC-Arabic documents using BPSO+KNN as a feature selector and J 48 decision tree as a classifier. As it is clear from the table, the results are as the following: the best classification is in “Newspapers” class with precision of 87.3, recall of 88.9 and F1-Measure of 88.0. The second performance rank of classes is the “Misc.” with precision of 83.9, recall of 89.6 and F1-Measure of 86.6. There is a convergence in the outcome of both “Word News” and “Sport” with a little outperforming in recall of 85.4 for “Word News” class. The worst two classes were the “Science” and the “Middle East” classes with precision of 62.7, recall of 86.1 and F-Measure of 72.5 for “Science” and the worst precision with 67.3 and F-Measure with 68.4 for “Science” class. Then we have the second data-set (Alkhaleej News Dataset) with the same previous experiment, Table (5.2) shows the results as the following:

Table (5.2): System **A** with J 48 tree applied on Alkhaleej News Dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Local News	75.8	78.4	77
International News	74.6	72.3	73.4
Economy	65.2	60.7	62.8
Sport	81.3	87.5	84.2
Average	**74.2**	**74.7**	**74.3**

Table (5.2) shows the classification of Alkhaleej News Dataset documents using BPSO+KNN as a feature selector and J 48 decision tree as a classifier. The best F-Measure is for “Sport” class with 84.2, and the worst F-Measure is for “Economy” class with 62.8. The next table, Table (5.3), describes the results of the Aljazeera News dataset by the same system.

Table (5.3): System **A** with J 48 tree applied on Aljazeera News dataset

Class	Precision%	Recall%	F1-Measure%
Politics	57.3	67.7	62
Science	88.4	87.9	88.1
Sport	86.8	88.1	87.4
Economy	68.2	75.6	71.7
Art	71.2	52.7	60.5
Average	**74.3**	**74.4**	**73.9**

Table (5.3) shows the classification of Aljazeera News dataset documents using BPSO+KNN as a feature selector and J 48 decision tree as a classifier. The best F-Measure achieved by J 48 is for “Science” class with 88.1, then we have the second rank of measuring, the “Sport” class with F-Measure of 87.4 and the worst F-Measure is for “Art” class with 60.5.

5.2 System A.$_B$ (“BPSO+KNN”/SVM)

The experimental results of system **A** with SVM classifier are shown by Tables (5.4), (5.5), and (5.6)using the previous three datasets (BBC Arabic, Aljazeera, and Alkhaleej datasets as the following:

Table (5.4): System **A** with SVM classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	88.3	79.7	83.7
World News	81.7	87.3	84.4
Business	84.5	92.4	88.2
Sport	87.2	79.7	83.2
Newspapers	86.4	88.2	87.3
Science	81.4	85.6	83.4
Misc.	89.4	95.6	92.3
Average	**85.5**	**86.9**	**86**

Table (5.4) shows the classification of BBC-Arabic documents using BPSO+KNN as a feature selector and SVM as a classifier. As it is clear from Table (5.4), the results are as the following: the best classification is for “Misc.” class with precision of 89.4, recall of 95.6 and F1-Measure of 92.3. The second performance rank of classes is the “Business” with precision of 84.5, recall of 92.4 and F1-Measure of 88.2. There is a convergence in the F1-Measure outcome of both “Middle East” and “Science” with F1-Measure of 83.7 and 83.4 gradually. The worst class is the “Sport” with precision of 87.2, recall of 79.7 and F-Measure of 83.2. Now we will apply system **A (**the same previous experiment with SVM**)** on the second data-set (Alkhaleej News Dataset) and Table (5.5) shows the results as the following:

Table (5.5): System **A** with SVM classifier applied on Alkhaleej News Dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Local News	86.1	90.4	88.1
International News	82.4	81.7	82
Economy	91.6	87.8	89.6
Sport	95.3	89.5	92.3
Average	**88.8**	**87.3**	**88**

Table (5.5) shows the classification of Alkhaleej News Dataset documents using BPSO+KNN as a feature selector and SVM as a classifier. The best F-Measure is for “Sport” class with 92.3, and the worst F-Measure is for “International News” class with 82. The next table, Table (5.6), describes the results of the Aljazeera News dataset by the same system.

Table (5.6): System **A** with SVM classifier applied on Aljazeera News dataset

Class	Precision%	Recall%	F1-Measure%
Politics	96.4	92.1	94.2
Science	79.1	94.5	86.1
Sport	95.7	96.2	95.9
Economy	69.8	88.2	77.9
Art	82.5	77.8	80
Average	**84.7**	**89.7**	**86.8**

Table (5.6) shows the classification of Aljazeera News dataset documents using BPSO+KNN as a feature selector and SVM as a classifier. We can see that the best F-Measure achieved by SVM classifier is for “sport” class with 95.9, then we have the second rank of measuring, the “Politics” class with F-Measure of 94.2 and the worst F-Measure is for “Economy” class with 77.9, that’s because of the low precision with 69.8. In general, SVM classifier does well when applied on Aljazeera News dataset after using feature selection by BPSO+KNN.

5.3 System A.$_C$(“BPSO+KNN”/REP-Tree)

The third combination of system **A** is our proposed classifier REP-Tree which has recently been used in English text classification as mentioned previously in the past chapters. Here, the REP-Tree is a classifier used to classify a group of feature resulting from the operation of features selection by BPSO+KNN. The experimental results of system **A** with REP-Tree classifier are shown by Tables(5.7), (5.8), and (5.9) using the previous three datasets (BBC Arabic, Aljazeera, and Alkhaleej datasets as the following:

Table (5.7): System **A** with REP-Tree classifier applied on BBC-Arabic Dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Middle East	87.7	91.5	89.5
World News	85.9	85.7	85.7
Business	86.1	90.6	88.2
Sport	80.3	72.2	76
Newspapers	89.2	88.7	88.9
Science	83.8	87.8	85.7
Misc.	79.2	72.3	75.5
Average	**84.6**	**84.1**	**84.2**

Table (5.7) shows the classification of BBC-Arabic documents using BPSO+KNN as a feature selector and REP-Tree as a classifier. As it is clear from Table (5.7), the results are as the following: the best classification is for "Middle East" class with precision of 87.7, recall of 91.5 and F1-Measure of 89.5. The second rank of performance is for classes "Newspapers" with precision of 89.2, recall of 88.7 and F1-Measure of 88.9. We can detect the convergence between the previous class performance and the "Business" class performance with precision of 86.1, recall of 90.6 and F1-Measure of 88.2. The worst performance was the "Misc." class with precision of 79.2, recall of 72.3 and F-Measure of 75.5. As in all previous experiments we'll apply the REP-Tree classifier on the other datasets. Now we will apply system **A (**the same previous experiment with REP-Tree**)** on the second data-set (Alkhaleej News Dataset) and Table (5.8) shows the results as the following:

Table (5.8): System **A** with REP-Tree classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	88.4	91.5	89.9
International News	93.2	85.2	89
Economy	80.1	83.6	81.8
Sport	92.7	82.7	87.4
Average	**88.6**	**85.7**	**87**

Accuracy results were comparable between REP-Tree and SVM with average F1-Measure of 87% for REP-Tree and 88% for SVM. For more details of the results the best F-Measure is for “Local News” class with 89.9, and the worst F-Measure is for “Economy” class with 81.8.

The next table, Table (5.9), describes the results of the Aljazeera News dataset by the same system.

