Republic of Iraq
Ministry of Higher Education and Scientific Research
University of Babylon
College of Information Technology
Software Department

Enhancement of Feature Selection Using Arabic Stemming Method

A Thesis

Submitted to the Council of the College of Information Technology for Postgraduate Studies of University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology - Software

By

Sebria Mohammed Hussien Aboob

Supervised by

Dr . Hazim J. Hassan Aburagheef

2022 A.D. 1444 A.H.

﴿ إِنَّ اللَّهَ وَمَلَائِكَتَهُ يُصَلُّونَ عَلَى النَّبِيِّ ۚ يَا أَيُّهَا الَّذِينَ آمَنُوا صَلُّوا عَلَيْهِ وَسَلِّمُوا تَسْلِيمًا ﴾

[الأحزاب: ٥٦]

قل سيروا في الأرض فانظروا كيف بدأ الخلق ۚ ثم الله ينشئ النشأة الآخرة ۚ إن

الله على كل شيء قدير ۚ

صدق الله العظيم [العنكبوت 20]

Declaration

I hereby declare that this dissertation entitled "**Enhancement of Feature Selection Using Arabic Stemming Method**, submitted to University of Babylon in partial fulfilment of requirements for the degree of Master in Information Technology \ Software, has not been submitted as an exercise for a similar degree at any other University. I also certify that this work described here is entirely my own except for experts and summaries whose source are appropriately cited in the references.

Signature:

Name: **Sebria Mohammed Hussein**

Date: / /2022

Supervisor Certification

I certify that the thesis entitled "**Enhancement of Feature Selection Using Arabic Stemming Method**" was prepared under my supervision at the department of Software/ College of Information Technology/the University of Babylon as partial fulfillment of the requirements of the degree of Master in Information Technology - Software.

Signature:

Supervisor Name: Dr . Hazim J. Hassan

Date: / /2022

The Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled "**Enhancement of Feature Selection Using Arabic Stemming Method**" for debate by the examination committee.

Signature:

Asst.prof.Dr. Ahmed Saleem

Head of Software Department

Date: / /2022

Certification of the Examination Committee

We, the undersigned, certify that (**Sebria Mohammed Hussien**) candidate for the degree of Master in Information Technology - Software, has presented his thesis of the following title (**Enhancement of Feature Selection Using Arabic Stemming Method**) as it appears on the title page and front cover of the thesis that the said thesis is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: 6 \ 9 \ 2022.

Signature:
Name: Dr.Abbas H Hassin Alasadi
Title: Professor
Date: / / 2022
(Chairman)

Signature:
Name: Dr. Rafid Sagban Abbood
Title: Professor Assistant
Date: / / 2022
(Member)

Signature:
Name: Dr. Sura Zaki Alrashid
Title: Professor Assistant
Date: / / 2022
(Member)

Signature:
Name:Dr. Hazim J. Hassan
Title : Lecturer
Date: / / 2022
(Member)

Signature:
Name: Dr. Hussein Atiya Lafta
Title: Professor
Date: / / 2022
(Dean of Collage of Information Technology)

Dedication

This work is dedicated to…

the owner of the age and time to hasten the appearance of the faraj come on.And the martyrs of Iraq of all sects and nationalities.

My father,

I will always be your daughter who is proud of you, and I will not disappoint you, my beloved father.

My mother,

I ask Allah to protect you from all evil and I will not disappoint you.

My Husband Abdul Hassan And Children

Whom I can't force myself to stop loving.

My beloved brothers and sisters; particularly my dearest brothers, Sajad And Asia Mercy and forgiveness for your pure souls.

Who stands by me when things look very difficult.

Acknowledgement

Foremost, I am highly grateful to Allah (God) for His unlimited blessings that continue to flow into my life.

To my family, and my friends: I give you all thanks for your belief in me, your constant support, encouragement, and cooperation at all times.

The Head of the Department Certification Asst.prof.Dr. Ahmed Saleem very very thank you.I hope from the God to give your above dgree for your life.

I would like to thank all the kind, helpful and lovely people who helped me directly or indirectly to complete this work and apologize to them for not being able to mention them by name here, but they are in my heart.

Abstract

Stemming operation is an essential step in pre-processing step. Its operation is used to reduce the dimensional of features by reducing the number of features. It returns each word into root words. It is the process of removing the collection of prefixes and suffixes in the Light stem method and summation of the word to produce the root in the Complex stem method. It converts plural to vocabulary.

Accordingly, this thesis aims Enhancement of feature selection using Arabic stemming, applying Light-based stemmer and Root-based stemmer approachs and text mining techniques.

The proposed system consists of five major stages. The first is data abtain from sanad subset. The second is data pre-processing that was performed by applying different methods. The third stage involves features extraction and selection by using Term Frequency-Inverse Document Frequency (TF-IDF) approach. Fourth, classification methods were applied using Random Forest (RF). Moreover, the findings of the classification algorithms were evaluated based on performance metric using accuracy, precision, recall, and F1-score measurements. The accuracy in light , precision, recall, and F1-score are 0.9662%, 0.9659%, 0.9659%, and The accuracy in complex, precision, recall, and F1-score are 0.9613%, 0.9611%, 0.9611%.

