



Sudan University of Science and Technology
College of Graduate Studies



**Partial Research for Master Degree in Information
Technology Entitled:**

**A Comparative Study for Two Stemming
Algorithms for Arabic Wikipedia Documents
Classification Based on Similarity Measures**

دراسة مقارنة لخوارزميتي تحليل الجذور لتصنيف ملفات الويكيبيديا العربية بناءً على

مقاييس التشابه

Submitted by:

Mohamed Idris Ali khamis

Supervised by:

Dr. Ali Ahmed Alfaki Abdalla

November 2018

الآيه

قال تعالى :

(وَوَصَّيْنَا الْإِنْسَانَ بِوَالِدَيْهِ إِحْسَانًا حَمَلَتْهُ أُمُّهُ كُرْهًا وَوَضَعَتْهُ كُرْهًا وَحَمَلُهُ
وَفَصَالُهُ ثَلَاثُونَ شَهْرًا حَتَّىٰ إِذَا بَلَغَ أَشُدَّهُ وَبَلَغَ أَرْبَعِينَ سَنَةً قَالَ رَبِّ أَوْزِعْنِي
أَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي أَنْعَمْتَ عَلَيَّ وَعَلَىٰ وَالِدَيَّ وَأَنْ أَعْمَلَ صَالِحًا تَرْضَاهُ
وَأَصْلِحْ لِي فِي ذُرِّيَّتِي إِنِّي تُبْتُ إِلَيْكَ وَإِنِّي مِنَ الْمُسْلِمِينَ)

صدق الله العظيم

الآيه (15) سورة الأحقاف

Dedication

This dissertation is dedicated to:

My Sweetheart my, dear father,

My Soul, My lovely mother,

The source of my happiness, my brother and

sisters,

My friends and colleagues,

Ask God to bless them and prolong their life

Acknowledgments

*In the name of Allah, the Most Gracious and the Most Merciful Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. And I Would like to express my gratitude and thankfulness to my Supervisor **Dr. Ali Ahmed Alfaki** for his continued efforts during all phases of the research Ask God to make this work in the balance of his good deeds,*

I am very appreciative to thank all who provide a helping hand to achieve this study specifically my teachers at the college of computer sciences and information technology Sudan University of sciences and technology

A lot of thanks to the University of Kordofan for offering me the scholarship to study this program

Abstract

Text mining is an important field in information retrieval; it organizes a large number of text documents that are available on the internet to facilitate the retrieval process and increase efficiency. Text classification is automatically determining the category of new or unseen documents that depends on the content of the document itself. In text classification, text preprocessing is a fundamental step to obtain a better result. Arabic text processing depends on stemming algorithms to achieve high accuracy. This research aims to compare between two stemming algorithms: the snowball approach (snowball light) and the root approach (Shereen Khoja) using three similarity measures: Euclidean distance, cosine similarity, and Pearson correlation distance. This research uses the Arabic Wikipedia dataset and TF-IDF as a weight scheme to construct the vector space model to represent the weight of selected features of text. For evaluation measures, the research applies overall accuracy, average recall, average precision, and average F1 measure to assess the results of the classified text documents. The collection of documents is divided into training and test documents according to three experimental splits: (85% – 15%), (80% – 20%), and (90% – 10%) for training and test documents respectively. The results showed that the overall accuracy of the Shereen Khoja stemmer is better than the Snowball stemmer in all experiments, except for cosine similarity in the first experiment and Euclidean distance in the third experiment, which has a better accuracy when using the Snowball stemmer.

المستخلص

تنقيب النص حقل مهم جدا في مجال استرجاع المعلومات من خلاله يمكن تنظيم العدد الكبير من الملفات النصية المتاحة عبر الانترنت لتسهيل عملية استرجاعها و زيادة الفعالية عن طريق تصنيف الملفات النصية. تصنيف الملفات النصية هي عملية آلية يتم عبرها تحديد الفئة التي ينتمي اليها ملف غير معروف فئته إعتقاداً على محتوى الملف نفسه. في مجال تصنيف الملفات النصية، عملية معالجة النصوص تعتبر خطوة اساسية للحصول على دقة عالية . هذا البحث يهدف لمقارنة بين طريقتين من طرق معالجة النص العربي هما محلل الجذور Snowball و محلل Shereen Khoja بإستخدام ثلاث من مقاييس التشابه (Euclidean Distance, cosine similarity, and Pearson Correlation distance). في هذا البحث استخدمت مجموعة بيانات الويكيبيديا ، وايضا تم استخدام IDF-TF كطريقة أو دالة لحساب الأوزان للخصائص التي تم إختبارها من النص لإنشاء نموذج فضاء المتجه لتمثيل خصائص النص المختارة. لقياس أداء نموذج تصنيف النص تم تطبيق المقاييس التالية : الدقة ، و الاستدعاء precision ، و fl measure وذلك لتقييم نتائج مقاييس التشابه. تم تقسيم مجموعة البيانات إلى ملفات نصية للتدريب وخرى للإختبار بناءً على ذلك تم تصميم ثلاثة تجارب الاولى اعتمدت التقسيم (85% - 15%) و الثانية (80% - 20%) والثالثة (90% - 10%) للتدريب والإختبار على التوالي. أظهرت النتائج أن طريقة محلل الجذور Shereen Khoja أكثر دقة من طريقة محلل الجذور Snowball ماعدا عند إستخدام مقياس التشابه cosine similarity في التجربة الاولى و مقياس Euclidean Distance في التجربة الثالثة فإنهما أكثر دقة عند إستخدام طريقة محلل الجذور Snowball .

Table of Contents

الآيه	I
Dedication	II
Acknowledgments	III
Abstract(English)	IV
Abstract(Arabic)	V

CHAPTER I: INTRODUCTION

1.1 Background	1
1.2 Problem statement.....	2
1.3 Research objectives	2
1.4 The scope of the study	3
1.5 Significant of the study	3
1.6 Thesis organization	3

CHAPTER II: LITERATURE REVIEW

2.1 Introduction	4
2.2 Arabic language structure	4
2.3 TEXT MINING AND TEXT MINING TECHNIQUES	5
2.3.1 Information Retrieval.....	5
2.3.2 Information Extraction.....	6
2.3.3 Summarization	6
2.3.4 Clustering.....	6
2.3.5 Categorization.....	7

2.4 APPLICATIONS OF TEXT MINING.....	7
2.4.1 Academic and Research Field.....	7
2.4.2 Business	8
2.4.3 Anti-Spam Filtering of Emails.....	8
2.4.4 Biomedical applications.....	8
2.4.5 Sentiment analysis	9
2.5 Related works	9
2.6 Summary	16

CHAPTER III: RESEARCH METHODOLOGY

3.1 Introduction	17
3.2 RESEARCH PHASES	17
3.2.1 phase 1 text preprocessing	17
3.2.1.1 Arabic stop words	18
3.2.1.2 snowball stemmer	20
3.2.1.3 Shereen Khoja stemmer	22
3.2.2 Phase 2 Vector Space Model Construction(VSM)	24
3.2.3 PHASE 3 TEXT CLASSIFICATION BASE ON SIMILARITY MEASURES ..	26
3.2.3.1 Euclidean Distance	26
3.2.3.2 Cosine Similarity	27
3.2.3.3 Pearson Correlation distance.....	28
3.2.4 Phase 4 Evaluation Measures	28
3.3 Dataset	30
3.4 TOOLS AND TECHNOLOGIES	32
3.4.1 Python programming language	32

3.4.2 RapidMiner	32
3.4.3 MATLAB	32
3.5 Summary	33

CHAPTER IV: IMPLEMENTATION AND RESULTS

4.1 Introduction	34
4.2 Pre-processing	34
4.3 Implement of VSM	39
4.4 Experimental Design	39
4.5 Results of the first experimental	41
4.6 Results of the second experimental	47
4.7 Results of the third experimental	48
4.8 Discussion	49

CHAPTER V: RECOMMENDATIONS AND FUTURE WORKS

5.1 Recommendation and Future Works	51
5.2 Conclusion	51
5.3 References	52

List of Abbreviations

AraNLP	Arabic Natural Language Processing
BOW	Bag-Of-Word”
CCA	Corpus of Contemporary Arabic
EFCM	Enhanced Fuzzy c – Means
FAC	Facility
FCM	Fuzzy C – Means
FN	False Negative
FP	False Positive
FRAM	Frequency Ratio Accumulation Method
FS	Feature Selection
GPE	Geo-Political
IDF	Inverse Data Frequency
IE	Information Extraction
IR	Information Retrieval
ISRI	Institute of Scrap Recycling Industries
kNN	k-Nearest-Neighbor
LOC	Location
LR	Logistic Regression
LSI	Latent Semantic Indexing
ML	Machine learning
MSA	Modern Standard Arabic
NB	Naïve Bayesian
NLP	Natural Language Processing
NLTK	Natural Language Tool Kit
NLTK	Natural Language Tool Kit
NOT	Not-Named-Entity
ORG	Organization
PER	Person
PNN	Polynomial Neural Network

PRO	Product
RBF	Radial Basis Function networks
SGML	Standard Generalized Markup Language
SPA	Saudi Press Agency
SVD	singular value decomposition
SVM	Support Vector Machine
TC	Text Classification
TF	Term Frequency
TP	True Positive
VEH	Vehicle
VSM	Vector Space Model
WEA	Weapon

List of table

Table 2.1 Summarize the literature review	15
Table 3.1 kinds of affixes attached to the word	20
Table 3.2 suffixes and prefixes for light stemmer	21
Table 3.3 procedure for root (Khoja) stemming	23
Table 3.4 light and root stemmer's weakness	23
Table 3.5 the number of the document in the data set.....	31
Table 4.1 implement the text preprocessing steps on the Arabic dataset	35
Table 4.2 number of documents after removing	36
Table 4.3 the number of words included in each stemmer	36
Table 4.4 the number of words after calculating TFIDF	37
Table 4.5 the number of documents after calculating TFIDF	37
Table 4.6 Number of training and testing documents using Snowball	38
Table 3.7 Number of training and testing documents using Snowball	38
Table 4.8 the Confusion matrix for Euclidean Distance with Shereen Khoja stemmer ..	41
Table 4.9 Recall for Snowball and Shereen Khoja stemmers	42
Table 4.10 precision for both stemmers using three similarity measures.....	43
Table 4.11 F1 measure for both stemmers with three similarity measures	45
Table 4.12 the results of the first experimental	46
Table 4.13 the results of the second experimental.....	47
Figure 4.6 the result of the second experimental using Shereen Khoja and Snowball stemmers.....	47
Table 4.14 the results of the third experimental.....	48

List of figure

Figure 3.1 the list of NLTK Arabic stop words.....	19
Figure 3.2 the steps of Arabic text processing.....	24
Figure 3.3 snapshot of raw text.....	31
Figure 4.1 snapshot of the VSM – using root stemmer dataset.....	39
Figure 4.2 Recall for both stemmers with three similarity measures	42
Figure 4.3 Precision for both stemmers with three similarity measures	44
Figure 4.4 F1 measure for both stemmers with the three similarity measures.....	45
Figure 4.5 the result of the first experimental using Shereen Khoja and Snowball stemmers	46
Figure 4.6 the result of the second experimental using Shereen Khoja and Snowball stemmers	47
Figure 4.7 the result of the third experimental with Shereen Khoja and Snowball stemmers	48

CHAPTER: I

Introduction

1.1 Background

The advanced technology in the domain of the communications and information technology has led to daily increase in the volume of the documents, and doubling the use of searching engines on the internet that make the retrieve of documents the very difficult; therefore, the text of documents is needed to be classified. The text classification is the technique of organizing texts into classes to make the retrieving process most effective and efficient. The goal of data (text) classification system makes the data of the text easy to be found and retrieved. The documentation of roles and guidelines for data (text) classification must define categories and criteria of classification. The organization will classify data and assign the roles and tasks for the staffs inside the organization, concerning with data stewardship. when a data classification structure has been done, the security ethics which suit with the handling practices that describing data's lifecycle requirements must be addressed (Otair, 2017). Text Classification (TC) or text categorization automatically determines the class or category of the new text which actually referred to as belonging to one class such as Economic, technological or political class. The large amount of text documents that existing are available on wide world web and these documents have different human languages from all over the world; need fast, perfect, powerful, and capable tool that is automatically classifying documents. TC has many applications including spam filtering, information retrieval, topical crawling, online news, digital libraries, and this not all of its' applications . (M. Al-Tahraw, 2016)

TC depends on natural language processing. This research will be using Arabic language which is the formal language of the Arab world countries, the language of the Holy Quran, and has fundamental importance in the life of every Muslims around the world. It carries many ideas and values it is basis of the Arab nation and keeps it from loss. More over Arabic is a rich language of extensive values, terminologies, vocabulary, and sentences structures. It is the first language accordingly to the number

of terms, meanings and sentence structure. The Arabic language is characterized as one of the Semitic languages; the Semitic languages consist of Asiatic and African languages. It's one of the six formal languages of the United Nations. Arabic contains 28 letters. There are many porcessing perform on the words called Natural Language Processing (NLP), it's an Artificial Intelligence (or Machine Learning) field that is capable of recognizing and understanding human speeach. It is a track in computer science witch related to computational linguistics process, and It has interactions between computers and human natural languages. The challenges in this track is focusing on how the computers understanding and analyzing the human natural languages. (Otair, 2017)

1.1 Statement of the problem

Lately, there is an exponential growth of available documents in several fields on the internet and these documents have different categorizations, therefore; classifying these documents manually is very difficult and time-consuming. (Elhassan & Ahmed, 2015b)

Recently, there are many stemmers used in pre-processing for Arabic text based on stem, root and statistical approaches stemmer (Madani & Kissi, 2017a). The overall result of using any of the above mentioned stemmers differs in term of classification, accuracy and other metrics depending on the nature of stemmers. On another hand, most of the latest comparative studies using machine learning algorithms to compare between different stemmers- [light and root stemmers].