Table (5.9): System **A** with REP-Tree classifier applied on Aljazeera News dataset

Class	Precision%	Recall%	F1-Measure%
Politics	60.7	71.4	65.6
Science	96.5	92.3	94.4
Sport	83.9	82.2	83
Economy	95.6	96.7	95.8
Art	88.1	89.7	88.8
Average	**84.9**	**86.4**	**85.5**

Table (5.9) shows the classification of Aljazeera News dataset documents using BPSO+KNN as a feature selector and REP-Tree as a classifier. We can see that the best F-Measure achieved by REP-Tree classifier is for Economy class with 95.8 then we have the second rank of measuring, the Science class with F-Measure of 94.4 and the worst F-Measure is for Politics class with 65.6 that’s because the low precision and recall with 60.7 and 71.4 .Also, here we can note that REP-Tree classifier do well when applied on Aljazeera News dataset after using feature selection by BPSO+KNN.

5.4 System B.$_A$ (“BPSO+SVM”/J 48)

The experimental results of system **B** with J 48 tree are shown by Tables (5.10), (5.11), and (5.12) using the previous three datasets (BBC-Arabic news dataset, Alkhaleej News dataset, and Aljazeera News dataset):

Table (5.10): System **B** with J 48 tree applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	70.4	72.6	71.4
World News	88.3	83.1	85.6
Business	77.5	71.2	74.2
Sport	87.7	78.5	82.8
Newspapers	85.2	87.3	86.2
Science	61	77.4	68.2
Misc.	82.5	87	84.6
Average	**78.9**	**79.5**	**79**

Table (5.10) shows the classification of BBC-Arabic documents using BPSO+SVM as a feature selector and J 48 decision tree as a classifier. As it is clear from the table, the results are as the following: the best classification performance is “Newspapers” class with precision of 85.2, recall of 87.3 and F1-Measure of 86.2. The second rank of classification performance is the “World News” with precision of 88.3, recall of 83.1 and F1-Measure of 85.6. We can see that the worst classes are the “Middle East” and the “Science” classes with precision of 70.4, recall of 72.6 and F-Measure of 71.4 for “Middle East” and the worst precision with 61.0 and F-Measure with 68.2 for “Science” class. Here we can be quite sure that the J 48 tree failed in the classification accuracy of “Science” class by 31.8% according to its F-Measure. Now we have the second data-set

(Alkhaleej News Dataset) with the same previous experiment, Table (5.11) shows the results as the following:

Table (5.11): System **B** with J 48 tree applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	49.8	52.4	51
International News	93.3	62.4	74.7
Economy	67.1	77.5	71.9
Sport	85.3	69.8	76.7
Average	**73.8**	**65.5**	**68.5**

Table (5.11) shows the classification accuracy of Alkhaleej News Dataset documents using BPSO+SVM as a feature selector and J 48 decision tree as a classifier. The best F-Measure is for "Sport" class with 76.6, and the worst F-Measure is for "Local News" class with 51. Also here we can be quite sure that the J 48 tree failed in the classification accuracy of "Sport" class by 49% according to its F-Measure.

The next table, Table (5.12), describes the results of the Aljazeera News dataset by the same system.

Table (5.12): System **B** with J 48 tree applied on Aljazeera News dataset

Class	Precision%	Recall%	F1-Measure%
Politics	52.5	79.7	63.3
Science	97.2	45.2	61.7
Sport	98.9	59.1	73.9
Economy	84.6	40.8	55
Art	39.8	70.7	50.9
Average	**74.6**	**59.1**	**60.9**

Table (5.12) shows the classification accuracy of Aljazeera News dataset documents using BPSO+SVM as a feature selector and J 48 decision tree as a classifier. The best F-Measure achieved by J 48 is for "Sport" class with 73.9 then we have the second rank of measuring, the “Politics” class with F-Measure of 63.3 and the worst F-Measure is for “Art” class with 50.9.

5.5 System B.$_B$ (“BPSO+SVM”/SVM)

The experimental results of system **B** with SVM classifier are shown by Tables(5.13), (5.14) and (5.15) using the previous three datasets (BBC Arabic, Aljazeera, and Alkhaleej datasets as the following:

Table (5.13): System **B** with SVM classifier applied on BBC-Arabic Dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Middle East	67.9	88.7	76.9
World News	98.7	90.3	94.3
Business	87.9	89.3	88.5
Sport	60.3	80.7	69
Newspapers	79.8	84.2	81.9
Science	99.2	85.6	91.8
Misc	90.4	98.8	94.4
Average	**83.4**	**88.2**	**85.2**

Table (5.13) shows the classification of BBC-Arabic documents using BPSO+KNN as a feature selector and SVM as a classifier. As it is clear from Table (5.13), the results are as the following: the best classification is for “Misc.” class with precision of 90.4, recall of 98.8 and F1-Measure of 94.4. The second performance rank of classes is the “World News” with precision of 98.7, recall of 90.3 and F1-Measure of 94.3. The worst class is the “Sport” with precision of 60.3, recall of 80.7 and F-Measure of 69. Now we will apply system B **(**the same previous experiment with SVM**)** on the second data-set (Alkhaleej News Dataset), and Table (5.14) shows the results as the following:

Table (5.14): System **B** with SVM classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	83.2	88.6	85.8
International News	88.5	85.7	87
Economy	96.6	90.9	93.6
Sport	90.3	89.7	89.9
Average	**89.6**	**88.7**	**89**

Table (5.14) shows the classification of Alkhaleej News Dataset documents using BPSO+SVM as a feature selector and SVM as a classifier. The best accuracy (F-Measure) is for Economy class with 93.6 and the worst F-Measure is for "Local News" class with 85.8. The next table, Table (5.15), describes the results of the Aljazeera News dataset by the same system.

Table (5.15): System **B** with SVM classifier applied on Aljazeera News dataset

Class	Precision%	Recall%	F1-Measure%
Politics	77.3	92.8	84.3
Science	92.9	96.6	94.7
Sport	94.3	94.5	94.3
Economy	90.8	87.4	89
Art	91.5	80.9	85.8
Average	**89.3**	**90.3**	**89.6**

Table (5.15) shows the classification of Aljazeera News dataset documents using BPSO+SVM as a feature selector and SVM as a classifier. We note that the best F-Measure achieved by SVM classifier is for "Science" class with 94.7then we have the second rank of measuring, the "sport" class with F-Measure of 94.3 and the worst F-Measure is for "Politics" class with 84.3, that's because the low precision with 77.3. Also here we can say that SVM classifier do well when applied on Aljazeera News dataset after using feature selection by BPSO+SVM.