A result of evaluation appears the complex stemmer is best than light stemmer. It reduces vector size from total term size 152749 to 15628 then 50776 in light stemmer, but in accuracy, light stemming is stay best than complex stem. It's more than complex stemmer in 0.01.

Acknowledgement

My father taught me that all men are angels and you are one of them. This work is an ongoing charity for your pure soul. Al-Fatiha with prayers.

Table Of Contents

List of Tables

List of Figures

List of Abbreviations

DC	Document Classification
DF	Document Frequency Threshold
DM	Data Mining
DT	Decision Tree
FN	False Negative
FP	False Positive
FS	Feature Selection
ML	Machine Learning
NLP	Natural Language Processing
PCA	Principle component analysis
RF	Random Forest
DT	Decisen Tree
SA	Sentiment Analysis
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TM	Text Mining
TN	True Negative
TP	True Positive

List of Algorithms

CHAPTER ONE

Introduction

INTRODUCTION

1.1 Overview

Text mining is an area of computer science that applies data mining, information retrieval, machine learning, statistics, natural language processing, and knowledge management approaches to text. Categorization of newspaper articles or web pages, e-mail filtering, library organization, customer complaints (or feedback) handling, marketing focus group programs, competitive intelligence, market prediction, extraction of topic trends in text streams, discovering semantic relations between events, and customer satisfaction analysis are just a few examples of typical applications. Text mining entails tasks like text categorization, phrase extraction, document summarization (single or multi-document), clustering, association rules mining, and sentiment analysis [1].

The rapid expansion of the data dimensionality creates problems in handling it. Sometimes, dataset contain redundant, irrelevant, and noisy data, which slow down the model and minimize the classification accuracy [2]. Feature selection could solve this problem, Different methods have been used for feature selection, such as filter, wrapper, and embedded.

Many methods have been used for handling different types of optimization problems. Stemming can be used to improve the problem of feature selection. In Arabic text pre-processing, stemming algorithms can be used to reduce many versions of a word to a single form (root or stem). The current stemmers aren't particularly efficient.

1.2 Stemming:[3]

Stemming can be defined in a variety of ways. According to Shereen Khoja's definition, stemming for Arabic language is limited to root extraction. She describes the stemming procedure as "... The process of removing all of a word's affixes to produce the stem or root is known as stemming". This refers to remove of prefixes, suffixes, and infixes pores in Arabic. Also suggested that a stemmer can use a combination of these methods Light stemming and root stemming are the two basic techniques to stemming in Arabic. The affix removal strategy, also known as stemming, refers to the practice of removing a small number of prefixes and/or suffixes in order to find the root of the word.

1.2.1 Match and Truncate (MT) [3]

The beginning of a word is eliminated, and the rest of the words are longer than three letters.

1. Remove and Check (R.C.): If a match is found and the remaining word is found in the document collection, the first letter of the word is removed.
2. RW (Remove With Other Letters): removing a combination of particles as well as the definite and related article Larkey divided the Arabic stemmers into four types.Dictionaries that were created by hand Prefixes and suffixes are removed using algorithmic light stemmers. Analyses morphologically based on the search for roots.
3. Statistical stemmers use clustering algorithms to group words variations.

4. Parallel corpora and new statistical approaches.

1.2.2 Explaining Essential Terms Used Throughout The Proposal:

1.**Dataset:** is a collection of documents.

2.A document is a set of tokens.

3.**Token**: is an instance of a sequence of one or more characters creating separate, independent regions in plain text separated from other regions using a delimiter used as a functional semantic unit for processing.

4.Delimiters: is defined as one of the following items in the set:

{ "”", "’", "+", "-", "_", "‹", "*" ,"۞" ,"،" ,"؟" ,"؛"}

5.Encoding: To process Arabic characters, the encoding system must be compatible with both the language and the system used to process the

6.text; Unicode (UTF-8) is chosen as the encoding system for its broad compatibility.

7.Arabic Character: is a Unicode (UTF-8) encoding character with a decimal value smaller or equal to 1610 or bigger or equal to 1568.

8.Non-Arabic character is any other character not in the range of values defined in the step above.

9.Term: is a class of all tokens containing the same character sequence normalized and included in the system dictionary of terms.

10.Feature: is an individual measurable property or characteristic of a text like a pattern of a numerical representation that distinguishes it from other text; in this case, it is a vector of numerical value extracted using an algorithm to represent the statistical frequency of terms used in a document. [1].

1.3 Research Problem

The number of problem can be seen in the thises:

1. Using the Light-based Techniques to Enhance Words in Arabic Text in order to shorten the stemming time.
2. stemming and root-based stemming are used.
3. The phase (pattern matching) is to extract the stem and root of the word by comparing it to the Arabic patterns without its affixes. This phase is used to extract the word's stem and root, Let "فعل " be the length of the word after prefixes and suffixes have been removed, and the patterns with lengths equal to "فعل " have been chosen.
4. The stem will be matched with the particular pattern for each selected pattern in order to compute the repetitions for each word.