This study proposes text classification model base on Euclidean Distance, cosine similarity, and Pearson Correlation distance similarity measures to compare between Snowball and Shereen khoja.

1.3 The objectives of the Study

The main objective of this study is to develop an automated document categorization method that is capable of automatically organizing and classifying documents through the following sub objectives:

- Construct vector space model based on TF-IDF weighting scheme.

- Apply (Euclidean distance, cosine similarity, and correlation distance) similarity measures.
- Performance evaluation based on Accuracy Recall, Precision, F-Measure.

1.4 The scope of the Study

This research focus on the study that compare between two popular Arabic stemmers Snowball, and Shereen Khoja stemmers algorithms base on the offline Arabic text classification using Arabic Wikipedia dataset.

1.5 Significant of the Study

The significance of this study is to provide method based on the computer systems for classification text that depends on bag of words vector. Toward automatically organize the documents into their corresponding classes or categories. In order to be a process of retrieve related documents more efficiently.

As well the significance of this study conduct to use two above stemmers to determine which one is better than other using Similarity Measures

1.6 Thesis organization

The thesis is organized as follows in Chapter II that covers Theoretical framework about the domain of the research and also presents the related works. Chapter III that describes the proposed method for the Arabic text classification and cover the text pre-processing and describe the overview about the technologies that used in this study. Chapter IV discusses the design experimental and results and discussion. In last Chapter IIV include the conclusion and recommendations.

CHAPTER: II

Literature review

2.1 Introduction

The Arabic language is a common language around the world and it considers the Arab mother tongue language; therefore the content of the Arabic language on the web is growing, and that increase the numbers of online Arabic documents. Accordingly to that, we need different classification algorithms to do texts classification for various purposes. Through this chapter, present the theoretical framework starting with the Arabic language structure in section (2.2) and concept of text mining and text mining techniques in section (2.3) followed by description of the applications of text mining in section (2.4) , describing related works in section(2.5).

2.2 Arabic language structure

Arabic language is one of the commonly spoken languages in the world, as it is the sixth most spoken language about 320 million speakers around the world.(Ababneh, and etal, 2014). The Arabic language belongs to the Semitic family of languages and it is the formal language of the Arab countries. It is the language of the Holy Quran. There are two types of spoken Arabic language: Classical Arabic alfushaa (الفصحى), and Modern Standard Arabic (MSA) which is based on the classical language. Unlike, Latin-based alphabets, Arabic alphabet are written from the right to the left. It has 28 characters, and there is no capitalization in Arabic text. The Arabic alphabets are linked by preceding or following letter except for the six non-connectors letters [ا, و, د, ر, ذ, ز] as they are linked only when in the medial of the word or the end on its right side. (Elhassan & Ahmed, 2015a)

The semivowel letters و, ا, and ي, sometimes act as consonants, and as vowels in others depending on the context, while the rest of the letters are constants. There are two genders: masculine (muzkir مذكر) and feminine (muanith مؤنث) which is presented by adding the suffix (ة) at the end of the word, and on the numerical context there are singular (mufrad مفرد), dual (mathnaa مثني) and plural (jame جمع) numbers. Plural numbers are divided to

regular (جمع الصحيح) or broken (جمع تكسير). Arabic grammarians divided the word into three types: noun, verb, particles. Grammatically the verb has three states: nominative (الرفع), accusative (النصب), and genitive (الجر) (Elhassan & Ahmed, 2015a).

2.3 TEXT MINING AND TEXT MINING TECHNIQUES

The text mining is emerged as new technique that attempts to gather the novel concept of the information from natural languages. In many cases the process that performed on the text led to eliminate some text features when the process of analyzing text to extract information that is valuable for certain purposes. The text is unstructured, formless, and there are challenging to process the text algorithmically. But, this time, the text is most public formal exchange of information in all societies around the world. The field of text mining frequently deals with texts that have communication of realistic information or sentiments. The term of “text mining” is mostly used to denote any system that analyzes large amount of natural languages texts or documents and set lexical , linguistic that usage patterns in an attempt to extract probably useful (Tsai, 2011). Text mining is a multi-disciplinary field that depending on information retrieval, data mining, machine learning, statistics and computational linguistics such as NLP (Talib, Hanif, Ayesha, & Fatima, 2016).

2.3.1 Information Retrieval

The most well-known information retrieval (IR) systems; for example Google search engines considers as one of the most popular IR application which recognize the documents on the internet that are clustered with a set of certain words. Document retrieval is, processing to extract the valuable information for the user. Consequently document retrieval is tracked by the text summarization step that emphasizes on the query posed by the user or an information extraction stage. IR in the broader sense include the wide range of information processing, from information retrieval to knowledge retrieval. The IR is related to an old research area. The first try for building automatic indexing system was in 1975. After that it’s extended , increased and growing on the internet and emerge to the classic search engines. (Dang & Ahmad, 2015)

2.3.2 Information Extraction

The information extraction is a method to convert semi-structured and unstructured text to a structured format. That's mean extraction is meaningful for a large volume of text corpus. Information extraction software detects key phrases and relationships through the text. It performs by matching predefined sequences in the text, this process named pattern matching. (Gupta & Lehal, 2009) it has much valuable information like name of somebody, place and organization or company etc... are extracted without understanding the natural of the text. IE is focusing on the extraction of semantic information from the text (Dang & Ahmad, 2015).

2.3.3 Summarization

Text summarization is a process of collecting and creating a short representation from raw text documents. Pre-processing and processing operations that performed on the raw text for summarization is fundamental step to get more accurate results (Talib et al., 2016). With enormous amount of texts, text summarization software processes and summarizes the document in the period that would take the user to read the first paragraph. The key to summarization is to decrease the length and redundant detail of a document while just keep its main points and the complete meaning(Gupta & Lehal, 2009).

2.3.4 Clustering

Clustering is an unsupervised technique unlike classification clustering is classifying the text documents or data in clusters by applying different clustering algorithms depend on similarity measures. In cluster, similar terms and patterns are grouped to take out from many documents. Clustering is achieved on top-down and bottom-up manner. From NLP point of view, several types of mining techniques and tools are applied to normalize and analysis of unstructured text before starting to clustering step. The power of clustering comes from the power of similarity between objects in the same cluster and dissimilarity between other objects in other clusters. There is various techniques of

clustering such as hierarchical, centroid, density, distribution, and k-mean(Talib et al., 2016).

2.3.5 Categorization

Text categorization or text classification is referred to the family of supervised learning techniques. where the classes or categories are predefine and then start the stage of training document. Then, not all text documents selected. Instead of that we use some key words based on the weight of words schema form the selected document. In the Nineties the field of classification is fully developed with the availability of continuous increasing numbers of text documents on the internet and many digital libraries. This led to organize these documents in order to facilitate their use. Categorization is the determining the label or category of documents to predefined categories according to document's content. And also it is a collection of text documents, the process of discovering the accurate category or class for each document. Nowadays automatic text classification is applied on different contexts from classical automatic or semiautomatic indexing of texts it's include the following applications: spam filtering , categorization of Web page under hierarchical catalogs, topic tracking, automatic metadata generation, detection of text genre and many others. The learning of automated text categorization starts in the early 1960s. It is hot topic in machine learning nowadays (Dang & Ahmad, 2015).

2.4 APPLICATIONS OF TEXT MINING

2.4.1 Academic and Research Field

In the education field, different text mining tools and techniques are used to analyze the educational trends in the specific province. Student's interest in the particular field and employment percentage (Al-Hashemi, 2010). Text mining also has been used in scientific research field help to find and classify research papers, thesis, dissertation and related material or works of different fields at one place. The use of k-means clustering and other techniques aid to recognize the attributes of relevant information. Student's

performance in different courses can be accessed and how many attributes affects on the selection of subjects(Talib et al., 2016).

2.4.2 Business

The companies have the biggest consumers and producers of information. the large amount of stored information lies unexploited. Most companies are using this information continually to develop their business into perfection, and customer's satisfaction. Before making changes, it would be useful to judge Staff effort or Customer feedback (opinion). This is done by different opinion poll like quantitative or descriptive. In the case of descriptive, questions take natural language answers, which are very difficult to process. For this Summarization can help. In this sentiment analysis (Cohen & Hersh, 2005)

2.4.3 Anti-Spam Filtering of Emails

The growing of unwanted e-mails, more frequently recognized as spam, over the last years has been discouragement continually the usability of e-mails. One solution is offered by anti-spam filters. The most commercially available filters use black-lists and hand-crafted rules. On the other hand, the success of machine learning methods in text categorization offers the opportunity to achieve anti-spam filters that is quickly, efficiently and may be adapted to new types of spam(Gupta & Lehal, 2009)

2.4.4 Biomedical applications

The medical field has generated a large volumes of information. For example hospital records, clinical trials, studies, research reports and doctors's notes. Most of this information is in a text form. Furthermore there is a lot of focusing in the research of genes and proteins like other medical fields. To read and analyse all the information manually is very diffecult. That's require tools to fetch all information from a large database. These tools are built completely to mine medical or scientific literature or information. Some tools capture the communication between cells, molecules, and proteins, and others extract biological facts from scientific articles. Thousands of these

facts can be automatically analysed for similarities or relationships(Cohen & Hersh, 2005).

2.4.5 Sentiment analysis

The sentiment analysis is considered as a computing effective area uses in text mining. Student assessments, movie evaluations, opinion polls, kid's stories, and science fiction stories are the applications of sentimental computing. WordNet and ConcepNet are the popular applications of sentiment analysis. (Dutta, 2017)

2.5 Related works

In the last decade Arabic TC article has grown rapidly, researchers addressed the issue of Arabic TC using different classification techniques, datasets, and pre-processing operations, as no standards exist regarding free benchmarking Arabic corpora or pre-processing tools to processing the Arabic text.

This section discusses a number of studies in the field of Arabic TC. M. Al-Tahraw, (2016) many works that related with Arabic text classification are based on keywords while other works used semantic web ontology such as (M., M., & Hussein, (2016). The Similarity Measures has been wildly used in pattern recognition, for example, Nguyen & Bai, (2011) and (Kaur & Aggarwal, 2013) the first Similarity Measures have been used (cosine similarity) for face Verification and then for image content based retrieval system.

This research, the focus will be on the linguistic systems such as text classification (Li & Han, 2013) and (Wartena & Brussee, 2008). Some of the researchers use Similarity Measures for document clustering such as (Wartena & Brussee, 2008) all of them used Similarity Measures on English documents, Few researchers used it for Arabic documents such as (Al-Anzi & AbuZeina, 2017). Many researches focused on comparative study between Arabic stemmer's algorithms by using text classification or text clustering.

(Previtali, Arrieta, & Ermanni, 2015) have presented a comparison of the five machine learning algorithms, studied the effects of different Arabic stemmers (light and root-based stemmers), and made comparison between different data mining software (Weka and RapidMiner).The results explain that the best accuracy provided by the Support Vector

Machine (SVM) classifier, particularly when used with the light10 stemmer algorithm, the WEKA tool showed better result than RapidMiner software, RapidMiner have memory limitation problems specifically when the size of data set is huge. Therefore the data set in this work must be reduced to get the best result.

(Madani & Kissi, 2017a) have developed a new stemming algorithm (origin-stemmer) in which each term of a given document is represented by its root then they compared their stemming algorithm (origin-stemmer) and other stemmer called Khoja's stemmer on Arabic text classification. This investigation was performed using chi-square as a feature selection to reduce the size of the feature space and decision tree as a classifier. This study used a corpus that consists of 5070 documents classified into six categories: sport, entertainment, business, Middle East, switch, and world using WEKA toolkit. The recall, f-measure and precision measures are used to compare the performance of the obtained models. The experimental results showed that text classification using origin-stemmer, outperforms the classification using Khoja's stemmer. The f-measure was 92.9% in the sports category and 89.1% in the business category. This work applied one classification algorithm is not enough to provide good result, and so more than one classification algorithm should be used to enhance the result of this study.

(Hadni, and etal, 2013) proposed TC system-an effective hybrid method- for Arabic text stemming. Naïve Bayesian (NB) classifier and the SVM classifier were used to build TC system. The dataset is extracted from a large Arabic corpus (Corpus of Contemporary Arabic (CCA)). This corpus is classified into 12 categories: economics, politics, education, science, health and medicine, interview, recipes, religion, sociology, spoken, sports, and tourist and travel. The result found that hybrid Stemmer is showing good results in most class.

(Al-Kabi, and etal, 2015) introduced a new light and heavy Arabic stemmer, compared between two popular Arabic stemmers (Ghwanmeh and Khoja stemmers), and used benchmark tests of different Arabic stemmer's data set. A dataset consisting of 6081 Arabic words was driven from native Arabic three letter verbs is used to evaluate their proposed Arabic stemmer algorithm in relevance to the other two stemmers. The Results showed that the accuracy of their stemmer is slightly better than the accuracy yielded by the others. This

study compare just Arabic root base stemmer and does not cover the Arabic light base stemmer.

(A. Otair, 2013) compared and analyzed many of Arabic stemmer's algorithms specifically light stemmers in terms of affix lists, and discussed the main Arabic language characteristics. The evaluation of the algorithms shows that Arabic stemming algorithm is one of the biggest information retrieval and text classification challenges. The results showed that the light10 stemmer outperformed the other stemmers. The researcher did not mention other stemmers which he used to compare with light10 stemmer.