5.6 System B.$_C$ (“BPSO+SVM”/REP-Tree)

The third combination of system **B** is our proposed classifier REP-Tree as we mentioned in the previous experiments which has recently been used by (Kalmegh, 2015), (Patel and Upadhyay, 2012) in English text classification and by (Naji and Ashour, 2016) in Arabic text classification (a previous paper related to the thesis), as mentioned previously in the past chapters specifically in the first chapter. Here the REP-Tree is a classifier, which is used to classify a group of features resulting from the operation of features selection by BPSO+SVM. The experimental results of system **B** with REP-Tree classifier are shown by Tables (5.16), (5.17), and (5.18) using the previous three datasets (BBC Arabic, Aljazeera, and Alkhaleej datasets as the following:

Table (5.16): System **B** with REP-Tree classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	77	89.4	82.7
World News	98.3	96.1	97.1
Business	87.2	78.5	82.6
Sport	79.5	75.8	77.6
Newspapers	88.2	88.9	88.5
Science	85.4	87.1	86.2
Misc	89	69.4	77.9
Average	**86.3**	**83.6**	**84.6**

Table (5.16) shows the classification of BBC-Arabic documents using BPSO+SVM as a feature selector and REP-Tree as a classifier. As it is clear from Table (5.16), the results are as the following: the best classification is for “World News” class with precision of 98.3, recall of 96.1 and F1-Measure of 97.1. The second rank of performance is for classes “Newspapers” with precision of 88.2, recall of 88.9 and F1-Measure of 88.5. We can detect the convergence between the “Middle East” class performance and the “Business” class performance with F1-Measure of 82.7 and 82.6. The worst performance was the “Sport” class with precision of 79.5, recall of 75.8 and F-Measure of 77.6. As in all previous experiments we'll apply the REP-Tree classifier on the other datasets. Now we will apply system **B (**the same previous experiment with REP-Tree**)** on the second data-set (Alkhaleej News Dataset), and Table (5.17) shows the results as the following:

Table (5.17): System **B** with REP-Tree classifier applied on Alkhaleej News Dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Local News	72	78.3	75
International News	89.6	92.2	90.8
Economy	87.3	88.3	87.7
Sport	95.4	87.5	91.2
Average	**86**	**86.5**	**86.1**

From Table (5.17) we see that the best accuracy of REP-Tree F1-Measure is 91.2 for “Sport” class, and the worst F-Measure is for “Local News” class with 75. We note that the results were comparable with SVM classifier. Now we will apply the REP-Tree on another data-set. The next table, Table (5.18), describes the results of the Aljazeera News dataset by the same system.

Table (5.18): System **B** with REP-Tree classifier applied on Aljazeera News dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Politics	96	98.7	97.3
Science	74.5	70.9	72.6
Sport	87.9	85.2	86.5
Economy	98.2	99.1	98.6
Art	90.3	95.5	92.8
Average	**89.3**	**89.8**	**89.5**

Table (5.18) shows the classification of Aljazeera News dataset documents using BPSO+SVM as a feature selector and REP-Tree as a classifier. We can see that the comparable accuracy or the(equality in classification average F-Measure)between REP-Tree and SVM (Table 5.15) on the same data-set (Aljazeera News dataset), with the average F-Measure of 89.5 for REP-Tree and 89.6 for SVM. The REP-Tree classifier outperforms the other classifiers in time needed to build and split the model of training data from the original dataset. Because of the short time it takes the REP-Tree to build a model of learning, from our point of view, we can overlook the difference between two classifiers.

5.7 System C.$_{A}$ (“BPSO+REP-Tree”/J 48)

System C consists of Binary PSO as a feature selector and the proposed REP-Tree as an evaluator to check the best group of features then we use the three previous classifiers (J 48, SVM, and REP-Tree) to build the classification model; the classification in the resultant group of features in the training set to reduce the dimension of the original data-set and then apply the classifiers on the test data-set. We have previously noted that REP-Tree has recently been used by (Kalmegh, 2015), (Patel and Upadhyay, 2012) to classify English text and by (Naji and Ashour, 2016) in Arabic text classification.

The experimental results of system **C** with J 48 tree are shown by Tables (5.19), (5.20), and (5.21) using the previous three datasets (BBC-Arabic news dataset, Alkhaleej News dataset, and Aljazeera News dataset):

Table (5.19): System **C** with J 48 tree applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	88.7	83.3	85.9
World News	90.4	87.4	88.8
Business	75.2	70.5	72.7
Sport	84.8	74.2	79.1

Newspapers	80.1	83.8	81.9
Science	79.8	78.3	79
Misc.	77.6	85.7	81.4
Average	**82.3**	**80.4**	**81.2**

Table (5.19) shows the classification of BBC-Arabic documents using BPSO+REP-Tree as a feature selector and J 48 decision tree as a classifier. As it is clear from the table, the results are as the following: the best classification performance is “World News” class with precision of 90.4, recalling of 87.4 and F1-Measure of 88.8. The second rank of classification performance is the “Middle East” with precision of 88.7, recall of 83.3 and F1-Measure of 85.9. We can note that the worst class was the “Business” class with precision of 75.2, recall of 70.5 and F-Measure of 72.7. Here we can be quite sure that the J 48 tree failed in the classification accuracy of “Science” class by 27.3% according to its F-Measure.

Now we have the second data-set (Alkhaleej News Dataset) with the same previous experiment, Table (5.20) shows the results as the following:

Table (5.20): System **C** with J 48 tree applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	60.3	56.8	58.4
International News	68.6	70.9	69.7
Economy	90.4	75.9	82.5
Sport	84.8	72.5	78.1
Average	**73.5**	**69**	**72.1**

Table (5.20) shows the classification accuracy of Alkhaleej News Dataset documents using BPSO+REP-Tree as a feature selector and J 48 decision tree as a classifier. The best F-Measure is for “Economy” class with 82.5, and the worst F-Measure is for “Local News” class with 58.4. Also here we can be quite sure that the J 48 tree failed in the classification accuracy of “Local News” class by 47.6% according to its F-Measure. The next table, Table (5.21), describes the results of the Aljazeera News dataset by the same system.

Table (5.21): System **C** with J 48 tree applied on Aljazeera News dataset

Class	Precision%	Recall%	F1-Measure%
Politics	39.2	70.7	50.4
Science	90.7	70.5	79.3
Sport	45.8	56.3	50.5
Economy	86.2	49.6	62.9
Art	97.9	66.8	79.4
Average	**71.9**	**62.7**	**64.5**

Table (5.21) shows the classification accuracy of Aljazeera News dataset documents using BPSO+REP-Tree as a feature selector and J 48 decision tree as a classifier. The best F-Measure achieved by J 48 is for both “Art” and “Science” classes with 79.4 and 79.3 gradually, and the worst F-Measure which is for “Politics” and “Sport” class with 50.4 and 50.5.

5.8 System C.$_{B}$ (“BPSO+REP-Tree”/SVM)

The experimental results of system **C** with SVM classifier are shown by Table (5.22), (5.23) and (5.24) using the previous three datasets (BBC Arabic, Aljazeera, and Alkhaleej datasets as the following:

Table (5.22): System **C** with SVM classifier applied on BBC-Arabic Dataset

Class	Precision%	Recall%	F1-Measure%
Middle East	98.6	94.4	96.4
World News	68.2	88.9	77.1
Business	82.3	85.7	83.9
Sport	64.6	78.5	70.8
Newspapers	81.4	82.8	82

Science	97.2	87.1	91.8
Misc.	92.5	96.9	94.6
Average	**83.5**	**87.7**	**85.2**

Table (5.22) shows the classification of BBC-Arabic documents using BPSO+REP-Tree as a feature selector and SVM as a classifier. From Table (5.22) we note the equality in F-Measure average value using the same classifier SVM with a different features selection combination (BPSO+REP-Tree). The current results have been compared with Table (5.13) (BPSO+SVM features selection). We get here an average F-Measure of 85.2 and 89.6 for SVM (the same classifier but different feature selector). As usual, we will apply system C **(**the same previous experiment with SVM**)** on the second data-set (Alkhaleej News Dataset), and Table (5.23) shows the results as the following:

Table (5.23): System **C** with SVM classifier applied on Alkhaleej News Dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Local News	97.2	93.7	95.4
International News	94.5	82.9	88.3
Economy	90.3	95.5	92.8
Sport	79.5	80	79.7
Average	**90.3**	**88**	**89.05**

Table (5.23) shows the classification of Alkhaleej News Dataset documents using BPSO+REP-Tree as a feature selector and SVM as a classifier. The best accuracy (F-Measure) is for “Local News” class with 95.4 and the worst F-Measure is for “Sport” class with 79.7. In this experiment, we note the equality and convergence in the classification process results using the same classifier SVM with a different features selection combination (BPSO+REP-Tree).