1.3 Research Questions

The following questions are addressed in this thesis:

1. How to get an Arabic text free of stop words, non-Arabic words, symbols and numbers.
2. How to get an Arabic text with words free of prefixes and suffixes.
3. What is the method used to get the verb in the past form.

1.4 Thesis Objectives

The study aims to improve the features selection method as well as testing the proposed algorithm's performance.The objects can describe as:

1. Stemming understanding methods includes light-based and root-based stemming

2. Improving of root-based stemming and reduce error generation .The stem will be matched with the particular pattern for each selected pattern in order to compute the repetitions for each word.
3. Feature selection reducing by enhance stemming operation,is using (Flter,PCA).
4. Identify the optimal set of features that will enable the creation of a model that will be beneficial in classroom settings.
5. Improving text classification by enhance stemming operation.

1.5 Research Significance and Contributions

There are a variety of outcomes from this endeavor that can be summarized as follows:

Arabic stemming comprehension and careful examination of various ways and Arabic stemming examines the failures of past techniques. And then, analysis of the impact of the stemming process on feature selection and text classification.

Arabic Pattern are used to design appropriate strategies for treating Arabic stemming errors.

1.6 Research Challenges

Stemming is divided into two types: light stemming and root-based stemming. Prefix and suffix are removed from words using light stemming. Root- based stemming return words into root.

1. Light stemming: there are two main problems with light stemming, first with word length less or equal three latter's and second with words length more than six latter.

2. Root-based stemming has major drawbacks in interpreter root words, such as:

a. Depending on its place and context in the text, a single word might have multiple meanings (ذهب, gold, or went).

b. Some words have the same root but various meanings; all of the following terms (مدرس, teacher) have the same root can be show in Figure (1.1).

Figure (1.1): the same root.[4].

c. Another root-based stemming issue with broken plurals.

1. 8 Related Work

Arabic stemming methods can describe in briefly as:

1. Khoja, S. and Garside, R. [5] (2001). Eliminated the affixes from the word and extracting the root or stem of the word The root-based strategy, the light stemmer approach, and the statistical stemmer approach are the three different stemming approaches for Arabic (n- garm). Using a dictionary of 4,748 triliteral and quadriliteral words, the stemmer achieves a 97 percent accuracy rate on Arabic terms.

2. Waed Al-Abweeny and Nahid Abu Zaid [6],2018. provided a method for eliminating the word's affixes and extracting the root or stem . The stemmer's job is to diminish the size of the stem. Due to the complexity of the language, Arabic stemming is a

difficult undertaking. Morphological versions of words that aren't usually thematically related This is a paper about provides an Arabic stemmer system that extracts trilaterals using Arabic rules (three radicals),Quadrilateral (four radicals), quintuple (five radicals), and hexagonal (six radicals) if the number of radicals is four. Available It was tested on four Arabic native speakers and found a 96.8% accuracy rate.

3. Yasir Alhanini and M. Aziz, [7], Existed stemmers appeared to have ignored the processing of multi-word phrases and the recognition of Arabic names. Enhance is used stemming, which is based on light stemming and a dictionary-based stemming approach, to extract the stem of Arabic words. The improved stemmer can now handle multi-word expressions and detects named entities. Arabic dataset is used with ten documents to evaluate the enhanced stemmer. The results of the accuracy tests were presented. The average accuracy in improved stemmer on the corpus is 96.29 percent for each document's enhanced stemmer, light stemmer, and dictionary-based stemmer.

4. Rouhia M. Sallam, et al 2016,[8], The Frequency Ratio Accumulation Method (FRAM) is a classifier used in the suggested method. Its features are expressed using the bag of words technique, and the features are chosen using an enhanced Term Frequency (TF) technique. Known datasets are used to test the suggested approach. The trials are conducted without, with one of, and with both stemming and normalizing. In comparison to current procedures, the proposed methodology often produces better outcomes. Accuracy, Recall, Precision, and F-measure were taken into consideration as performance aspects of the suggested Arabic Text Categorization technique (F1). When normalization is

used, the averages of the findings obtained are 97.50%, 97.50%, 97.51%, and 97.49%, respectively.

5. Ali Alnaieda et al at [9], (2020), The Arabic Morphology Information Retrieval (AMIR) method, which is a novel method for Arabic stem generation and extraction, is discussed in this paper. AMIR applies a set of rules regarding the relationship between Arabic letters to identify the root or stem of the corresponding words that are used as indexing terms for text searches in Arabic retrieval systems. His emphasize the advantages of the suggested rules for various Arabic information retrieval systems to show the value of the provided method. In terms of mean average precisions, we have evaluated the AMIR system by contrasting it with that of LUCENE, FARASA, and the no-stemmer counterpart system. According to the results, AMIR has a mean average precision of 0.34%, compared to LUCENE's 0.27%, FARASA's 0.28%, and no stemmer's 0.21. This proves AMIR's existence to improve Arabic stemmer and increases retrieval as well as being strong against any type of stem.