All the researchers are using machine learning algorithm to compare the performance of Arabic stemmer algorithms, while there is no comparative study between Arabic stemmers algorithm with a similarity measure. This research aims to compare between Arabic light (Snow ball) and root (Khoja) stemmer's algorithms.

M. Al-Tahraw,(2016) compares Polynomial Neural Networks (PNNs) with five famous classification algorithms in TC used Aljazeera-News Arabic dataset collected from the website of Al-Jazeera Arabic News channel. All experiments used the similar text settings, like preprocessing, Feature Selection (FS), Chi-Square and reduction criteria, feature weighting, and classifier performance evaluation measures. These algorithms are SVM, NB, k-Nearest-Neighbor (kNN), Logistic Regression (LR) and Radial Basis Function (RBF). The five classifiers tested in their experiments are evaluated using recall, precision, and F1- measures. The study found that the PNN is a competitive classifier in the field of Arabic TC. Use one stemmer (root base stemmer) and was not applied on the other stemmer.

(Dutta, 2017) proposed method to Enhance Fuzzy C – Means (EFCM) which is unsupervised learning base on clustering technique. The text corpus is classified into categories: Education, Agriculture, Politics, Entertainment, Geography, and others were used the J48, Fuzzy C – Means (FCM), and k-means were compared with the proposed method. Dataset collected from Aracorpora website (www.aracorpora.e3rab.com) having 1.3 million words with a number of occurrences. The text pre-processed was given an input to the system and the output is classifying the Arabic text into particular cluster according to

(1 to 6). The result shows the accuracy for clusters having metrics Cosine similarity, Euclidean distance and Dice coefficient for each experimented method with the dataset. The accuracy of EFCM is better than the other methods. The researcher doesn't expose what Arabic stemming method is used.

(Mohammad, and etal,2016) compare between three classification algorithms. (K-NN,C4.5) and Rocchio algorithm. Are Using root stemming for Arabic text classification. Root in the Arabic language available in three, four, five and six letters. Also over 80% of Arabic words can be mapped into a three-letter root. Dataset that collected from Aljazeera news website, Saudi Press Agency (SPA) and Al-Hayat Dataset consists of 1400 Arabic documents belongs to eight categories and divided into two parts: First part is used for training and it consist 920 documents (66%) and the second part consist of 480 documents are used for testing (34%). recall, precision, and F1 have been used as evaluation measures performance. The results of K-NN and Rocchio can work well on Arabic data set C4.5.

(Al-Tahrawi , 2013) improves the accuracy of Polynomial Networks algorithm in English Text Categorization by using the role of rare or infrequent terms. Use the Reuters dataset Corpus. The Porter Stemmer was used for text processing steps that performed on the datasets, Only letters, hyphens “-” and underscores “_” are kept; any other character is eliminated, and converts all letters to low case (capitalization). And ignoring list of more than 1000 stop words that to reduce the number of terms. Chi-Square (χ^2) was used to compute the strength of each term in the corpus. Chi-square was yield good results in classification, compared to others FS and use three different methods to reduce the result of the terms. The PNN classifier performance was evaluated by computing its accuracy. The result shows that the best result comes with keeping rare terms rather than removing it in all experiments.

(Ghwanmeh, and etal,2007), propose a framework for ontology-based information retrieval system for the Arabic language. The system consists of four main modules: namely the query parser, indexer, search and a ranking module. Their approach includes building a semantic index by linking ontology concepts to documents, including an annotation weight for each link, to be used in ranking the results. they also improved the framework with an automatic document categorizer, which enhances the overall documents ranking. they built

three Arabic domain ontologies: Sports, Economic, and Politics. As an example for the Arabic language, and building a knowledge base that's consists of 79 classes and more than 1456 instances. The system is evaluated by using the precision and recall metrics. In addition to they have done many retrieval operations on a sample of 40,316 documents with a size of (320 MB) of pure text. The results show that the semantic search enhanced text classification gives a better performance results. The processes of text processing not involve in the stemming process, and stemming processing is fundamental process in Arabic text to get better results.

(Bijalwan, and etal, 2014), developed an information retrieval system based on text classification The data set in the SGML files form dataset called Reuters-21578 dataset. Have 21578 documents split: 9603 documents for training, 3299 for test, and 8676 unused documents. Performing various techniques in text pre-processed phases including: Bag of words, Stop word removal, TF-IDF, Case Folding, and Normalization. After that, they applied KNN, Term Graph algorithm, and Naïve Bayes algorithms to build classification model and classify each document into one of these classes (exchange, organization, people, places, and topics). And the Accuracy metric has been used to evaluate the accuracy of the classifier model and compares the results of the three algorithms above. The results show that KNN is maximum accuracy when comparing it with the NB, and Term-Graph. The drawback for KNN is that its time complexity is high but gives a better accuracy to the other algorithms.

(Elhassan & Ahmed,2015), The main purpose of this work is to describe and determine the effectiveness of the data preprocessing on a full word in the term of accuracy of both training model and classifier. The dataset include 750 documents from the local newspaper (Akhir Lahza and Alyoum Altali), and international newspaper (Al-Raya, Asharq Al-Awsat, and Al-Hayat) from the websites during the period from January 2001 to January 2015. These documents divided into five categories: economy, politics, religion, sport, and technology. Every category contains 150 documents; each document belongs to one category only. For every category there are 105 used for training and the rest used for testing. Use Precision, recall, and F1 to evaluate the classification model. In the text preprocessed phase using two approaches; first corpus and optimized corpus (with

elimination the stop words). The second approaches enhanced the accuracy of the training models. The result of experiments shows that the average of accuracy of SVM algorithm better than other algorithms in the training stage and also in the tested stage. The text processing in this work is not including stemming process, and researcher conducts that.

(Al-Anzi & AbuZeina, 2017), try to enhance Arabic text classification using Cosine similarity measure and (VSM). They commonly used as a model to represent textual information as numerical vectors However, Latent Semantic Indexing (LSI) is a better textual representation method as it maintains semantic information between the words and the singular value decomposition (SVD) method to extract textual features based on LSI. This study conducts a comparison between ML algorithms: Naïve Bayes, k Nearest Neighbors, Neural Network, Random Forest, Support Vector Machine, and decision tree. They Used a corpus contains 4,000 documents within ten categories (400 documents for each topic). The corpus contains 2,127,197 words with about 139,168 unique words. The testing set contains 400 documents, 40 documents for each topic. As a weighing scheme, they used Term Frequency Inverse Document Frequency (TF-IDF). The results show the classification methods that use LSI features significantly outperform the TF-IDF based methods. It also reveals that KNN (based on cosine measure) and support vector machine are the best performing classifiers.

(M. et al., 2016), proposed approach to enhancement the accuracy overall in Arabic text classification by using the Frequency Ratio Accumulation Method (FRAM) as a classifier. They use three different datasets: dataset without stemming, a dataset with Tashaphyne Light Arabic Stemmer and dataset with ISRI Stemmer. And text representation use Bag-Of-Word (BOW). It is the most popular document representation scheme in text categorization. It Uses selection feature to reduce the size of the training file, and split the three data sets into training data and test data. And calculate the accuracy of the classifier by using classification measures such as accuracy, precision, recall, and f-measure. The result shows that the text classification with normalizes achieved highest classification accuracy than Tashaphyne Light Arabic Stemmer and ISRI Stemmer.

Table2.1 Summarize some of the literature review above

Table 2.1 Summarize the literature review

#	Paper	Dataset	Number of class	Method used	The best Results
1	Previtali et al., (2015)	sole Arabic dataset, 2700 documents	9	NB,VSM,KNN , Decision table ,Decision tree	VSM, light stemmer
2	Madani & Kissi, (2017b)	open source Arabic corpora (OSAC) 5070 Arabic documents	6	Khoja's stemmer, rout stemmer	rout stemmer
3	Hadni et al., (2013)	(Corpus of Contemporary Arabic (CCA)	12	Stemmers: Khoja, light, N-gram, hybrid Classifiers: NB, VSM	Hybrid stemmer And VSM classifier
4	Al-Kabi et al., (2015)	6081 Arabic words	-	New stemmer, Khoja stemmer, Ghwanmeh stemmer	New stemmer
5	(M. Al-Tahraw, 2016)	AljazeeraNews,1 50 documents	5	PN,SVM, NB, kNN, LR, RBF,	PN
6	(Dutta, 2017)	Aracorpora website	5	J48, FCM, K – means, EFCMC	EFCMC
7	(Mohammad et al., 2016)	Three different sources Dataset 1400 documents	8	K-NN, C4.5, Rocchio algorithm	K-NN
8	(Bijalwan et al., 2014)	Reuters-dataset. 21578 documents	5	KNN, Term Graph algorithm, and Naïve Bayes	KNN
9	(Elhassan & Ahmed, 2015a)	750 documents from five difference	5	SVM, NB, J48, and kNN	SVM

		resources			
10	(Al-Anzi & AbuZeina, 2017)	the corpus that contains 4,000 documents	10	SVM, Cosine, LR, k-NN, NN, RF, NB, CT, CN2	SVM
11	(Syiam, Fayed, & Habib, 2005)	BBC, CNN Arabic corpora	-	Naïve Bayesian, DMNBtext, and C4.5.	DMNBtext

2.6 Summary

This chapter provided an overview of the theoretical framework, that includes a brief description of text mining. Also, present the different text mining techniques that depend on a number of factors. Applications of text mining have been explained. Moreover, describing the Arabic language structure. After that present the related work that study compares the difference between Arabic stemmer's algorithms and also mentioned some works related by Arabic text classification which use different text processing, feature generation, and machine learning algorithms.

The next chapter provides an overview of the text processing steps and describes the methodology of research.

Remove the non-Arabic characters such as English words

Remove diacritic

- ◌ْ | # shadda
- ◌َ | # Fatha
- ◌ِ | # Tanwin Fath
- ◌ُ | # Damma
- ◌ُ | # Tanwin Damm
- ◌ِ | # Kasra
- ◌ِ | # Tanwin Kasr
- ◌◌ | # Sukun
- # Tatwil/Kashida

Text processing include (Tokenization, Normalization, Remove the punctuation mark, Remove the non-Arabic characters and Remove diacritic) have been implemented using pre-processing Arabic text (motazsaad python code).

3.2.1.1 Arabic Stop words

are tokens and repeated excitingly in the whole of documents collection using stop words list. The advantage of stop words removal is reducing the size of the feature selection from a text document. In this research, the python NLTK stops words list has been used to filter out stop words from Wikipedia dataset. The list of stop word shown in the figure below, Fugue 3.1 explains the Arabic stop words within NLTK python's library that used to normalization the Arabic text.

3.2.1.2 Snowball stemmer

Snowball is a small string processing language aimed for creating stemming algorithms for use in Information Retrieval. The Snowball compiler translates a Snowball script into another language - currently ISO C, C#, Go, Java, JavaScript, Object Pascal, Python and Rust are supported. The Arabic stemming written on Snowball framework language by Assem Chelli and Abdelkrim Aries. It proposals light stemming and text normalization. depend on stem base approach.

The stem base is not try to give the linguistic root pattern for the word, instead, its main effort to remove the most common suffixes and prefixes. There are different types of Light Stemming. Many studies have considered this approach(Kalita, 2015)(Alabbas, and etal, 2016). The literature, in general, gives argument that light stemming provide best result in information retrieval applications. (Alabbas etal., 2016)

There are four types of affixes: antefixes, prefixes, suffixes and postfixes that can be attached to words. An Arabic word can represent a phrase in English, for example, the word ليخبرونهم which mean “to speak with them” is decomposed as shown in Table(Almusaddar, 2014)

Table 3.1 kinds of affixes attached to the word

Antefix	Prefix	Root	Suffix	Postfix
ل	ي	خبر	ون	هم
Preposition meaning “to”	A letter meaning the lense and the person of conjugation	News	Temination of conjugation	A pronoun meaning “them”

There are several versions of light stemming, all of the following have the same steps:

1. Tokenization
2. Replace initial, ا, اِ, اُ with .ا (Normalization)
3. Stop-words removal.
4. Remove punctuation, non-letters, and diacritics
5. Eliminate و (“and”) for light2, light3, and light8 if the remains of the word is 3 or more characters long. it is important to eliminate و , and it is also problem, because many common Arabic words begin with this character, hence the stricter length criterion here than for the definite articles.
6. Go through the list of suffixes once in the (right to left)inorder to indicated in Table3.2 ,removing any of the ffixes that are found at the end of the word, if its leaves 2 or more characters.

In table below describe the prefixes and suffixes for each types of light stemmer (Larkey, and etal, 2007)

Table 3.2 suffixes and prefixes for light stemmer

Light stemmer	Remove prefixes	Remove Suffixes
Light stemmer1	ال ، وال، بال، قال	None
Light stemmer2	ال ، وال، بال، قال، و	None
Light stemmer3	ال ، وال، بال، قال، و	هـ، ة
Light stemmer8	ال ، وال، بال، قال، و	ها، ان ، ات، ون، ين، يه، ية، هـ، ة، ي
Light stemmer10	ال ، وال، بال، قال، و، لل	ها، ان، ات، ون، ين، يه، ية، هـ، ة، ي

The snowball arabic light stemmer	أ، فال، وال، ف، و، لل، ال، ف، و، لل، ال بال، كال، ب، ك، ل، ب، ك، ك، سي، ست، سن، شأ، يست، نست، تست،	ه، ك، ي، نا، كم، ها، هن، هم، كما، هما، ا، ي، ون، ات، ت، ة، ي ه، ك، ني، نا، ها، هم، هن، كم، كن، هما، كما، كمو ت، ا، ن، ي، نا، تا، تن، ان، هن، ين، تما، وا، تم، و، تموي
-----------------------------------	--	--

this research used python programming language NLTK library to implement snowball stemmer on arabic text. At the end of this section we introduce the overview about this programming language.