As usual we will apply system C **(**the same previous experiment with SVM classifier**)** on the third data-set (Aljazeera News Dataset), and Table (5.24) shows the results as the following:

Table (5.24): System **C** with SVM classifier applied on Aljazeera News dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Politics	97	90.6	93.6
Science	91.6	98.3	94.8
Sport	75.3	85.1	79.9
Economy	92.5	92.7	92.5
Art	90.7	80.8	85.4
Average	**89.4**	**89.5**	**89.2**

Table (5.24) shows the classification of Aljazeera News dataset documents using BPSO+REP-Tree as a feature selector and SVM as a classifier. We note that the best F-Measure achieved by SVM classifier is for “Science” class with 94.8, then we have the second rank of measuring, the “Politics” class with F-Measure of 93.6 and the worst F-Measure is for “Sport” class with 79.9.

5.9 System C.$_C$ (“BPSO+REP-Tree”/REP-Tree)

The third combination of system C consists of Binary PSO as a feature selector and the proposed REP-Tree as an evaluator then we use REP-Tree as a classifier, as we mentioned in the previous chapter System C subsection. The experimental results of system **C** with REP-Tree classifier are shown by Tables (5.25), (5.26), and (5.27) using the previous three datasets (BBC Arabic, Aljazeera, and Alkhaleej datasets as the following:

Table (5.25): System **C** with REP-Tree classifier applied on BBC-Arabic Dataset

Class	**Precision%**	**Recall%**	**F1-Measure%**
Middle East	97.2	95.3	96.2
World News	88.6	78.5	83.2
Business	87.3	88.6	87.9
Sport	79.9	75.9	77.8
Newspapers	86.1	98.4	91.8
Science	80	86.9	83.3
Misc.	82.5	92	86.9
Average	**85.9**	**87.9**	**86.7**

Table (5.25) shows that the REP-Tree has been effective enough in the classification for BBC-Arabic documents using BPSO+REP-Tree as a feature selector and REP-Tree as a classifier. The results are as following: the best classification is for “Middle East” class with precision of 97.2, recall of 95.3 and F1-Measure of 96.2. Next we have the second classification performance the “Newspapers” with precision of 86.1, recalling of 98.4 and F1-Measure of 91.8. The third classification accuracy is the “Business” with F-Measure of 87.9. We can detect the convergence between the “Science” class performance and the “World News” class performance with F1-Measure of 83.3 and 83.2. The worst performance was the Sport class with F-Measure of 77.8.

As usual, we will apply the REP-Tree classifier on the other datasets. Now, we will apply system **C,** the same previous experiment with REP-Tree**,** on the second data-set (Alkhaleej News Dataset), and Table (5.26) shows the results as the following:

Table (5.26): System **C** with REP-Tree classifier applied on Alkhaleej News Dataset

Class	Precision%	Recall%	F1-Measure%
Local News	98	97.4	97.6
International News	91.3	92.5	91.8
Economy	85.7	87.1	86.3
Sport	93.8	89.6	91.6
Average	**92.2**	**91.6**	**91.8**

From Table (5.26), we see that the best accuracy of REP-Tree (F1-Measure) is 97.6 for "Local News" class, and the worst F-Measure is for "Economy" class with 86.3. The average accuracy of the REP-Tree in this experiment was 91.8.

Now, we will apply the REP-Tree classifier on the last data-set. The next table, Table (5.27), describes the results of the last experiment, the Aljazeera News dataset by the same system.

Table (5.27): System **C** with REP-Tree classifier applied on Aljazeera News dataset

Class	Precision%	Recall%	F1-Measure%
Politics	90.7	94.6	92.6
Science	93.5	90.7	92
Sport	70.2	67	68.5
Economy	98.6	97.5	98
Art	96.2	95.8	95.9
Average	**89.8**	**89.3**	**89.4**

Table (5.27) shows the classification of Aljazeera News dataset documents using BPSO+REP-Tree as a feature selector and REP-Tree as a classifier. We can see that the REP-Tree classifier works well in classifying Arabic data-set (Aljazeera News dataset) with average Precision of 89.8, Recall of 89.3 and F-Measure of 89.4 compared with the existing feature selection combinations like (BPSO+KNN) and (BPSO+KNN)where applied on Aljazeera News dataset. The REP-Tree classifier also outperforms the other classifiers in time needed to build and split the model of training data from the original dataset.

5.10 Performance of the Three Systems

In this subsection, we will make a comparison between the previous results on the previous three datasets (BBC-Arabic, Alkhaleej, and Aljazeera dataset) before adding some enhancements to each system in the preprocessing phase. Both Table (5.28) and Figure (5.1) show the results of this comparison.

Table (5.28): Comparison between the F-Measure averages of the three systems

Datasets	**System A (BPSO+KNN)%**	**System B (BPSO+SVM)%**	**System C (BPSO+REP-Tree)%**
BBC-Ar (J48)	79	79	81.2
BBC-Ar (SVM)	86	85.2	85.2
BBC-Ar (REP)	84.2	84.6	86.7
Alkhaleej(J48)	74.3	68.5	72.1
Alkhaleej(SVM)	88	89	89
Alkhaleej(REP)	87	86.1	91.8
Aljazeera (J48)	73.9	60.9	64.5
Aljazeera(SVM)	86.8	89.6	89.2
Aljazeera (REP)	85.5	89.5	89.4

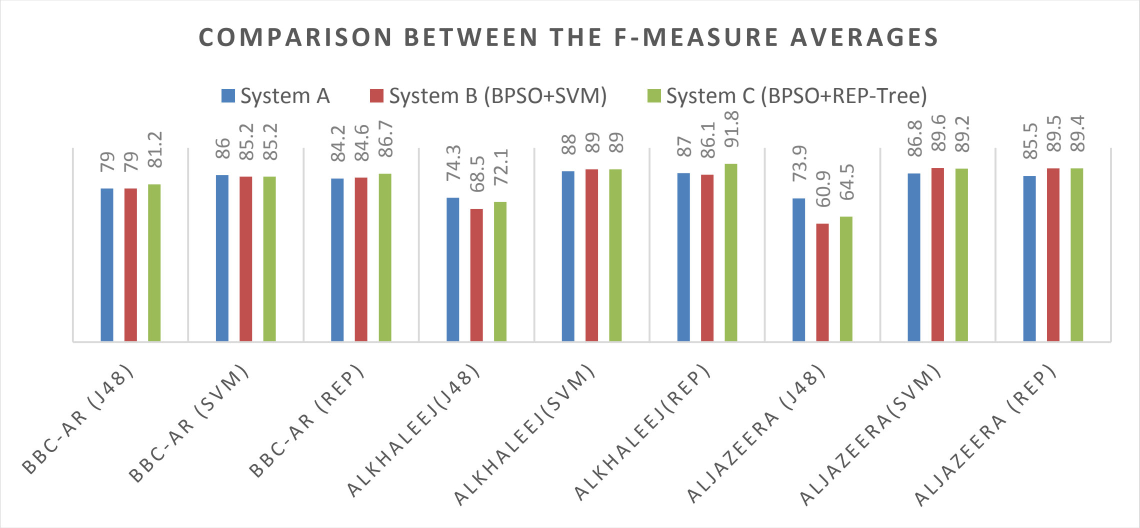


Figure (5.1): Comparison between the accuracy of the three systems

From Table (5.28) and Figure (5.1), we draw the overall results of all the experiments, calculate the average for F1-Measure values, and compare all the systems with each other.