6. Rafea Mohammed [10] (2016),The root-based method and the stem-based approach are the two basic types of Arabic stemming algorithms. Both forms have issues that have been handled in the proposed stemmer, which uses Arabic patterns (Tafealat1) to discover the extra letters and combines the rules of both types.

7. Essam K.Mohammed Al-Yasir, [11]. (2019), In the suggested system, there are five basic steps. Data collection was the first phase, which required getting tweets from Twitter. The second phase was data pre-processing, which was accomplished in a variety of methods. In the following stage, The Term

Frequency-Inverse Document Frequency (TF-IDF) technique is used to extract features. As a result, the text was converted into a vector of features. In addition, the Linear Support Vector Classifier (LSVC), Decision Tree (DT), and Random Forest classification algorithms were used (RF). These three strategies were utilized to determine which was the most accurate for classifying SNSs data. This research contributes to a better understanding of terrorist organizations' use of social media in the Arab world. Accuracy is 96 %.

8. Aqil M. Azmi and Huda Abdularhman Almuzaini [12] (2020), them provided seven deep learning-based systems in number eight. They're known as Convolutional Neural Networks. For word representation, they employed CNN-LSTM, CNN-GRU (Gated Recurrent Units), BiLSTM (Bidirectional LSTM), Att-LSTM (Attention-based LSTM), and Att-GRU, as well as the word embedding technique (Word2Vec). They tested our strategy on two big datasets with six and eight categories using ten-fold cross-validation.

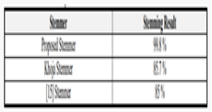
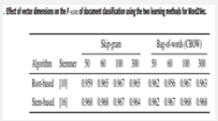
9. Riyad Alshammari [4], 2018.To achieve high accuracy in Arabic text categorization, the current pre-processing procedures utilized to prepare the data set are crucial. As a result, this study investigates the impact of pre-processing methodologies on the performance of three machine learning algorithms, namely Nave Bayesian, DMNBtext, and C4.5. In categorizing Arabic text, the DMNBtext learning system outperformed other machine learning algorithms.according to the results.

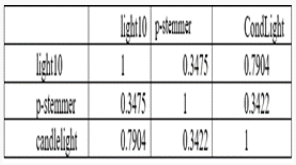
10. Sabria M. Hussien, Hazim J. Aburagheef [13]. Arabic stemming is an important part of the natural language preparation process (NLP). It takes affixes out of words. It improves text

classification (TC) as well as retrieval of information (IR). Light-based stemming and root-based stemming are the two types of stemming. The energy required for light-based stemming is more than that required for root-based stemming. Only suffixes and prefixes have been removed from the terms. The light10 stemmer, the p-stemmer, and conditional light stemming are three well-known light stemming methods (CondLight). Under certain circumstances, Light10 stemmers remove prefixes and suffixes. The P-stemmer removes only prefixes, but the CondLight stemmer adds eight conditions to the Light10 stemmer. The stemmers were used to determine how far Arabic TC had progressed. The Support Vector Machine (SVM), the k-nearest neighbor Three classifiers usage, and statistical similarity measurement. The stemming result shows a slight improvement (about 2 percent improvement).

Table (1.1) : Related work

num	Name 0f resercher	year	Titel of reserch	Research Object	ACCURACY
1	**Khoja, S. and Garside, R.[6]**	**(1999)**	**Stemming Arabic Text". Computing Department, Lancaster University, Lancaster**	using a process of removing the affixes from the word and extracting the word root or stem For Arabic Language, there are three different Stemming approaches: the root-based approach (exp Khoja[5]); the light stemmer approach (Larkey[7]); and the statistical stemmer approach (n- garm).	a 97 percent accuracy rate
2	Essam Kazem Mohammed Al-Yasiri[12]	2019.	,“Arabic Sentiment Analysis for Identifying Terrorism Supporters on Twitter Using Data Mining Techniques”,	The proposed system consists of five major stages. The first is data collection which focused on fetching tweets from Twitter. The second is data pre-processing that was performed by applying different methods. Another stage involves features extraction by using Term Frequency-Inverse Document Frequency (TF-IDF) approach.	99 % in light stemming