3.2.1.3 Shereen Khoja stemmer

The Khoja Arabic stemmer algorithm introduced by Khoja and Garside (1999). They relied on morphological analysis to improve their stemmer by eliminating layers of prefixes and suffixes in the first and then checking a set of roots and patterns to specify whether the remainder was a known root with a known pattern. (Althobaiti, and etal, 2014a) More than 80% of the Arabic words can be mapped into three-letter root pattern, reducing a word to its root pattern could decrease the number of words from hundreds of thousands to as little as (Alabbas et al., 2016). This research uses the Java AraNLP project (Althobaiti, and etal, 2014b) to extract root from Arabic text using root stemming. AraNLP applies different techniques for preprocessing tasks on Arabic text based on the work needs in a sequence form and then save the preprocessed text into text files of UTF-8 encoding to ensure that the process of writing data on the file has been done correctly. (Alqarout, n.d.) A summarization of the Khoja's stemming procedure is shown in the table below (Almusaddar, 2014). Table 3.3 describes the steps of text processing using Khoja.

Table 3.3 procedure for root (Khoja) stemming

<p>Algorithm 3.2: Khoja Root-Based Stemming Algorithm</p> <p>Purpose: Stemming Arabic Words</p> <p>Input:</p> <ul style="list-style-type: none"> • Dataset • Stop-word list • Assets and patterns files <p>Output: Stemmed Dataset</p>
<p>Procedure:</p> <ol style="list-style-type: none"> 1. Tokenization 2. Replace initial, ! , ! , ! with .! 3. Stop-words removal. 4. Remove punctuation, non-letters, and diacritics. 5. Remove definite articles from the beginning of the word. 6. Remove the letter (ﺝ) from the beginning of the word and (ﻩ) from the end of the word. 7. Remove prefixes and suffixes 8. Comparing the resulting word to patterns stored in the dictionary, if the resulting root is meaningless the original word is returned without changes.

The table 3.4 describes the weakness of light and root stemmer when used

Table 3.4 light and root stemmer's weakness

Stemmer	Weakness
Shereen Khoja	- need to the updated dictionary - some of words product incorrect root, For example, the word (منظمات) which mean (organizations) is stemmed to (ظماً) which means (he was thirsty) instead of (نظم). (Hadni et al., 2013)
Snowball	-In many cases remove suffixed led to truncate the words, and change the form of the word.

Figure 3.2 Arabic text processing steps

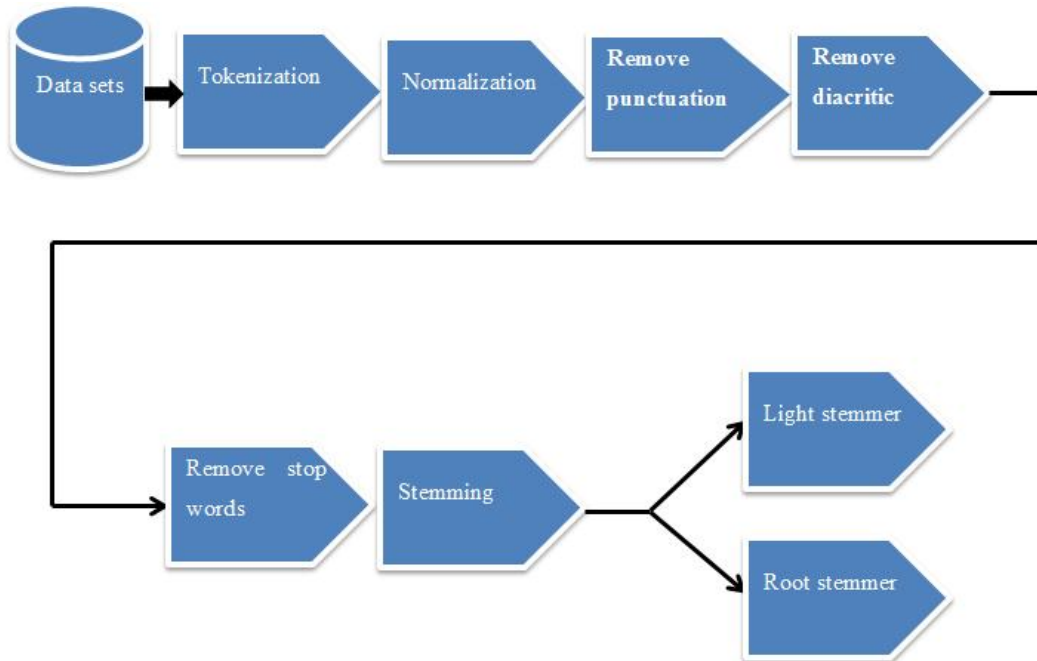


Figure 3.2 the steps of Arabic text processing

After the text processing phase has been done. We have two dataset light stemmer (Snowball) and another hand root stemmer (Shereen Khoja). Now two dataset is ready to generate the TFIDF and construction the vector space model.

3.2.2 Phase 2 Vector Space Model Construction (VSM)

The Vector Space Model (VSM) proposed by Salton (Salton, 1968) is a public technique for document representation in text classification. In this technique, each document is represented as a vector of features. Each feature is associated with a weight. Usually, these features are simple words. The feature weight can be simply a Boolean specifying the occurrence or nonappearance of the word in the document, its occurrence number in the document or it can be calculated by the formula number (3.3). (Chag heri, and etal, 2011)

Term Frequency (TF):

It gives us the occurrence of the word in each document in the corpus. It is a number of the word time's appearance in a document compared to the total number of words in that document. It increases when the number of occurrences of that word within the document frequently. Each document has its own TF.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad (3.1)$$

Inverse Data Frequency (IDF):

It Used to calculate the weight of rare words across all documents in the corpus the words that occur infrequently in the corpus have a high IDF score. It is given by the equation below (3.2).

$$idf(w) = \log\left(\frac{N}{df_t}\right) \quad (3.2)$$

Combining these two we come up with the TF-IDF score (w) for a word in a document in the corpus. It is the product of TF and IDF:

$$w_{i,j} = tf_{i,j} * \log\left(\frac{N}{df_i}\right) \quad (3.3)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing j

N = total number of documents

This research used rapid Miner to generate of TF-IDF in last section we provide overview about this software.

3.2.3 PHASE 3 TEXT CLASSIFICATION BASE ON SIMILARITY MEASURES

The similarity measure is defined as the distance between various data points. The performance of many classification or clustering algorithms depends upon selecting a good distance function over input dataset. While similarity is an amount that reflects the strength of the relationship between two data items, dissimilarity deals with the measurement of divergence between two data items(Patidar, Agrawal, & Mishra, 2012). Here, presenting a brief overview of the similarity measure functions used in this research.

3.2.3.1 Euclidean Distance

Euclidean distance is a standard metric for geometrical problems field. It is the regular distance between two points, and can be simply measured with a ruler in two or three dimensional space. Euclidean distance is broadly used in clustering problems, including text clustering. It is also the default distance measure used with the K-means algorithm. Measuring distance between text documents, give us two documents D_a and D_b represented by their term vectors \vec{t}_a and \vec{t}_b respectively, the Euclidean distance of the two documents is calculated in the equation below (Huang, 2008)

$$D_E(\vec{t}_a, \vec{t}_b) = \left(\sum_{t=1}^m |w_{t,a} - w_{t,b}|^2 \right)^{1/2} \quad (3.4)$$

Where the term set is $t = \{t_1, \dots, t_m\}$ mentioned previously, we use the *tfidf* value as term weights, that is $w_{t,a} = \text{tfidf}(d_a, t_a)$.

3.2.3.2 Cosine Similarity

When documents are represented as term vectors, the similarity of two documents matches the correlation between the vectors. This is measured as the cosine of the angle between these vectors, it's called cosine similarity. Cosine similarity is one of the most common similarity measure uses for text classification, Such as in numerous information retrieval applications and clustering too. Providing two documents \vec{t}_a and \vec{t}_b , can calculate cosine similarity using the following equation (3.5)

$$SIM_c(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a \cdot \vec{t}_b}{|\vec{t}_a| * |\vec{t}_b|} \quad (3.5)$$

Where \vec{t}_a and \vec{t}_b are multi-dimensional vectors over the term set $T = \{t_1, \dots, t_m\}$. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and limited between [0, 1].

An important property of the cosine similarity that its independence of document length. For example, combining two identical copies of a document d to get a new pseudo document d' , the cosine similarity between d and d' is 1(document itself), which means that these two documents are observed to be identical. Meanwhile, given another document l , d and d' will have the same similarity value to l , that is, $\text{sim}(\vec{t}_d, \vec{t}_l) = \text{sim}(\vec{t}_{d'}, \vec{t}_l)$. In other words, documents with the same composition

but different totals will be treated identically. when the term vectors are normalized to a unit length such as 1, and in this case, the representation of d and d' is the same.(H uang, 2008).

3.2.3.3 Pearson Correlation distance:

Pearson's correlation distance is another similarity measure of the extent to which two vectors are related. (Patidar et al., 2012) The distance measure could be mathematically specified as words W_a and W_b represented can calculate the correlation between them using the following equation (3.6)

$$SIM(X, Y) = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{(\sum x^2 - \frac{(\sum x)^2}{n})(\sum y^2 - \frac{(\sum y)^2}{n})}} \quad (3.6)$$

This is also a similarity measure. However, unlike the other measures, it ranges from -1 to +1 and it is 1. (Froud, Lachkar, & Ouatik, 2012)

TO apply these similarities measures and implement ask of text classification method in this research using MATLAB R2016a in last chapter we introduce overview about MATLAB.

3.2.4 Phase 4 Evaluation Measures

The classifiers model evaluated to determine classifier efficiency that means the average time to build classifier model or classification efficiency (the average time to classify an unseen document) or effectiveness (the average correctness classification)(Alnoukari, Alzoabi, & Sheikh, 2008). In this research evaluated the effectiveness of similarity measures using accuracy, recall, precision, and F1 measure. Which are widely used to evaluate supervised learning algorithms in TC (M. Al-Tahraw, 2016)(Alnoukari et al., 2008)(Al-tahrawi & Al-khatib, 2015)(Hadni et al., 2013).

Accuracy: the accuracy of class C_i , Acc_i is calculating using the following formula:

$$Acc_i = \frac{TP_i}{TP_i + FN_i + FP_i} \quad (3.7)$$

Where:

TP_i True Positive the number of documents classified correctly by the classifier as belong to the class C_i .

FP_i False Positive the number of documents classified incorrectly by the classifier as belong to the class C_i .

FN_i False Negative the number of documents classified incorrectly by the classifier as not belong to the class C_i .

Precision: the precision refers to the ratio of test files categorized into a class that belong to that class. The Precision of class C_i , P_i is calculating using the following formula:

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (3.8)$$

The recall is the ratio of test files belonging to a class and was claimed by the classifier as belonging to that class. The Recall of class C_i , R_i is calculating using the following formula:

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (3.9)$$

F1 measure is a balance between precision and recall the following formula is calculated the F1 measure.

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (3.10)$$

OR

$$F1 = \frac{2 \times recall \times precision}{recall + precision} \quad (3.11)$$

3.3 Dataset

The majority of researchers fully depend on collecting their data sets from online documents available on the internet that gathered from Al-Jazeera and other Arabic News channels

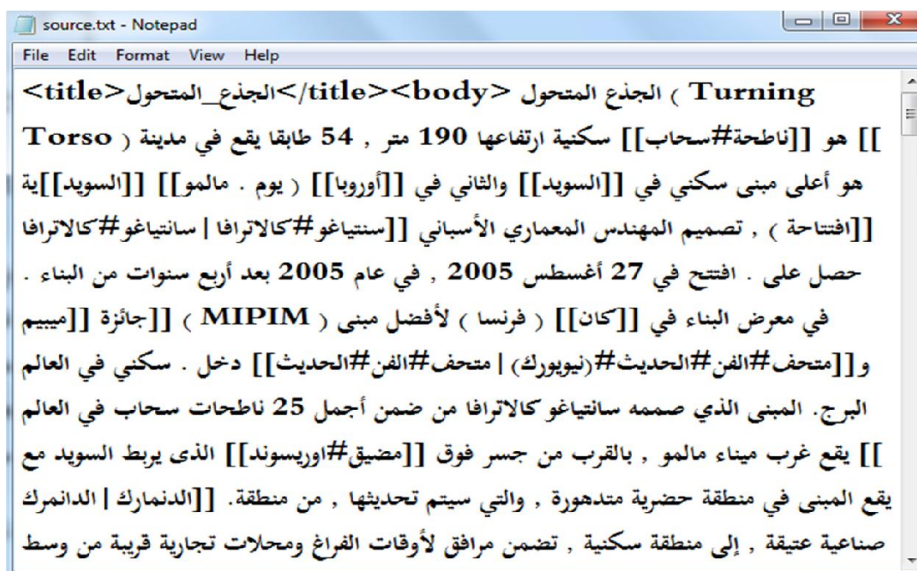
Arabic Wikipedia is the one of the benchmark dataset that available for the Arabic language. This corpus is consists of 4000 of Arabic Wikipedia articles that classified into nine classes (Facility (FAC), Geo-Political (GPE), Location (LOC), Not-Named-Entity (NOT), Organization (ORG), Person (PER), Product (PRO), Vehicle (VEH), and Weapon (WEA))(Alotaibi & Lee, 2012) named entities taxonomy. This dataset can be used in document classification tasks. (Yahya & Salhi, 2014) the table below determines the number of the document in each category.