Now we try to apply the same experiment on part C of every system, but with reducing the dataset to the half to study the effect of data size when use REP-Tree as an evaluator and a classifier. Now we chose part C to show the using of REP-Tree, note that system:

A.c is the BPSO+KNN/ REP-Tree. (REP-Tree here is a classifier).

B.c is the BPSO+SVM/ REP-Tree. (REP-Tree here is a classifier).

C.c is the BPSO+REP-Tree/ REP-Tree. (REP-Tree here is an evaluator and a classifier).

Results are shown in the following table, Table (5.29).

Table (5.29): Comparison between the accuracy of A.c, B.c, and C.c

Datasets	System A (BPSO+KNN)%	System B (BPSO+SVM)%	System C (BPSO+REP-Tree)%
BBC-Ar (J48)	79	79.3	81.2
BBC-Ar (SVM)	86	85.2	85.2
BBC-Ar (REP)	85.9	84.9	88.2
Alkhaleej(J48)	74.3	78.5	72.1
Alkhaleej(SVM)	88	89	89
Alkhaleej(REP)	89.1	88.2	91.9
Aljazeera (J48)	73.9	60.9	64.5
Aljazeera(SVM)	86.8	89.6	89.2
Aljazeera (REP)	85.7	89.5	91.5

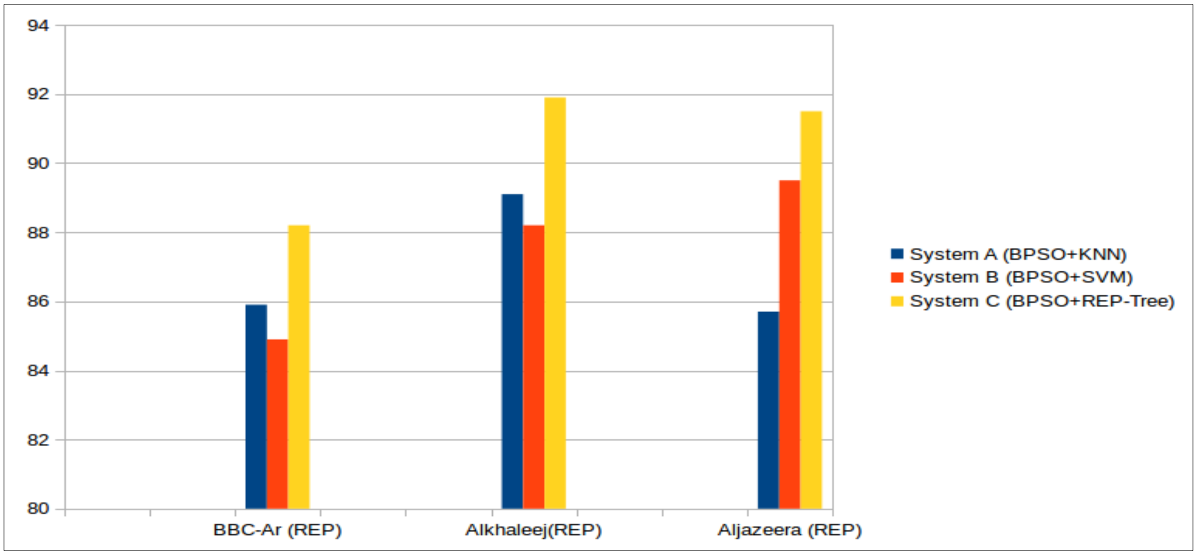


Figure (5.2): Comparison between the accuracy (F1-Measure) of A.C, B.C, and C.C

Both Figure (5.2) and Table (5.29) draw the overall results of all the experiments of system A.C, B.C, and C.C, we note that the accuracy of REP-Tree increased when we reduced the amount of dataset. System C,c with the green color values outperforms the other systems, but also the three systems accuracy is increased. So REP-Tree do well with features selection phase (as evaluator with BPSO). In addition we note that there is a direct correlation between the decrease in the volume of data and the effectiveness of REP-Tree.

Next, we will review the impact of adding some improvements -so to speak- on these systems with the original datasets and come up with new results.

5.11 Performance of the Three Systems with Adding Enhancements

We are previously talked about adding some enhancements such as Stemming and Normalization to the preprocessing phase and studying the effects of these additions. Both of Table (5.30) and Figure (5.3) show the results of the previous experiments of the three systems A, B and C after adding the aforementioned enhancements.

Table (5.30): Comparison between the accuracy of the three systems after enhancements

Datasets	**System A (BPSO+KNN)%**	**System B (BPSO+SVM)%**	**System C (BPSO+REP-Tree)%**
BBC-Ar (J48)	82	79	77.6
BBC-Ar (SVM)	87.3	86.4	87
BBC-Ar (REP)	86.2	85.3	86.9
Alkhaleej(J48)	72.3	76.7	71.5
Alkhaleej(SVM)	87.7	95.8	89.8
Alkhaleej(REP)	87.6	88.9	94.7
Aljazeera (J48)	74.4	69.9	67.5
Aljazeera (SVM)	85.2	86.1	89.2
Aljazeera (REP)	85.9	90.4	89.4