3	Y. Alhanini and M. Aziz,[8]	2011,	"The Enhancement of Arabic Stemming by Using Light Stemming and Dictionary-Based Stemming,"	The Enhancement of Arabic Stemming by Using Light Stemming and Dictionary-Based Stemming," Journal of Software Engineering and Applications	96.29
4	Rafea Mohammed[11] ,	2016.	" New Arabic Stemming based on Arabic Patterns"	“Algorithms for Arabic stemming available in two main types which are root-based approach and stem-based approach. Both types have problems which have been solved in the proposed stemmer	
5	**Huda Abdulrahman Almuzaini and Aqil M. Azmi [13],**	2020	The goal was to see how different stemming tactics and word embedding effect classification.	The objective were to study how the classification is affected by the stemming strategies and word embedding. They present seven deep learning-based algorithms to classify the Arabic documents. Them applied the word embedding technique (Word2Vec). Them tested our approach on two large datasets-with six and eight categories-using ten-fold cross-validation.	
6	**Waed Al-Abweeny 1 and Nahed Abu Zaid 2[6]**	**2018**	**” Arabic Stemmer System based on Rules of Roots “**	**Extracting the word root or stem . Stemmer is an automated process that generates a base string in an attempt , and it's the first stage in many sorts of data processing applications like text mining, information retrieval, and natural language processing.**	**96.8% accuracy**
7	Riyad Alshammari [4]	2018	“ Arabic Text Categorization using Machine Learning Approaches “,	The DMNBtext learning algorithms are achieving higher performance when Khoja and Light stems with boolean, idf, tfidf, wc and wc-norm preprocessing methods on the BBC data set.	The DMNBtext attained an accuracy of 99% on the BBC data set. On the other hand, evaluating the DMNBtext on the CNN data sets, it produced higher performance using Light stem and Raw Text with boolean, idf, tf, tfidf and wc-norm preprocessing methods. The DMNBtext achieved an accuracy higher than 93% on the CNN data set
8	Ali Alnaieda et al [9]	2020	An intelligent use of stemmer and morphology analysis for Arabic information Retrieval	"The Arabic Morphology Information Retrieval AMIR applies a set of rules regarding the relationship between Arabic letters to identify the root or stem of the corresponding words that are used as indexing terms for text searches in Arabic retrieval systems	AMIR has a mean average precision of 0.34%, compared to LUCENE's 0.27%, FARASA's 0.28%, and no stemmer's 0.21. This proves AMIR's existence to improve Arabic stemmer and increases retrieval as well as being strong against any type of stem
9	4. Rouhia M. Sallam, et al at [8]	2016	Comparative Experimental Study of NLP Tools and Language	The Frequency Ratio Accumulation Method (FRAM) is a classifier used in the suggested method. Its features are expressed	When normalization is used, the averages of the findings obtained are 97.50%, 97.50%, 97.51%, and 97.49%, respectively.

			Resources for Arabic“,	using the bag of words technique, and the features are chosen using an enhanced Term Frequency (TF) technique..	
10	Sabria M. Hussien, Hazim J. Aburagheef[13]	2021	Arabic light-based stemming: a comparative study among ligh10 stemmer, P-stemmer, and Conditional light stemmer.	The light10 stemmer, the p-stemmer, and conditional light stemming are three well-known light stemming methods (CondLight).	

	light10	p-stemmer	CondLight
light10	1	0.3475	0.7904
p-stemmer	0.3475	1	0.3422
candleLight	0.7904	0.3422	1

This present thesis, however, extends previous literature and improves the accuracy of identifying Feature selection. This is clear as we use our dataset without using previously available data where the research dataset was downloaded based on a high variety of keywords. Such keywords help in obtaining better identification of language used by Arabic texts. Moreover, this study achieves high accuracy in Arabic text analysis. A possible reason for that is the use of several pre-processing techniques before classifying texts. This might help clean data from unnecessary information. Thus, this study adds to previous research by combining the identification of the best feature from the Arabic texts. where it is based on stemmer and Evaluations. [11].

1.7 Thesis Organization

There are five chapters in this thesis. Each chapter opens with a brief summary that provides a broad overview of the topic.The key contents of other chapters are as follows:

1. **Chapter one :” Introduction” .**

2. **Chapter Two**: entitled **" Theoretical Background"**. This chapter covers data mining ideas, social networking sites, S.A., text mining, text mining methods, classification algorithm evaluation.
3. **Chapter Three:** entitled **"The Proposed System"**. It contains information about the proposed system as well as its.
4. **Chapter Four**: entitled **"Results and Discussion"**. The outcomes of the suggested system and the research experiments are presented in this chapter. It also goes over how to assess the system's performance.
5. **Chapter Five**: entitled **"Concluisons and Works Future "**. This chapter summarizes the findings of the study as well as potential research directions that could be pursued to improve the work.

CHAPTER TWO

Theoretical Background

Theoretical Background

1.8 Introduction

This chapter explains the theoretical background regarding the concepts of Arabic text stemming and natural language processing (NLP). Furthermore, it discusses the main steps in classifying texts. Data pre-processing, features extraction, and feature selection techniques. Arabic text Concepts. [14].

Arabic stemming affects the performance of a variety of natural language processing applications, including part of speech tagging, syntactic parsing, machine translation systems, and information retrieval systems. Computational stemming is a significant topic in Arabic Natural Language Procesing since Arabic is a strongly inflected language. Multi-word expressions and Arabic nams were not handled well by the stemmers. The stem of Arabic words is extracted using enhanced stemming, which is based on light steming and a root-based steming approach. The upgraded stemmer now recognizes named entities and can handle multi-word expressions. In order to assess the upgraded stemmer. [7].