Table 3.5 the number of the document in the data set

Category	Number of documents
FAC	129
GPE	611
LOC	98
NOT	1099
ORG	338
PER	1387
PRO	249
VEH	46
WEA	45
Total	4002

The figure below shows the Arabic text Wikipedia dataset before text processing phase . this text document has taken from FAC category.

Figure 3.3 snapshot of raw text



3.4 TOOLS AND TECHNOLOGIES

3.4.1 Python programming language

The python is an open source and high level programming language. it has standard libraries and dynamic typing and binding, contribute of rapid developing of programs, packages and integrating systems more professionally. also, it supports other libraries and allowances available on the web without any fees that gives it ability to be more fruitful and powerful tool in different fields like web development, game , Enterprise resources planning deployment, desktop programming, big data analysis, and ML applications. Python offers increased productivity, it does not need a code compiling; the maintenance progression is very fast. Any bug or incompatible input will never cause a failure. Instead, the interpreter will raise an exception and if the program does not catch the exception, the interpreter prints a stack trace. The python debugger allows checking of local and global variables, validating expressions, setting breakpoints, stepping through the code line by line, and more. Even with the fast debugging the python offers, the fastest debugging is by adding a few print statements to the source.(Alqarout, n.d.) The python language offer two different version python 2, and python 3. This research using python 3.6.5

3.4.2 RapidMiner

The RapidMiner tool offers much success within the research community and it is well known by most researchers for its easy to use and contains large collection of algorithms (Hadni et al., 2013). This research used RapidMiner to generate **TF-IDF**. **TF-IDF** stands for “Term Frequency-Inverse Data Frequency”.

3.4.3 MATLAB

MATLAB is a high-representation level language for technical performance. It integrates computation, contemplate, and programming in an easy-to-working area where problems and solutions are state in familiar notations. Standard utilize are following: Mathematics and estimation, creating algorithms, Data obtain and extensive used in Modeling and prototyping Data analysis. MATLAB is the computational utilize

choice for research, development and obtain, it has an image processing tools which are used in processing. MATLAB is the level of contemplate environment for most working fields. MATLAB has several other tools that are used in multi discipline such as mathematical, engineering, scientific ...etc. It also provide a Graphic User Interface (Sharma, 2014)

3.5 Summary

This section explains the text processing steps and introduce an overview of research methodology that Representative in similarity measures. Describe the calculation TFIDF and construct the VSM. Also, explain the evaluation of similarity measures using a confusion matrix. Finally, overview of dataset and mentioning the software tools and technologies that used in this research.

CHAPTER: IV

Implementation and results

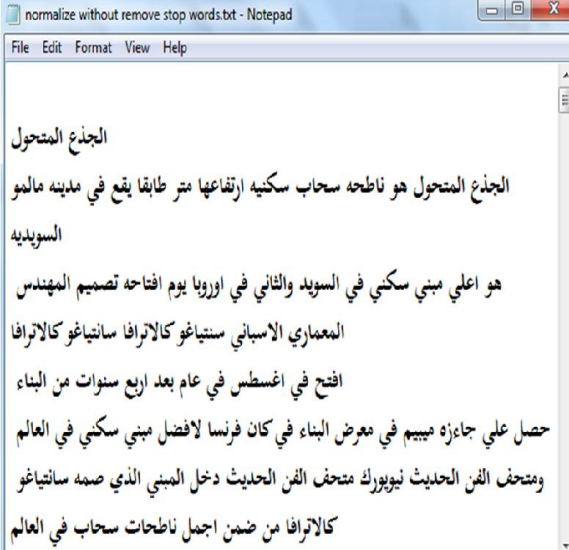
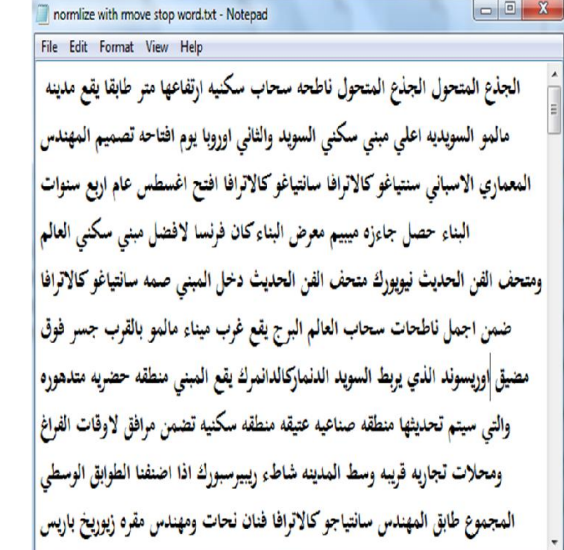
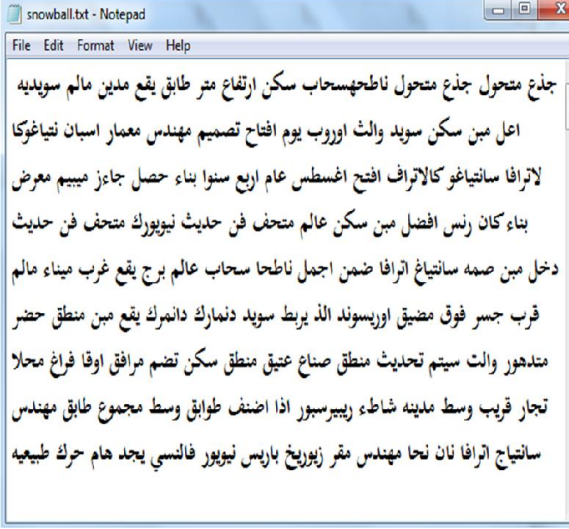
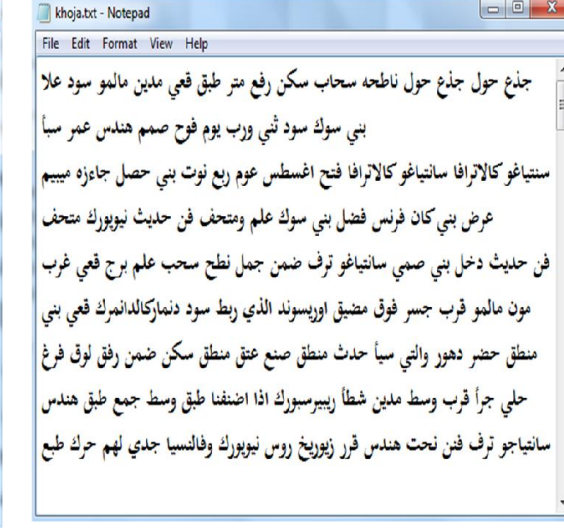
4.1 Introduction

This chapter discusses implementation model and how can evaluate the result according to the technique that mentioned in the previous chapter. The first section presents the pre-processing that described the implement of text preprocessing steps on the arabic dataset and generate TF-IDF wieght schema base on vector space model. The second section including the experimental design for root and stem stemmer approaches. the third section shows the result for both light and root stemmer corresponding to three similarity measures Euclidean distance, cosine similarity and correlation distance using evaluation measures recall, precision and f1 measure. Provide a summary and discussion of the result in the last section.

4.2 Pre-processing

All the steps of text pre-processing that mentioned in chapter three section one have been applied on the Arabic Wikipedia dataset. illustrate the following Table 4.1 his graph has been taken from text file within FAC category.

Table 4.1 implement the text preprocessing steps on the Arabic dataset

Step 1: Text normalization	Step 2 : step 1 with remove stop words
 <p>normalize without remove stop words.txt - Notepad</p> <p>الجذع المتحول الجذع المتحول هو ناطحه سحاب سكنيه ارتفاعها متر طابقا يقع في مدينه مالمو السويديه هو اعلي مبني سكني في السويد والثاني في اوربا يوم افتاحه تصميم المهندس المعماري الاسباني سنياغو كالاترافا افتح اغسطس عام اربع سنوات البناء حصل جائزه مبيم معرض البناء كان فرنسا لافضل مبني سكني العالم ومتحف الفن الحديث نيويورك متحف الفن الحديث دخل المبني صمه سانياغو كالاترافا ضمن اجمل ناطحات سحاب العالم البرج يقع غرب ميناء مالمو بالقرب جسر فوق مضيق اوريسوند الذي يربط السويد الدنماركالدانمرك يقع المبني منطقته حضرته متدهوره والتي سيتم تحديثها منطقته صناعيه عتيقه منطقته سكنيه تضمن مرافق لاوقات الفراغ ومحلات تجاره قريه وسط المدينه شاطئ ريبيرسورك اذا اضفنا الطوابق الوسطي المجموع طابق المهندس سانياغو كالاترافا فان نحات ومهندس مقره زيوريخ باريس</p>	 <p>normlize with move stop word.txt - Notepad</p> <p>الجذع المتحول الجذع المتحول ناطحه سحاب سكنيه ارتفاعها متر طابقا يقع مدينه مالمو السويديه اعلي مبني سكني السويد والثاني اوربا يوم افتاحه تصميم المهندس المعماري الاسباني سنياغو كالاترافا افتح اغسطس عام اربع سنوات البناء حصل جائزه مبيم معرض البناء كان فرنسا لافضل مبني سكني العالم ومتحف الفن الحديث نيويورك متحف الفن الحديث دخل المبني صمه سانياغو كالاترافا ضمن اجمل ناطحات سحاب العالم البرج يقع غرب ميناء مالمو بالقرب جسر فوق مضيق اوريسوند الذي يربط السويد الدنماركالدانمرك يقع المبني منطقته حضرته متدهوره والتي سيتم تحديثها منطقته صناعيه عتيقه منطقته سكنيه تضمن مرافق لاوقات الفراغ ومحلات تجاره قريه وسط المدينه شاطئ ريبيرسورك اذا اضفنا الطوابق الوسطي المجموع طابق المهندس سانياغو كالاترافا فان نحات ومهندس مقره زيوريخ باريس</p>
Step 2: with Snowball stemmer	Step 2: with Shereen Khoja stemmer
 <p>snowball.txt - Notepad</p> <p>جذع متحول جذع متحول ناطحه سحاب سكن ارتفاع متر طابق يقع مدين مالم سويديه اعل مين سكن سويد والث اورب يوم افتاح تصميم مهندس معمار اسبان نياغو كا لاترافا سانياغو كالاترافا افتح اغسطس عام اربع سنوا بناء حصل جائزه مبيم معرض بناء كان رنس افضل مين سكن عالم متحف فن حديث نيويورك متحف فن حديث دخل مين صمه سانياغ اترافا ضمن اجمل ناطحا سحاب عالم برج يقع غرب ميناء مالم قرب جسر فوق مضيق اوريسوند الذ يربط سويد دنمارك دانمرك يقع مين منطق حضر متدهور والت سيتم تحديث منطق صناع عتيق منطق سكن تضم مرافق اوقا فراغ محلا تجار قريه وسط مدينه شاطئ ريبيرسور اذا اضفنا طابق وسط مجموع طابق مهندس سانياغو ارافا نان نحا مهندس مقر زيوريخ باريس نيويور فالنسيا جدي لهم حرك طبيعيه</p>	 <p>khoja.txt - Notepad</p> <p>جذع حول جذع حول ناطحه سحاب سكن رفع متر طبق فعي مدين مالمو سود علا بني سوك سود ثني ورب يوم فوح صمم هندس عمر سبأ سنياغو كالاترافا سانياغو كالاترافا فتح اغسطس عوم ربع نوت بني حصل جائزه مبيم عرض بني كان فرنس فضل بني سوك علم ومتحف فن حديث نيويورك متحف فن حديث دخل بني صمي سانياغو ترف ضمن جمل نطح سحب علم برج فعي غرب مون مالمو قرب جسر فوق مضيق اوريسوند الذي ربط سود دنماركالدانمرك فعي بني منطق حضر دهور والتي سبأ حدث منطق صنع عتق منطق سكن ضمن رفق لوق فراغ حلي جراً قرب وسط مدين شطأ ريبيرسورك اذا اضفنا طبق وسط جمع طبق هندس سانياجو ترف فن نحت هندس قرر زيوريخ روس نيويورك وفالنسيا جدي لهم حرك طبع</p>

The Generate TF IDF requires a large memory to load all documents and products the VSM. Four gigabyte RAM has been used in this research to the generation of VSM. This size of RAM is not sufficient to handle all documents, therefore, some documents have been ignored. And the table below contains the number of documents after remove duplicate and large size file to increase the efficiency of TFIDF generation.

Table 4.2 number of documents after removing

Category	after remove duplicate and large size file
FAC	128
GPE	250
LOC	98
NOT	200
ORG	180
PER	200
PRO	150
VEH	46
WEA	45
Total	1298

Table 4.2 includes the number of words using light stemmer and root stemmer for each category after generate the TFIDF immediately.

Table 4.3 the number of words included in each stemmer

Category	Number of the word (light stemmer)	Number of the word (root stemmer)
FAC	13867	6623
GPE	7292	5580
LOC	10303	5580
NOT	8892	5600
ORG	8102	4603
PER	7096	4310
PRO	10034	4606
VEH	3937	2105
WEA	5969	3017
Total	75492	42024

After calculating the TF IDF in this research selecting the top hundred words for each category depending on the highest weight (TF IDF). That to reduce the amount of dimension and focus on important features. The table below explains the number of words after removing duplicates.