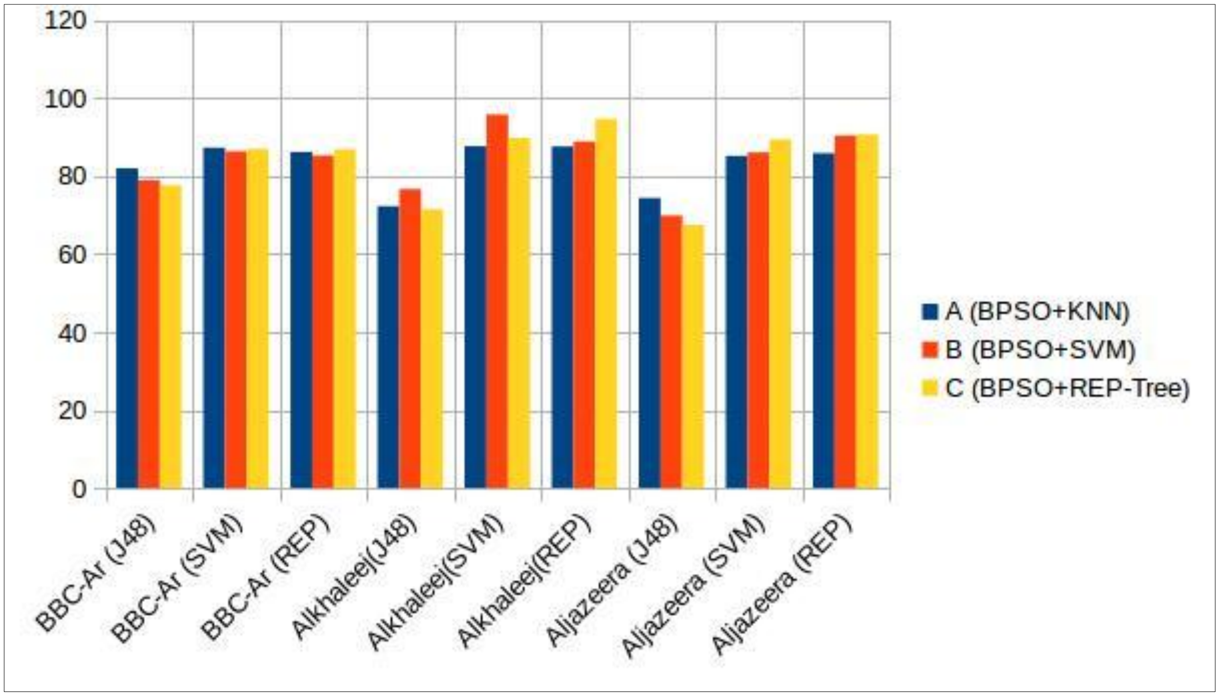


Figure (5.3): Comparison between the accuracy of the three systems after enhancements.

Both Table (5.30) and Figure (5.3) show the effect of adding the previous improvements to all experiments which are as following. We begin with system **A,** we note a rise in classification accuracy for each of (BBC-Ar (J48), BBC-Ar (SVM), BBC-Ar (REP), Alkhaleej (REP), Aljazeera (J48) and Aljazeera (REP)) and we observe a decrease in classification accuracy for each of (Alkhaleej (J48), Alkhaleej (SVM) and Aljazeera (SVM)).

For system **B** we note a rise in classification accuracy for each of (BBC-Ar (SVM), BBC-Ar (REP), Alkhaleej(SVM), Alkhaleej(REP), Aljazeera (J48), and Aljazeera (REP)) and we observe a decrease in classification accuracy for each of (BBC-Ar (J48), Alkhaleej(J48), and Aljazeera (SVM)).

Finally, we have a system **C.** we note a rise in classification accuracy for each of (BBC-Ar (SVM), BBC-Ar (REP), Alkhaleej (SVM), Alkhaleej (REP), and Aljazeera (J48)), and we observe a decrease in classification accuracy for each of (BBC-Ar (J48) and Alkhaleej(J48)) and we did not notice any changes in the value of each of Aljazeera (SVM) and Aljazeera (REP).

So we cannot say that the impact of enhancements was absolutely well because there are some changes in some of the values for the better and some for the worse and cannot adopt these additions unless there is an improvement of results at all. At the end, we can say that the increase in accuracy was more noticeable, and cannot neglect the use of some of the previous common enhancements such as stemming.

5.12 Summary

This chapter presented and clarified the previous experiments in the previous chapter (Chapter 4), also draw conclusions and comparisons with the previous three systems or the nine subsystems. The division of work is on three data sets, BBC-Arabic News Data set, Alkhaleej News Data set and Aljazeera News Data set using three classifiers SVM, KNN and the proposed classifier REP-Tree. Results demonstrated that the proposed classifier REP-Tree is promising in the field of classification of Arabic texts especially when combined with BPSO and worked as an evaluator to check the best subset of features (Terms) and also worked well when used as a final classifier with other systems of features selection like BPSO+KNN and BPSO +SVM after comparing it with other similar classifiers, J 48 and SVM.

Chapter 6

Conclusion

Chapter 6

Conclusion

This thesis proposed a new feature selection approach to select the best subset of features from the original Arabic document .We showed that the proposed approach works well in this area after extracting the experimental results. The proposed approach can be used in the field of Arabic search engines and classifying huge amounts of Arabic websites pages into hierarchal classes, labels.

We proposed the Reduced Error Pruning-Tree classifier, which was not used with Arabic text classification before for two purposes. The first one is an evaluator to evaluate the subset of features that resulted from the features selection algorithm Binary Particle Swarm Optimization BPSO. To evaluate this approach (BPSO+REP-Tree), we used three Arabic datasets, BBC-Arabic News dataset, Alkhaleej News dataset, and Aljazeera News dataset. The second purpose of the Rep-Tree is to use it as a classifier to build the learning model. We compare the first purpose (BPSO+REP-Tree) with two existing approaches, (BPSO+KNN) and (BPSO+SVM), and the second purpose (REP-Tree classifier) with two well-known classifiers, J 48 and SVM. We named the three features selection approaches with **A** for (BPSO+KNN), **B** for (BPSO+SVM), and **C** for (BPSO+REP-Tree)**.** After we get the experimental results, we concluded that the proposed approach System C is effective. We choose the F1-Measure to estimate the accuracy of the classification process which came from two factors, precision and recall factors.

The values of F1-Measure for system A with the classifier J 48 is in the range of 73% - 79%, with SVM is in 86% - 88% and with the proposed classifier REP-Tree is in the range of 84% - 87%. Next, we get the F1-Measure values of the second system (B) with the same classifiers as the following, with J 48 are in the range of 60.9% - 84.6%, with SVM is in 85.2% - 89.6% and with the proposed classifier REP-Tree is in the range of

84.6% - 89.5% and here is the last two algorithms which are comparable in the accuracy. Finally, we apply the experiments on our proposed approach system (C) in features selection domain and it gave these ranges of accuracy as the following, with J 48 was in the range of 69.5% - 79.6%, and with SVM is in 87% - 89.8% and with the proposed classifier REP-Tree is in the range of 86.7% - 91.8%.

Last but not least we added Stemming by Light Stemming and Normalization and studied the effects of them, where most of the classifiers accuracies were enhanced and some of them still without any changes. We finally studied the effect of reducing the three dataset to the half; the experimental results showed that the REP-Tree outperformed the others in this case. In general, a new approach of features selected had been added and it worked well in this domain as we showed by results.

REFERENCES

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523. doi:10.1016/0306-4573(88)90021-0

Li, S., Xia, R., Zong, C., & Huang, C. (2009).*A framework of feature selection methods for text categorization.* Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09. doi:10.3115/1690219.1690243

Saeys, Y., Inza, I., & Larranaga, P. (2007).A review of feature selection techniques in bioinformatics.*Bioinformatics, 23*(19), 2507-2517. doi:10.1093/bioinformatics/btm344

Al-Tahrawi, M. M., & Al-Khatibb, S. N. (2015).Arabic text classification using Polynomial Networks.*Journal of King Saud University - Computer and Information Sciences, 27*(4), 437-449. http://dx.doi.org/10.1016/j.jksuci.2015.02.003

Aimunandar, T., & Winarko, E. (2015). Regional Development Classification Model using Decision Tree Approach. *International Journal of Computer Applications IJCA, 114*(8), 29-34. doi:10.5120/20000-1755

Kalmegh, S. (2015).Analysis of WEKA Data Mining Algorithm REP-Tree, Simple Cart and RandomTree for Classification of Indian News.*PARIPEX Paripex - Indian Journal of Research, 2*(2), 438-446. doi:10.15373/22501991/feb2015.