To prepare vowelized and unvowelized Arabic texts for Natural Language Processing, use this method (NLP). The four phases of this rule-based technique are text tokenization, word light stemming, word morphological analysis, and text annotation. first isolates the words from the input text and represents them in a formal manner. Second uses a light

stemmer to remove prefixes and suffixes from each word, allowing for the extraction of the stem In the third stage, a rule-based morphological analyzer determines the root and morphological pattern of each extracted stem. The final phase aims to build an annotated document with definitions for each word. morphological characteristics The preprocessor described in this study can handle both vowelized and unvowelized words and provides the information words as well as the semantic data required by diverse applications. It's designed to work with a variety of NLP applications, including machine interpretation, text synopsis, text repair, data recovery, and more. programmed vowelization of Arabic text. [14].

1.9 Arabic Stemming

The process of turning texts into a format that may be classified is known as text pre-processing. Text pre-processing is critical because it may be used to reduce feature space and speed up the computing process, which can improve classification accuracy. To prepare texts for classification, many pre-processing processes are carried out. Non-Arabic words, special characters, symbols, usernames, repeated and elongated letters, numbers, digits, and punctuations, for example, are all eliminated. Unnecessary data can also be deleted, which affects the classification algorithm's efficiency and accuracy [11].

Stemming: Stemming can be defined in a variety of ways. According to Shereen Khoja's definition, stemming for Arabic language is limited to root extraction. She describes the stemming procedure as "... The process of removing all of a word's affixes to produce the stem or root is known as

stemming. Prefixes, suffixes, and infixes are all removed in Arabic. [3]. is the process back words to their root. For example, the common root for "**يلعبون, لعبه**" and **لاعب** "is '**لعب**'.

Many natural language processing applications, such as part of speech tagging, syntactic parsing, machine translation systems, and information retrieval systems, are affected by Light-based stemming and root-based stemming are the two basic ways to stemming in Arabic. [11].

2.2.1 light-Based Stemming

For the Arabic language, numerous stemming algorithms are used, one of which is the light stemmer algorithm. The root-based algorithm is a more aggressive method. The goal of this method is to remove the most common suffixes and prefixes, rather than producing the linguistic root of a particular Arabic surface form. In Arabic, unlike English, both prefixes and suffixes are deleted for efficiency, while infixes add an extra layer of difficulty.

Light-based stemmer is used in the stemming process. Larkey's mechanism is indeed used in such research. If the rest of the word has three or more letters, the additions at the beginning of the term, such as "" or "and," are removed. If there are two or more characters left, the definite articles must be eliminated. The strings that must be taken are shown in Table 2.1 Both definite articles and conjunctions are meant by the specified prefixes. The light stemmers are not allowed to remove any strings that contain Arabic prefixes.

Table (2.1) Arabic prefixes.

Prefixes	Suffixes
ال، وال، بال، كال، فال، لل، و	ها،ان،ات،ون،ين،هن،هم،تة،تي،ني،يه،يهٔ،ه،ة،ي

The light stemmer isn't a reference book. As a result, it's impossible to find an application that specializes in removing affixes from the remaining Arabic words. The popularity of light stemmers may be ascribed to the fact that their words change regardless of their presence in a word list. The trial's goal is to eliminate strings that could be used as affixes in the future, rather than only having the beginning or end of an Arabic stem with no affixes. [15].

2.2.2Root-Based Stemming [16]

Is the verb root generation depends on Arabic Pattern. arabic stemmer algorithm.

In Figure(2.1), we define sets of diacritical markings and affix classes to begin him description of our stemmer. his define sets of diacritical markings and affix classes. The stemmer is in charge of eradicating these blemishes. (Note that the س is not a diacritical mark; it is only included in set D as an example of a consonant with diacritics.) him must also construct certain pattern sets in Figure(2.3) The algorithm is described in chapter three. Stemming is done in the order listed below:

1. Removing vowel diacritics from the text.

2. Converting the hamza, which appears in multiple forms when combined with different letters, to a single form (c). This step is required to ensure that phrases like توكل (he eats) and اؤكـل (it eats) merge to the same root once their prefixes are removed.

3. Removing the length three and length two prefixes in that sequence.

4. Removing the connector و if it comes before a term that starts with و.

5. Normalizing آ , أ, to ا units. In this scenario, removing the hamza does not affect the root.

6. If the number of stems is fewer than or equal to three, return them. Trying to shrink stems even more leads to confusing stems.