Table 4.4 the number of words after calculating TFIDF

Category	Number of the word for Snowball stemmer	Number of the word Shereen Khoja stemmer
FAC	100	100
GPE	96	100
LOC	95	91
NOT	95	88
ORG	90	82
PER	92	77
PRO	94	80
VEH	94	84
WEA	88	82
Total	844	784

Table 2 explains the number of the document included in each category after selecting the top hundred words from them. Some documents removed because the weight of all words within those documents is equaled zero.

Table 4.5 the number of documents after calculating TFIDF

Category	documents of Snowball stemmer	documents of Shereen Khoja stemmer
FAC	128	125
GPE	144	189
LOC	94	98
NOT	152	194
ORG	160	176

PER	165	197
PRO	130	147
VEH	45	46
WEA	45	45
Total	1063	1217

This research has two datasets split training and testing documents. The training dataset has 85% from the light and root dataset and 15% for testing documents. As shown the following Table4.5

Table 4.6 Number of training and testing documents using Snowball

Snowball							
Category	Number of document	first Experiment		second Experiment		third Experiment	
		training	test	training	test	training	Test
FAC	128	108	20	102	26	115	13
GPE	144	122	22	115	29	129	15
LOC	94	79	15	75	19	84	10
NOT	152	129	23	121	31	136	16
ORG	160	136	24	128	32	144	16
PER	165	140	25	132	33	148	17
PRO	130	110	20	104	26	117	13
VEH	45	38	7	36	9	40	5
WEA	45	38	7	36	9	40	5
Sum	1063	900	163	849	214	953	110

Table 3.7 Number of training and testing documents using Snowball

Shereen Khoja							
Category	Number of document	first Experiment		second Experiment		third Experiment	
		training	test	training	test	training	Test
FAC	125	106	19	100	25	112	13
GPE	189	160	29	151	38	170	19
LOC	98	83	15	78	20	88	10
NOT	194	164	30	155	39	174	20
ORG	176	149	27	140	36	158	18

PER	197	167	30	157	40	177	20
PRO	147	124	23	117	30	132	15
VEH	46	39	7	36	10	41	5
WEA	45	38	7	36	9	40	5
sum	1217	1030	187	970	247	1092	125

4.3 Implement of VSM

The figure below explains the weight (TF-IDF) for every word and construct the vector space model using RapidMiner.

The screenshot shows a VSM matrix with 45 rows (examples) and 9 columns (words). The words are: بون، بوس، يوم، بوند، بون، بونف، بونع، بونق، بونر. The matrix contains numerical values representing TF-IDF weights, with most cells being 0. The interface includes a filter dropdown set to 'all' and a scroll bar on the right.

بونر	بونق	بونع	بونف	بون	بوند	يوم	بوس	بون
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.028	0	0.012
0	0	0	0	0	0	0.024	0	0.030
0	0	0	0	0	0	0.014	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0.054	0	0	0.028
0	0	0	0	0	0	0.082	0	0
0	0	0	0	0	0	0.049	0	0.013
0	0	0	0	0.086	0	0.033	0	0
0	0.009	0	0	0	0	0.021	0	0
0.019	0	0.009	0	0.004	0.015	0.015	0.009	0.004
0	0	0	0	0.023	0	0.036	0	0
0	0	0	0	0.089	0	0	0	0
0	0	0	0	0.013	0	0	0	0.013
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

Figure 4.1 snapshot of the VSM – using root stemmer dataset

4.4 Experimental Design

In this research, have three experimental according to the document segmentation percentage into training and testing documents. The first experimental use 85% training documents and 15% test documents, the second experimental use 80% for training and 20% for test. And 90% for training,10% for test in the third experimental. And each experimental has implemented using two stemmers that divides the data set into snowball stemmer data set, and Khoja stemmer data set. The first experimental use the data set it has 844 words as attribute after generating the TFIDF from 900 documents

for training and 163 documents for testing and use Snowball stemmer. And use Shereen Khoja with 784 words as attribute dimension also after generating the TFIDF from 1034 documents for training and 183 documents for testing. The second experimental use Snowball with 849 training documents and 214 test documents, Shereen Khoja 970 for training and 247 for test. In third experimental use 953 training document and 110 for test with Snowball stemmer and use Shereen Khoja stemmer with 1092 for training and 125 for test.

In each experimental using Euclidean Distance (Distance), cosine similarity (cosine), and Pearson Correlation distance (Correlation) with two stemmers (Snowball and Shereen Khoja) as classifiers.

This research using the accuracy, recall, precision, and f1 measure to evaluate the result. The confusion matrix has been presented to count the TP, FP, and FN. The figure below include matrix has two dimensions; the first dimension present the predicted class that it is predicted by classifier, and it has nine columns (predicted class label), and the second dimension present the Actual class that it is actually existing within dataset, and it has nine rows (Actual class label).

The table 4.8 describes the confusion matrix for proposed method, and explains the FN, FP, and TP.

TP for the model is sum the diameter of the matrix is colored in yellow, wither the TP for certain class is a cell that is colored in yellow within the row of actual class. For example the TP for (NOT) class is 22 .

FN: for each class can calculate the **FN** by sum the row it includes the actual class label except a cell of **TP** that colored in yellow. For example the **FN** for (NOT) class is the sum of cells that colored in red = 7 .

FP: for each class can calculate the **FP** by sum the column it includes the predicted class label except the cell of **TP** that colored in yellow. For example the **FP** for (NOT) class is the sum of cells that colored in blue = 4 .

Recall: for (NOT) class = $TP / (TP + FN) = 22 / (22 + 7) = 0.76$

Precision: for (NOT) class = $TP/(TP + FP) = 22/(22+4) = 0.85$

F1 Measure for (NOT) class = $2 * Recall * Precision / (Recall + Precision)$
 $2 * .76 * 0.85 / (0.76 + 0.85) = 0.8$

Overall accuracy = sum of TP divided by sum of all the cells. In table below overall accuracy = $117/182 = 0.64$

Table 4.8 the Confusion matrix for Euclidean Distance with Shereen Khoja stemmer

		Predicted class									
		FAC	GPE	LOC	NOT	ORG	PER	PRO	VEH	WEA	
Actual class	FAC	18	0	0	0	0	0	1	0	0	
	GPE	5	20	0	2	0	0	0	1	0	
	LOC	4	4	4	1	1	1	0	0	0	
	NOT	1	0	0	22	0	5	1	0	0	
	ORG	2	0	0	1	16	4	3	0	0	
	PER	4	0	0	0	3	21	2	0	0	
	PRO	6	0	0	0	0	2	14	0	0	
	VEH	3	1	2	0	0	0	0	1	0	
	WEA	0	2	1	0	1	0	1	0	1	

4.5 Results of the first experimental

Table 4.9 contains the recall for each category using similarity measure distance, cosine and correlation the average of recall for each one (0.5 0.83 0.83) respectively. The (NOT) category in cosine and correlation are achieve similar result. in the Snowball stemmer the (PRO) category in distance measure were achieved a highest recall. (VEH) category have lower recall 0.29 when use distance measure. On the other hand Shereen Khoja stemmer for correlation measure achieves a higher average recall value 0.8 than other measures. The distance achieve less average recall value of 0.55 The cosine and correlation have the same recall for all categories except GPE and PER categories. NOT

category has the same recall value for all similarity measures. The GPE category has a high recall value 0.96 in the correlation measure. And VEH has a lower recall value 0.14 in the distance measure. The Figure 4.2 display the presentation graph.

Table 4.9 Recall for Snowball and Shereen Khoja stemmers

Category	Shereen Khoja			Snowball		
	Euclidean	cosine	correlation	Euclidean	cosine	Correlation
FAC	0.95	0.89	0.89	0.74	0.95	0.95
GPE	0.71	0.93	0.96	0.32	0.73	0.68
LOC	0.27	0.6	0.6	0.36	0.86	0.86
NOT	0.76	0.76	0.76	0.48	1	1
ORG	0.62	0.88	0.88	0.63	0.92	0.92
PER	0.7	0.83	0.9	0.56	0.56	0.56
PRO	0.64	0.86	0.86	1	0.8	0.8
VEH	0.14	0.86	0.86	0.29	0.86	0.86
WEA	0.17	0.5	0.5	0.67	0.83	0.83
average	0.55	0.79	0.8	0.56	0.83	0.83

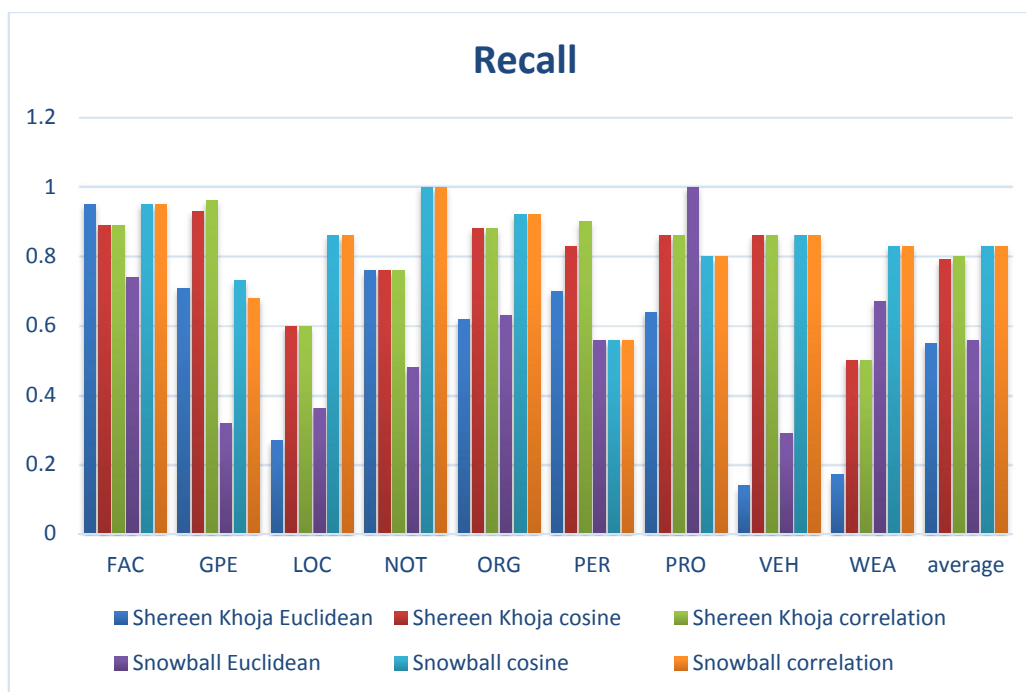


Figure 4.2 Recall for both stemmers with three similarity measures

Table 4.10 shows the result of precision for three similarity measures in each category. The average precision for distance is 0.85, correlation 0.84 and cosine achieved higher average precision is 0.85. In all category the average precision is same for cosine and correlation Except GPE, PER. And PRO. LOC, ORG has the same top precision value in cosine and correlation measures. PRO category gets the lower precision 0.26 in distance measure. Either Shereen Khoja stemmer the correlation measure has achieved a high average precision value 0.84 while the distance has lower average precision value of 0.63. the precision in the WEA category for all similarity measure is equal and consider the highest value 1. the lower precision in FAC category with distance measure 0.42. Figure 4.3 presents the precision in presentation graph.

Table 4.10 precision for both stemmers using three similarity measures

Category	Shereen Khoja			Snowball		
	Euclidean	cosine	correlation	Euclidean	cosine	Correlation
FAC	0.42	0.85	0.94	1	0.53	0.53
GPE	0.74	0.87	0.87	1	0.76	0.88
LOC	0.57	0.75	0.75	0.71	1	1
NOT	0.85	0.88	0.92	0.76	0.88	0.88
ORG	0.76	0.85	0.85	0.75	1	1
PER	0.64	0.78	0.79	0.93	1	0.93
PRO	0.64	0.79	0.79	0.26	0.94	0.8
VEH	0.5	0.67	0.67	1	0.86	0.86
WEA	1	1	1	1	0.71	0.71
average	0.68	0.83	0.84	0.82	0.85	0.84

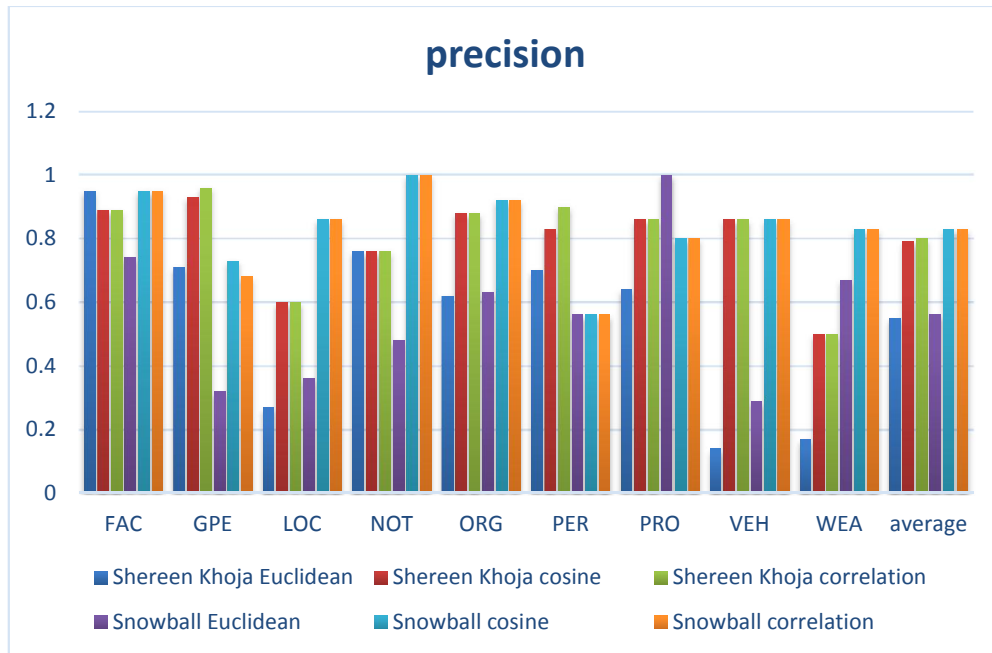


Figure 4.3 Precision for both stemmers with three similarity measures

Table 4.11 describes the F1 measure for each category depend on recall and precision. The cosine and correlation similarity measure have a similar result. ORG category achieves high F1 Measure value 0.96 in cosine and correlation measure. PRO category achieved less F1 Measure value 0.41 in the distance measure. The similarity measure distance, cosine and correlation have a similar value (0.7 0.72 0.7) respectively in PER category. And that explains in presentation graph Figure 2. On the other hand Shereen Khoja stemmer in FAC and GPE categories with correlation measure have the best F1 Measure value 0.91 together. VEH category and distance measure have lower result 0.22. The Figure 4.4 display presentation graph