L.gupta, D., Malviya, A. K., & Singh, S. (2012). Performance Analysis of Classification Tree Learning Algorithms. *International Journal of Computer Applications IJCA, 55*(6), 39-44. doi:10.5120/8762-2680

Patel, N., & Upadhyay, S. (2012). Study of Various Decision Tree Pruning Methods with their Empirical Comparison. *International Journal of Computer Applications, 60*(12), 20-25. doi:10.5120/9744-4304

Duwairi, R., Al-Refai, M. N., & Khasawneh, N. (2009). Feature reduction techniques for Arabic text categorization. *J. Am. Soc. Inf. Sci. Journal of the American Society for Information Science and Technology, 60*(11), 2347-2352. doi:10.1002/asi.21173

Duwairi, R., & El-Orfali, M. (2014).A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science, 40*(4), 501-513. doi:10.1177/0165551514534143

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña-López, L. A., & Perea-Ortega, J. M. (2011). OCA: Opinion corpus for Arabic. J. Am. Soc. Inf. Sci. *Journal of the American Society for Information Science and Technology, 62*(10), 2045-2054. doi:10.1002/asi.21598

Khreisat, L. (2009). A machine learning approach for Arabic text classification using N-gram frequency statistics. *Journal of Informatics, 3*(1), 72-77. doi:10.1016/j.joi.2008.11.005

El-Halees, A. M. (2007).Arabic text classification using maximum entropy.*TheIslamic University Journal (Series of Natural Studies and Engineering), 15*(2), 157-167.

Kourdi, M. E., Bensaid, A., & Rachidi, T. (2004).*Automatic Arabic document categorization based on the Naïve Bayes algorithm.* Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages - Semitic '04. doi:10.3115/1621804.1621819

Sawaf, H., Zaplo, J. I., & Ney, H. (2001).*Statistical Classification Methods for Arabic News Articles.* Workshop on Arabic Natural Language Processing, ACL'01, 177(5), 1083-1089. Retrieved June 26, 2016, from http://www.abdelali.net/ref/Sawaf_ArabicClassification.pdf

Al-Zoghby, A., Eldin, A. S., Ismail, N. A., & Hamza, T. (2007).*Mining Arabic text using soft-matching association rules.*Paper presented atInternational Conference on Computer Engineering & Systems. doi:10.1109/icces.2007.4447080

Harrag, F., El-Qawasmeh, E., & Pichappan, P. (2009).*Improving Arabic text categorization using decision trees.* Paper presented at First International Conference on Networked Digital Technologies. doi:10.1109/ndt.2009.5272214

Al-Shalabi, R., Kanaan, G., & Gharaibeh, M. H. (2006).Arabic Text Categorization Using KNN Algorithm. *University of California Irvin Data Collections Repository.*

Brahimi, B., Touahria, M., & Tari, A. (2016).Data and Text Mining Techniques for Classifying Arabic Tweet Polarity. *Journal of Digital Information Management*, *14*(1), 12-19.

Oraby, S. M., El-Sonbaty, Y., El-Nasr, M. A. (2013).*Exploring the Effects of Word Roots for Arabic Sentiment Analysis.* In International Joint Conference on Natural Language Processing, Nagoya, Japan (471-479).

Khoja, S. (1999).*Stemming Arabic Text, Lancaster, U.K, Computing Department, Lancaster University.*

Taghva, K., Elkhoury, R., & Coombs, J. (2005).*Arabic stemming without a root dictionary.* Paper presented at International Conference on Information Technology: Coding and Computing (ITCC'05). doi:10.1109/itcc.2005.90

Tashaphyne. (2010). *Arabic light stemmer, 0.2.* Available at https://pypi.python.org/pypi/Tashaphyne

Shoukry, A., & Rafea, A. (2012).*Sentence-level Arabic sentiment analysis.* International Conference on Collaboration Technologies and Systems (CTS). doi:10.1109/cts.2012.6261103

Al-Thwaib, E. (2014). Text Summarization as Feature Selection for Arabic Text Classification. *World of Computer Science and Information Technology Journal (WCSIT), 4*(7), 101-104.

Sakhr company website: Retrieved September 27, 2016, from http://www.sakhr.com/index.php/en/about-sakhr/overview.

Al-Hindi, K., & Al-Thwaib, E. (2013). A Comparative Study of Machine Learning Techniques in Classifying Full-Text Arabic Documents versus Summarized Documents. *World of Computer Science and Information Technology Journal (WCSIT), 2*(7), 126-129. Retrieved August17, 2016, from http://www.wcsit.org/pub/2013/vol.3.no.7/A Comparative Study of Machine Learning Techniques in Classifying Full-Text Arabic Documents versus Summarized Documents.pdf

Sallam, R., Mousa, H., & Hussein, M. (2016).Improving Arabic Text Categorization using Normalization and Stemming Techniques. *International Journal of Computer Applications IJCA, 135*(2), 38-43. doi:10.5120/ijca2016908328

Abu-Errub, A. (2014). Arabic Text Classification Algorithm using TF.IDF and Chi Square Measurements. *International Journal of Computer Applications IJCA, 93*(6), 40-45. doi:10.5120/16223-5674

Goweder, A., Elboashi, M., & Elbekai, A. (2013).*Centroid-Based Arabic Classifier.* The International Arab Conference on Information Technology (ACIT'2013), 108(3). Retrieved June 27, 2016, from http://acit2k.org/ACIT/2013Proceedings/108.pdf

Abidi, K., & Elberrichi, Z. (2012). Arabic Text Categorization: A Comparative Study of Different Representation Modes. *Journal of Theoretical and Applied Information Technology, 38*(1), 465-470. Retrieved May 21, 2016, from http://ccis2k.org/iajit/PDF/vol.9,no.5/2983-10.pdf

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. doi:10.14569/ijacsa.2015.060228

Mohammad, A. H., Al-Momani, O., & Alwada, T. (2016). Arabic Text Categorization using k-nearest neighbor, Decision Trees (C4.5) and Rocchio Classifier: A Comparative Study. *International Journal of Current Engineering and Technology,* *6*(2), 477-482. RetrievedMay29, 2016,from http://inpressco.com/wp-content/uploads/2016/03/Paper16477-482.pdf

Hussien, M. I., Olayah, F., AL-dwan, M., & Shamsan, A. (2011).Arabic text classification using SMO, naïve Bayesian, J48 Algorithms. *International Journal of Research and Reviews in Applied Sciences*, *9*(2), 306-316. Retrieved June 21, 2016, from http://www.arpapress.com/volumes/vol9issue2/ijrras_9_2_15.pdf

Kanan, T., & Fox, E. A. (2015).Automated Arabic text classification with P-Stemmer, machine learning, and tailored news article taxonomy. *Journal of the Association for Information Science and Technology J Assn Inf Sci Tec.* doi:10.1002/asi.23609

Gharib, T. F., Habib, M. B., & Fayed, Z. T. (2009).*Arabic Text Classification Using Support Vector Machines.* Faculty of Computer and Information Sciences, Ain Shams University, 1-8. Retrieved August 20, 2016, from http://wwwhome.cs.utwente.nl/~badiehm/PDF/ISCA2009.pdf

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., & Al-Rajeh, A. (2008).*Automatic Arabic Text Classification.* JADT 2008: 9 Es Journées Internationales D'Analyse Statistique Des Données Textuelles, 77-83. Retrieved May 1, 2016, from http://eprints.soton.ac.uk/272254/1/Arabic-Classification.pdf

Al-Anzi, F. S., & Abuzeina, D. (2016).Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. *Journal of King Saud University - Computer and Information Sciences.* doi:10.1016/j.jksuci.2016.04.001

Zubi, Z. S. (2009). *Using Some Web Content Mining Techniques for Arabic Text Classification.* Recent Advances on Data Networks, Communications, Computers, 73-84. doi:10.1109/mmcs.2009.5256696

Zrigui, M., Ayadi, R., Mars, M., & Maraoui, M. (2012). Arabic Text Classification Framework Based on Latent Dirichlet Allocation. *Journal of Computing and Information Technology CIT, 20*(2), 11-14. doi:10.2498/cit.1001770

Hotho, A., Nurnberger, A., & Paab, G. (2005).A brief survey of text mining.*LDV Forum-GLDV Journal for Computational Linguistics and Language Technology, 20,* 19-62. Retrieved from http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf

Witten, I. H., & Frank, E. (2005).*Data mining: Practical machine learning tools and techniques.* Amsterdam: Morgan Kaufman.