7. Consider the following four scenarios, each based on the length of the word:

a. Length = 4: If the term meets one of the categories, Extract the relevant stem from the Pro_w4 patterns (Table (2.2)) and return it. Otherwise, if the term is longer than three letters, delete length-one suffixes and prefixes from S1 and P1 in that order.

b. 5th Length: For words that match Pro_w53 patterns, extract stems with three characters are used. If no matches are found, delete suffixes and prefixes; otherwise, return the relevant length-three stem. If the word is still five characters long, it is compared to Pro_w54 to see if it has any stems that are four characters long. If a relevant stem is located, it is returned.

c. Length If the word matches a pattern from Pro_w63, extract three-length stems. Otherwise, try removing suffixes; if a suffix, Return the word to step 7b once it has been deleted and a phrase of length five has resulted. Otherwise, try removing one character prefix and, if successful, proceed on to step 7b with the length-five phrase.

d. The length of the string is 7. Suffixes and prefixes with only one character should be removed. Send the length-six term to step 7c if you were successful.

Step 7 basically takes larger words and trims single-character affixes one by one. If successful, it compares the shorter term to numerous patterns at various levels until it either fits a pattern and extracts the necessary term, or it becomes too short to be a valid stem. [16].Arabic definition article in Table (2.2) Arabic pattern.

Table (2.2) Arabic pattern and Root.[16].

set	description	examples
Pro_w4	length four patterns	فاعل فعول فعلة فعال فعيل مفعل
Pro_w53	length five patterns And length three patterns	تفاعل افتعل افعال افاعل فعالة فعلان فعولة تفعلة تفعيل مفعلة مفعول فاعول فواعل مفعال مفعيل افعلة فعانل منفعل مفتعل فاعلة مفاعل فملاع يفتعل تفتعل فعالي انفعل
Pro_w54	llength five patterns and length four patterns	تفعلل افعلل مفعلل فعللة فعلان فعالل
Pro_w63	Length six patterns and length three patterns	استفعل مفعالة افتعال افعوعل انفعل مستفعل

Pro_w64	length six patterns And length four patterns	افتعلل افعلال متفعلل

Table(2.3): Affix Sets [16].

set	description	examples
P3 P2 p1	prefixes of length three Length two prefixes length one prefixes	وللـ، وال، كال، بال ال، لل ل، ب، ف، س، و ي، ت، ن، ا
S3 S2 S1	Length three suffixes Length two suffixes Length one suffixes	تمل، همل، تان، تين، گمل ون، ات، ان، ين تن، گم، هن، نا يا، ها، تم، گن ني، وا، ما، هم ة،ه،ي،گ،ت،ا،ن
D	diacritics-vowelizations	سَ سِ سُ سْ سّ سً سٍ سٌ

Simple nouns and verbs are created by combining these roots with a variety of vowel patterns, to which affixes can be added for more sophisticated derivations [17]. Patterns in Arabic are a feature of the language's grammar. They're built around the Arabic root [18]. In Arabic lexicography and morphology, patterns are very important [17]. They result from the vocalization and affixation processes [19]. Each root can canonically combine with orthographically distinct patterns to form new words; for example, the root"بعل "is made up of three characters, the root " بعل" corresponds to the pattern "فعل " and the pattern preserves "فعل" and "in the *same order, where other letters can be added to form a new pattern. Several patterns, for example, are formed from the morpheme "لعف "'s

foundation pattern " بعل”,". "Form" is a pattern. the letter " بعلم” to the morpheme "م ". the word "بعل " by adding the letter "م" to the morpheme .

2.2.2.1 Khoja's stemmer[20]

Khoja's stemmer removes the longest suffix and prefix. The root is then extracted by comparing the remaining word to verbal and nominal patterns. The stemmer uses a list of all diacritic letters, punctuation characters, definite articles, and 168 stop-words, among other linguistic data files. Although it has several limitations, particularly with nouns, the Khojas method removes suffixes, infixes, and prefixes before extracting the roots using pattern matching. An overview of the Khojas origins. An overview of the Khojas origins. Algorithm (2.3) depicts the technique.

Algorithm 2.3: Khoja Root-Based Stemming Algorithm

Purpose: Stemming Arabic Words

Input:

· Dataset

· Stop-wordlist

· Assets and patterns files

Output: Stemmed Dataset

Procedure:

1. Replace initial آ, أ, إ with ا.

2. Stop-words removal.

3. Remove punctuation, non-letters and diacritics.

4. Remove definite articles from the beginning of the word.

5. Remove the letter (و)from the beginning of the word and (ة) from the end of the word.

6. Remove prefixes and suffixes

7. Comparing the resulting word to patterns stored in the dictionary, if the resulting root is meaningless the original word is returned without changes.

Will be summarized the advantages and disadvantages of the previous methods according to the Table 2.4 .

Table (2.4): advantages and disadvantages of stemming approach [20].

	Advantage	Disadvantage
Light Stemming	Stripping off a small set of prefixes and/or suffixes without dealing with infixes or	No absolute abundant lists of strippable prefixes and/ or suffixes , adjective mainly does not provide conflated specially with its singular

	recognize patterns and finds roots.	form, it fails to conflate broken plurals for nouns, conflate is not given by with the present forms of past tense
Root-Based Stemmers	Shortening the vocabulary space, thus drastically improving the size of the index.	Increases word ambiguity , all possible patterns are not involved
Statistical Stemmers	Not require language expertise, they employ statistical information from a large corpus of a given language to learn morphology of words	Little complex and may over stem the words sometimes.