Table 4.11 F1 measure for both stemmers with three similarity measures

Category	Shereen Khoja			Snowball		
	Euclidean	cosine	correlation	Euclidean	cosine	Correlation
FAC	0.58	0.87	0.91	0.85	0.68	0.68
GPE	0.72	0.9	0.91	0.48	0.74	0.77
LOC	0.37	0.67	0.67	0.48	0.92	0.92
NOT	0.8	0.82	0.83	0.59	0.94	0.94
ORG	0.68	0.86	0.86	0.68	0.96	0.96
PER	0.67	0.8	0.84	0.7	0.72	0.7
PRO	0.64	0.82	0.82	0.41	0.86	0.8
VEH	0.22	0.75	0.75	0.45	0.86	0.86
WEA	0.29	0.67	0.67	0.8	0.77	0.77
average	0.55	0.8	0.81	0.6	0.83	0.82

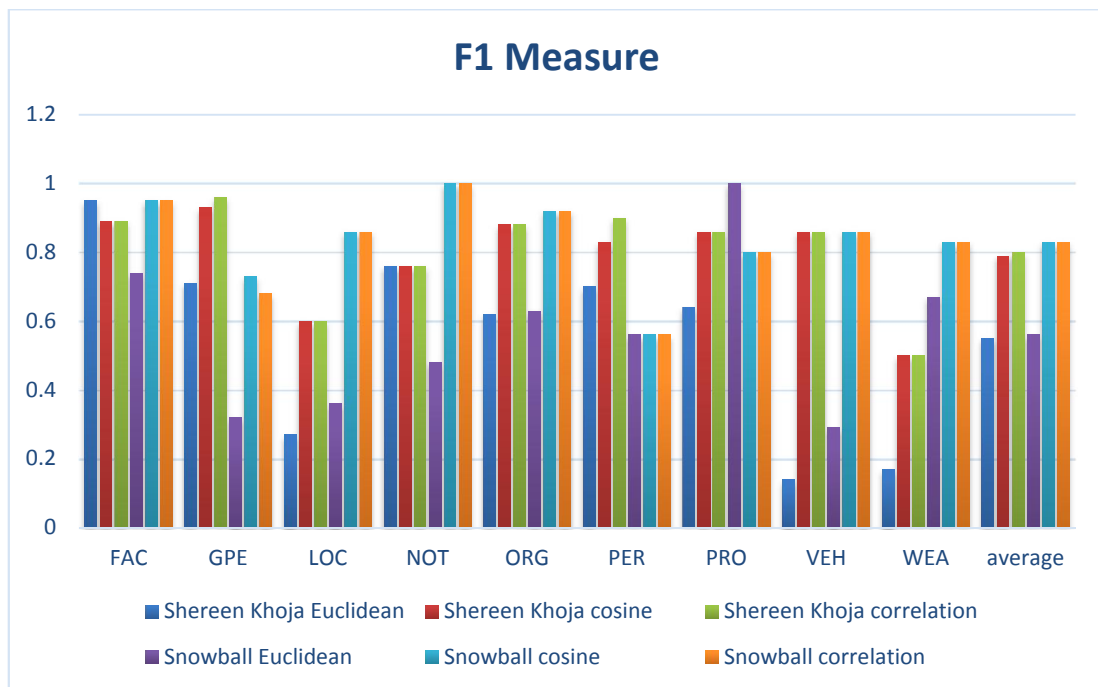


Figure 4.4 F1 measure for both stemmers with the three similarity measures

Table 4.12, shows the accuracy, recall, and F1 measure of similarity measures using Snowball stemmer, and Shereen Khoja stemmer. In Snowball stemmer the Cosine similarity achieved a high accuracy value 82.5% and distance similarity has low

accuracy 58.75%. Either in Screen Khoja stemmer the correlation measure achieves a high accuracy of 0.84. Either the distance measure has low accuracy 0.64. explain the result in presentation graph Figure 4.5

Table 4.12 the results of the first experimental

Snowball					Shereen Khoja			
measure	Overall Accuracy	Average of Recall	Average of Precision	Average of F1 measure	Overall Accuracy	Average of Recall	Average of Precision	Average of F1 measure
Euclidean	0.59	0.56	0.82	0.67	0.64	0.55	0.68	0.61
Cosine	0.83	0.83	0.85	0.84	0.82	0.79	0.83	0.81
correlation	0.82	0.83	0.84	0.83	0.84	0.8	0.84	0.82

The overall accuracy has been explained in Table 4.6

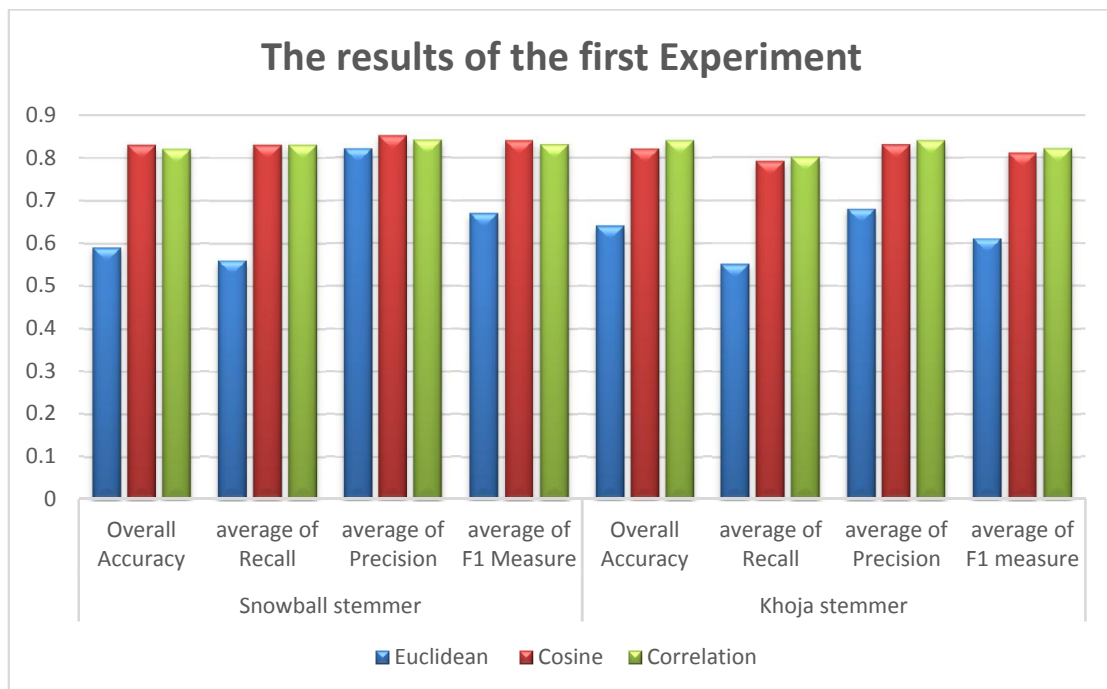


Figure 4.5 the result of the first experimental using Shereen Khoja and Snowball stemmers

4.6 Results of the second experimental

The table below summarize the results of overall accuracy this experimental is less than the first experimental. And the overall accuracy for cosine similarity have the same value 0.82 in two experimental when use Shereen Khoja stemmer. show presentation graph 4.6.

Table 4.13 the results of the second experimental

Measure	Snowball				Shereen Khoja			
	Overall Accuracy	Average of Recall	Average of Precision	Average of F1 measure	Overall Accuracy	Average of Recall	Average of Precision	Average of F1 measure
Euclidean	0.55	0.5	0.82	0.62	0.57	0.51	0.7	0.59
Cosine	0.81	0.81	0.84	0.82	0.82	0.78	0.83	0.8
correlation	0.81	0.81	0.83	0.82	0.83	0.79	0.84	0.81

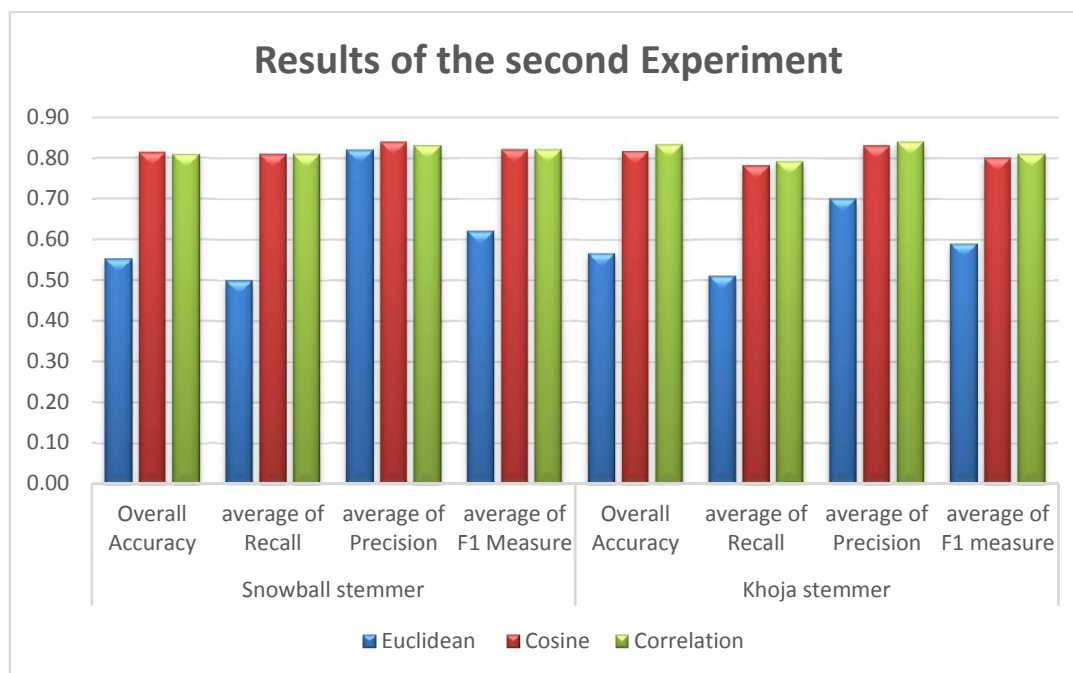


Figure 4.6 the result of the second experimental using Shereen Khoja and Snowball stemmers

4.7 Results of the third experimental

This experimental show that cosine and correlation measure achieved high overall accuracy 0.87 for each one. The results of Euclidean distance in this experimental better when comparing with two other experimental. Show presentation graph 4.7

Table 4.14 the results of the third experimental

Measure	Snowball				Shereen Khoja			
	Overall Accuracy	Average of Recall	Average of Precision	Average of F1 measure	Overall Accuracy	Average of Recall	Average of Precision	Average of F1 measure
Euclidean	0.73	0.71	0.86	0.78	0.70	0.62	0.63	0.62
Cosine	0.82	0.84	0.88	0.86	0.87	0.79	0.76	0.77
Correlation	0.81	0.83	0.86	0.84	0.87	0.79	0.76	0.77

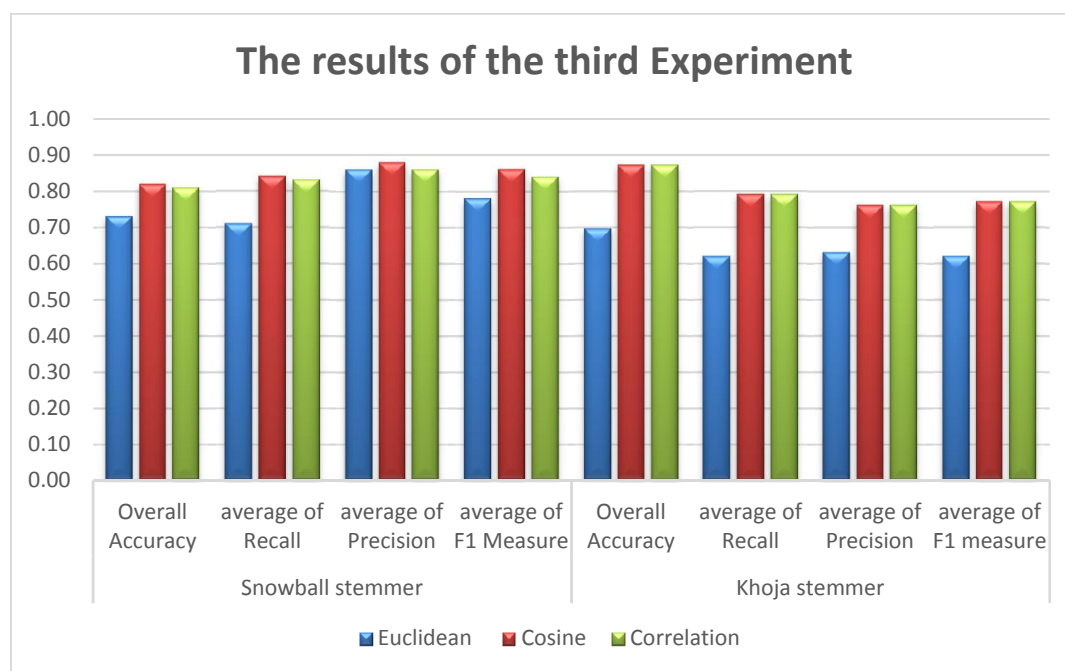


Figure 4.7 the result of the third experimental with Shereen Khoja and Snowball stemmers

4.8 Discussion

The results above have been divided into three experiments depend on documents segmentation percentage, use two stemmers with each one. The overall accuracy of Shereen Khoja stemmer are outperforming Snowball stemmer when use cosine similarity and Pearson Correlation have 0.87 for each one in the third experimental, and Euclidean Distance it is perform less overall accuracy in the second experimental. That means the overall accuracy is better when the gap between training set and test set is big.