Singh, M. P. (2005). *The practical handbook of Internet computing.* Boca Raton: Chapman & Hall/CRC.

Sebastiani, F. (2002).*Machine learning in automated text categorization* (1st ed), PP.1-47. ACM computing surveys (CSUR).

Wajeed, M. A., & Adilakshmi, T. (2009).Text Classification Using Machine Learning. *Journal of Theoretical and Applied Information Technology, 7*(2), 119-123.

Sebastiani, F. (2006).*Classification of text, automatic*. The Encyclopedia of Language and Linguistics (1st ed., Vol. 14, Pp.457-462).

Chen, H., & Ho, T. K. (1999).*Evaluation of decision forests on text categorization.* Document Recognition and Retrieval VII, 1-9. doi:10.1117/12.373494

Forman, G. (2002). *Choose Your Words Carefully: An Empirical Study of Feature Selection Metrics for Text Classification.* Principles of Data Mining and

Knowledge Discovery Lecture Notes in Computer Science, 150-162.doi:10.1007/3-540-45681-3_13

Duwairi, R., Al-Refai, M., & Khasawneh, N. (2007).*Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization.*2007 Innovations in Information Technologies (IIT). doi:10.1109/iit.2007.4430403

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007).*Light Stemming for Arabic Information Retrieval.* Text, Speech and Language Technology Arabic Computational Morphology, 221-243. doi:10.1007/978-1-4020-6046-5_12

Chantar, H. K., & Corne, D. W. (2011). *Feature subset selection for Arabic document categorization using BPSO-KNN.* Third World Congress on Nature and Biologically Inspired Computing. doi:10.1109/nabic.2011.6089647

Silva, C., & Ribeiro, B. (2003).*The importance of stop word removal on recall values in text categorization.* Proceedings of the International Joint Conference on Neural Networks, 2003. doi:10.1109/ijcnn.2003.1223656

Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern Anal Applic Pattern Analysis and Applications, 8*(1-2), 199-209. doi:10.1007/s10044-005-0256-3

Salton, G., Wong, A., & Yang, C. S. (1975).A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620.

Ramos, J. (2003*). Using TF.IDF to determine word relevance in document queries.* Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. doi:10.18653/v1/w16-2365

Mesleh, A. M., & Kanaan, G. (2008).*Support vector machine text classification system: Using Ant Colony Optimization based feature subset selection.*2008 International Conference on Computer Engineering & Systems. doi:10.1109/icces.2008.4772984

Torkkola, K. (2004). Discriminative features for text document classification. *Formal Pattern Analysis & Applications, 6*(4).doi:10.1007/s10044-003-0196-8

Martinez, A., & Kak, A. (2001).PCA versus LDA.IEEE Transactions on Pattern Analysis and Machine Intelligence IEEE Trans. Pattern Anal. *Machine Intell., 23*(2), 228-233. doi:10.1109/34.908974

Liu, T., Chen, Z., Zhang, B., Ma, W., & Wu, G. (2004).*Improving Text Classification using Local Latent Semantic Indexing.* Fourth IEEE International Conference on Data Mining (ICDM'04). doi:10.1109/icdm.2004.10096

Chen, L., Tokuda, N., & Nagai, A. (2003).A new differential LSI space-based probabilistic document classifier. *Information Processing Letters, 88*(5), 203-212. doi:10.1016/j.ipl.2003.09.002

Liu, H., & Yu, L. (2005).Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491-502.

Doan, S., & Horiguchi, S. (2004).*An Efficient Feature Selection Using Multi-Criteria in Text Categorization.* Fourth International Conference on Hybrid Intelligent Systems (HIS'04), 86-91. doi:10.1109/ichis.2004.20

Yang, Y., & Pedersen, J. O. (1997).*A comparative study on feature selection in text categorization.* Machine Learning International Workshop Then Conference, 412-420. Morgan Kaufmann Publishers, INC

Ng, H. T., Goh, W. B., & Low, K. L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. *ACM SIGIR Forum, 31*, 67-73.

Kennedy, J., & Eberhart, R. (1995).*Particle swarm optimization*. Proceedings of ICNN'95 - International Conference on Neural Networks, 4, 1942-1948. doi:10.1109/icnn.1995.488968

Kennedy, J., & Eberhart, R. (1997).*A discrete binary version of the Particle swarm algorithm*. Proceedings of The 1998World Congress on Computational Cybernetics and Simulation, 4, 4104-4108.

Shi, Y., & Eberhart, R. (1995).*A modified particle swarm optimization*. Proceedings of The 1998World Congress on Computational Intelligence, 6, 69-73. doi:10.1109/icnn.1995.488968

Yang, B., Chen, Y., & Zhao, Z. (2007).*Survey on Applications of Particle Swarm Optimization in Electric Power Systems.* IEEE International Conference on Control and Automation, 481-486. doi:10.1109/icca.2007.4376403

Mitchell, T. M. (1997). *Machine learning.*McGraw Hill. Retrieved May 3, 2016, from http://personal.disco.unimib.it/Vanneschi/McGrawHill_-_Machine_Learning_-Tom_Mitchell.pdf

Burges, C. J. (1988). A tutorial on Support Vector Machine for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121-167.

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine learning. 20*(3), 273-297.

Naji, H., Ashour, W. (2016). Text Classification for Arabic Words Using Rep-Tree. *International Journal of Computer Science and Information Technology IJCSIT, 8*(2), 101-108. doi:10.5121/ijcsit.2016.8208

Weka 3: *Data Mining Software in Java*. (2011). Retrieved September 07, 2016, from http://www.cs.waikato.ac.nz/ml/weka

JMLR: Machine Learning Open Source Software. (2010). Retrieved September 07, 2016, from http://www.jmlr.org

Arabic Corpora - Mourad Abbas. (2004). Retrieved September 07, 2016, from https://sites.google.com/site/mouradabbas9/corpora

Abbas, M., Smaili, K., (2005).*Comparison of topic identification methods for Arabic language*. International Conference on Recent Advances in Natural Language Processing, 14.

Arabic Corpora - Alj-News.(2004). Retrieved September 07, 2016, from https://filebox.vt.edu/users/dsaid/Alj-News.tar.gz.

Mendes, R., Kennedy, J., Neves, J. (2004). The fully informed particle swarm: simpler, maybe better. *IEEE Transactions on Evolutionary Computation*, *8*(3), 204-210.

Tsai, M., Su, C., Chen, K., & Lin, H. (2012).An Application of PSO Algorithm and Decision Tree for Medical Problem. *Neural Comput & Applic Neural Computing and Applications, 21*(8), 124-126. Retrieved September 07, 2016, from http://psrcentre.org/images/extraimages/31012565.pdf