In first experimental the overall accuracy of Shereen Khoja is better when use Euclidean Distance or correlation while cosine similarity is achieve high accuracy in Snowball stemmer. The average of recall and average of precision is better also when using Snowball stemmer. That means the Snowball stemmer is achieved high F1 measure than Shereen Khoja stemmer. The results of cosine similarity and correlation Euclidean distance in all experimental Very similar that referred to the method of calculating similarity for each one is depend on the calculate the correlation between the documents. In the results of Snowball stemmer Distance achieves high recall in the PRO category because the words in this category are high uniqueness than a word in other categories Table 4.10 Shows Distance achieve lower precision in PRO category refer to the category's false positive FP is high. the more false positive led to less precision. When false positive is high precision is low. In categories, FAC, GPE, VEH, WEA the false positive for all is zero that means the precision became the highest. Also In Shereen Khoja stemmer, the categories VEH WEA have less recall referred to datasets that includes a little number of the document for both that shown in Table 4.5 . The second experimental is provide bad results, clearly the overall accuracy is decrease while increase the amount of test documents. Offer less overall accuracy once use Euclidean Distance with Snowball is 0.55 and Shereen Khoja is 0.57, less average recall 0.5 and 0.51 and also less F1 Measure 0.59 . In this experimental the overall accuracy of Shereen Khoja stemmer is better.

The third experimental offer the highest overall accuracy 0.87 , recall 0.84 , precision 0.88 , F1 Measure 0.86. In all experimental when using cosine similarity is provide better results with Snowball stemmer in all term of recall and precision and F1 Measure

excluding the overall accuracy in the second and third experimental is better when use Shereen Khoja stemmer, also the less precision is 0.63 in this stemmer. And the overall accuracy of Shereen Khoja is better than Snowball.

CHAPTER: V

Recommendation and future works

5.1 Recommendations and Future Works

The huge text dataset will generate a large amount of text features (words). It's a difficult operation to calculate the TFIDF for each word. In this research, we recommend using a large size of computer RAM to handle all text features efficiently.

In future works compare using other Arabic stemmer's algorithms with other similarity measures and also using this study in other text mining application such as fraud detection, and email spam filtering.

5.2 Conclusion

The study aims to compare between two popular Arabic text stemmers algorithm Snowball, and Shereen Khoja stemmer using similarity measures Euclidean Distance, cosine similarity, and Pearson Correlation distance. Use Arabic Wikipedia dataset. It consists of 4002 Arabic Wikipedia articles classified into nine categories. Divided into two copies depending on stemming process; three experimental have been implemented the first divided into 85% documents for training and 15% document for the test. The second experimental have 80% training and 20% for test and the third experimental use 90% document for training and 10% document for test in each experiment use two stemmer (Snowball, Shereen Khoja). Generate the TFIDF weight schema and construct the vector space model. The Shereen Khoja stemmer achieves the best overall accuracy is 0.87 for cosine and correlation in the third experimental. While the overall accuracy of Snowball stemmer is 0.84 0.83 using cosine, correlation. My contribution is the root approach is more accurate when using similarity measures than stem approach and cosine similarity is the best similarity measure.

5.3 References

- A. Otair, M. (2013). Comparative Analysis of Arabic Stemming Algorithms. *International Journal of Managing Information Technology*, 5(2), 1–12. <https://doi.org/10.5121/ijmit.2013.5201>
- Ababneh, J., Almomani, O., Hadi, W., El-omari, N. K. T., & Al-ibrahim, A. (2014). Problem, 7(4), 219–223.
- Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. *Journal of King Saud University - Computer and Information Sciences*, 29(2), 189–195. <https://doi.org/10.1016/j.jksuci.2016.04.001>
- Al-Hashemi, R. (2010). Text Summarization Extraction System (TSES) Using Extracted Keywords. *International Arab Journal of E-Technology*, 1(4), 164–168.
- Al-Kabi, M. N., Kazakzeh, S. A., Abu Ata, B. M., Al-Rababah, S. A., & Alsmadi, I. M. (2015). A novel root based Arabic stemmer. *Journal of King Saud University - Computer and Information Sciences*, 27(2), 94–103. <https://doi.org/10.1016/j.jksuci.2014.04.001>
- Al-Tahrawi, M. M. (2013). The Role of Rare Terms in Enhancing the Performance of Polynomial Networks Based Text Categorization. *Journal of Intelligent Learning Systems and Applications*, 5(2), 84–89.
- Al-tahrawi, M. M., & Al-khatib, S. N. (2015). ig ht s y ig ht s. *Journal of King Saud University - Computer and Information Sciences*, 27(4), 437–449. <https://doi.org/10.1016/j.jksuci.2015.02.003>
- Alabbas, W., Al-Khateeb, H. M., & Mansour, A. (2016). Arabic text classification methods: Systematic literature review of primary studies. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)* (pp. 361–367). <https://doi.org/10.1109/CIST.2016.7805072>
- Almusaddar, M. Y. (2014). Improving Arabic Light Stemming in Information Retrieval Systems, 1.
- Alnoukari, M., Alzoabi, Z., & Sheikh, A. El. (2008). Y Ig Ht S Y Ig Ht S, 22(2), 295–322.
- Alotaibi, F., & Lee, M. (2012). Mapping Arabic Wikipedia into the Named Entities Taxonomy. In *Proceedings of COLING 2012* (pp. 43–52).
- Alqarout, B. O. (2017). *Parallel Text Classification Applied to Large Scale Arabic Text تصنيف النصوص العربية ذات النطاق الواسع على التوازي*
- Althobaiti, M., Kruschwitz, U., & Poesio, M. (2014a). AraNLP: a Java-based Library for the Processing of Arabic Text. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, (August 2015), 4134–4138. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/621_Paper.pdf

- Althobaiti, M., Kruschwitz, U., & Poesio, M. (2014b). AraNLP: a Java-based Library for the Processing of Arabic Text. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4134–4138. Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/621_Paper.pdf
- Bijalwan, V., Kumar, V., Kumari, P., & Pascual, J. (2014). KNN based machine learning approach for text and document mining. *International Journal of Database Theory and Application*, 7(1), 61–70. <https://doi.org/10.14257/ijta.2014.7.1.06>
- Chagheri, S., Calabretto, S., Roussey, C., & Dumoulin, C. (2011). Document Classification Combining Structure and Content. In *Proceedings of the 13th International Conference on Enterprise Information System (ICEIS)*. <https://doi.org/10.1006/37665>
- Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Information Retrieval*, 6(1), 57–72. <https://doi.org/10.1093/bib/6.1.57>
- Dang, S., & Ahmad, P. H. (2015). Text Mining : Techniques and its Application Text Mining : Techniques and its Application, (December 2014), 2–6.
- Dutta, D. kumar. (2017). Classification of Arabic text corpus using enhanced fuzzy c means algorithm, 116(10), 331–339.
- Elhassan, R., & Ahmed, M. (2015a). Arabic Text Classification on Full Word, 4(5), 114–120.
- Elhassan, R., & Ahmed, M. (2015b). Arabic Text Classification review. *International Journal of Computer Science and Software Engineering (IJCSSE)*, 4(1), 1–5.
- Froud, H., Lachkar, A., & Ouatic, S. (2012). A comparative study of root-based and stem-based approaches for measuring the similarity between arabic words for arabic text mining applications. *Advanced Computing An International Journal (ACIJ)*, 3(6), 55–67. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authType=crawler&jnl=2229726X&AN=84318567&h=3fA7TSXknVym/8I9+NDGIHP13a2PnGCKaXmexYZ0h+gjpXAS/oo3dYO8g0kNLdQTcGS7A/ChEzeVupMERrAwgg=&crl=c>
- Ghwanmeh, S., Kanaan, G., Al-Shalabi, R., & Ababneh, A. (2007). Enhanced Arabic Information Retrieval System based on Arabic Text Classification. In *Innovations in Information Technology, 2007. IIT '07. 4th International Conference on* (pp. 461–465). <https://doi.org/10.1109/IIT.2007.4430469>
- Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1), 60–76. <https://doi.org/10.4304/jetwi.1.1.60-76>
- Hadni, M., Ouatic, S. A., & Lachkar, A. (2013). Effective Arabic Stemmer Based Hybrid Approach for Arabic Text Categorization. *International Journal of Data Mining & Knowledge Management Process (IJDMP)*, 3(4), 1–14. <https://doi.org/10.5121/ijdkp.2013.3401>

- Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand*, (April), 49–56. Retrieved from http://nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf
- Kalita, D. (2015). Supervised and Unsupervised Document Classification-A survey. *International Journal of Computer Science and Information Technologies*, 6(2), 1971–1974. Retrieved from <http://ijcsit.com/docs/Volume6/vol6issue02/ijcsit20150602235.pdf>
- Kaur, S., & Aggarwal, D. (2013). Image Content Based Retrieval System using Cosine Similarity for Skin Disease Images. *Advances in Computer Science: An International ...*, 2(4), 89–95. Retrieved from <http://www.acsij.org/publications/acsij-2013-volume-2-issue-4/image-content-based-retrieval-system-using-cosine-similarity-for-skin-disease-images>
- Larkey, L., Ballesteros, L., & Connell, M. (2007). Light stemming for Arabic information retrieval. *Arabic Computational Morphology*, 221–243. <https://doi.org/10.1145/564376.564425>
- Li, B., & Han, L. (2013). Distance weighted cosine similarity measure for text classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8206 LNCS, 611–618. https://doi.org/10.1007/978-3-642-41278-3_74
- M. Al-Tahraw, M. (2016). Polynomial Neural Networks versus Other Arabic Text Classifiers. *Journal of Software*, 11(5), 418–430. <https://doi.org/10.17706/jsw.11.4.418-430>
- M., R., M., H., & Hussein, M. (2016). Improving Arabic Text Categorization using Normalization and Stemming Techniques. *International Journal of Computer Applications*, 135(2), 38–43. <https://doi.org/10.5120/ijca2016908328>
- Madani, A., & Kissi, M. (2017a). Arabic text classification using new stemmer for feature selection and decision trees, 12(May), 1475–1487.
- Madani, A., & Kissi, M. (2017b). Arabic text classification using new stemmer for feature selection and decision trees, (May).
- Mesleh, A. (2007). Support Vector Machines based Arabic Language Text Classification System : Feature Selection Comparative Study. *Science*, 228–233.
- Mohammad, A. H., Al-momani, O., & Alwada, T. (2016). Arabic Text Categorization using k-nearest neighbour , Decision Trees (C4 . 5) and Rocchio Classifier : A Comparative Study, 6(2), 477–482.
- Nguyen, H. V., & Bai, L. (2011). Cosine Similarity Metric Learning for Face Verification. *Computer Vision-ACCV 2010*, (Figure 1), 709–720. https://doi.org/10.1007/978-3-642-19309-5_55

- Otair, M. A. (2017). Classifiers, 7(1), 57–61.
- Patidar, A. K., Agrawal, J., & Mishra, N. (2012). Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach. *International Journal of Computer Applications*, 40(16), 975–8887. <https://doi.org/10.5120/5061-7221>
- Previtali, F., Arrieta, A. F., & Ermanni, P. (2015). Double-walled corrugated structure for bending-stiff anisotropic morphing skins. *Journal of Intelligent Material Systems and Structures*, 26(5), 599–613. <https://doi.org/10.1177/1045389X14554132>
- Sharma, V. (2014). Object Counting using MATLAB, 5(3), 614–616.
- Syiam, M., Fayed, Z., & Habib, M. (2005). {A}rabic text categorization using machine learning techniques. *Proc. of the 5th Conference on Language Engineering*, 9(3), 226–230.
- Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *IJACSA) International Journal of Advanced Computer Science and Applications*, 7(11), 414–418. Retrieved from www.ijacsa.thesai.org
- Tsai, F. S. (2011). *Text mining and visualisation of Protein-Protein Interactions. International Journal of Computational Biology and Drug Design* (Vol. 4). <https://doi.org/10.1504/IJCBDD.2011.041412>
- Wartena, C., & Brussee, R. (2008). Topic detection by clustering keywords. *Belgian/Netherlands Artificial Intelligence Conference*, 379–380. <https://doi.org/10.1109/DEXA.2008.120>
- Yahya, A., & Salhi, A. L. I. (2014). Arabic Text Categorization Based on Arabic Wikipedia, 13(1).