2.3Feature Extracting

Reducing the number of resources needed to describe a huge quantity of data is the goal of feature extraction. One of the main issues with analyzing complex data is the sheer amount of variables that are involved. A classification algorithm may overfit to training examples and perform poorly on fresh samples when an analysis with a high number of variables involves a lot of memory and processing resources. A way of creating combinations of the variables to get past these issues while still accurately describing the data is known as feature extraction. Many machine learning experts think that the secret to efficient model creation is appropriately tuned feature extraction. Using constructed sets of application-dependent attributes, often created by a specialist, results can

be enhanced. The technique of feature engineering is one of these. As an alternative, broad dimensionality reduction methods like:

1. Analysis of independent components
2. Kernel Isomap PCA
3. Analysis of latent semantics
4. Least squares in part
5. The principal component method
6. Reduced multifactor dimensionality
7. decrease of the dimensions nonlinearly
8. Autoencoder.
9. Semidefinite embedding.

Feature extraction is a process used in machine learning, pattern recognition, and image processing that starts with a set of measured data and creates derived values (features) meant to be informative and non-redundant. This process speeds up the learning and generalization processes and, in some cases, improves human interpretations. Dimensionality reduction and feature extraction are connected. When an algorithm's input data is too extensive to process and is thought to be redundant (such as when the same measurement is given in feet and meters or when pixels are used to represent images), it can be reduced to a smaller collection of characteristics (also named a feature vector). The process of selecting a portion of the first characteristics is known as feature selection. It is anticipated that the chosen characteristics will include the pertinent information from the input data, allowing the required

task to be carried out using this condensed representation rather than the entire starting data. [21].

2.3.1Term Frequency-Inverse Document Frequency (TF-IDF)[22]

The TF-IDF is a numerical statistic that depicts the importance of a word in a collection of documents . In text mining and information retrieval, the TF-IDF is often used as a weighting factor. The number of times a word appears in the text improves the TF-IDF value, while the term's repetition in the corpus counteracts this. This can assist in overcoming the issue that some terms are more familiar than others. The mathematical result of two measures, T.F. and IDF, is TF-IDF. The TF-IDF for each phrase can be calculated using Equation 2.1 [22].

$$\boldsymbol{TD - IDF(t,d) = TF(t,d) \times IDF(t)} \qquad \textbf{(0.1)}$$

The number of times a phrase t appears in a document d is defined as T.F. IDF, on the other hand, denotes the use of a statistical weight to determine the relevance of a term in a collection of documents . Equation 2.2. can be used to calculate this. [22].

IDF (t)= log [(n) / (df(t))] IDF (t)= log [(n) / (df(t)) IDF (t)= log [] (2.2)

df(t) is the document frequency of t, and n is the total number of documents in the document set. The number of documents in the document set that contain the phrase t is known as the document frequency. Machine learning algorithms can employ texts that have been converted using TF-IDF. [11]. This is a technique for quantifying words in a set of documents. In general, each word is scored to show its meaning in the document and corpus. This method is widely used in information retrieval and text mining. [23].

2.4 Feature Selection[24]

also known as dimension selection, is a machine learning approach for selecting a subset of attributes from a data collection in order to construct a stable learning model.such as Feature Selection (Filtering, PCA).

Unsupervised Techniques: These techniques can be used for unlabeled data.

From a taxonomic point of view, these techniques are classified as under:

A. Filter methods

B. Wrapper methods

C. Embedded methods

D. Hybrid methods.

2.4.1 Filtering method

The filtering approach captures the feature's distinctive qualities as determined by univariate statistics, not cross-validation performance. These methods are faster and need less computing power than wrapper methods. When working with high-dimensional data, the filter approach is less computationally expensive. Threshold for variation The variance threshold is a straightforward method for selecting features. Removes all features whose variance falls below a certain level. All zero distribution characteristics are disabled by default. The ability for all samples to have the same value. His think that features with a high variance contain more important information, but his overlook one of the limitations of filtering

methods: feature variables and the link between features and target variables. [24].

2.4.2Principal Component Analysis(PCA)[25]

PCA is a dimensionality-reduction technique for lowering the dimensionality of large data sets by transforming a huge collection of variables into a smaller set that preserves the majority of the data.

The dataset for the PCA technique has to be scaled. The outcome is influenced by the relative scale. It's a way of explaining data to non - technical people.

In technical terms, a main component is a linear combination of ideally weighted observable variables. PCA resulted in these primary components, and the number of original variables is less than or equal to the number of new variables. His can use less if them wish to eliminate or reduce the number of dimensions in our dataset.

Principal Component Analysis, when viewed in its most informative light, can provide the viewer with a lower-dimensional representation, a projection or "shadow" of this item.As in PCA in figure (2.4).

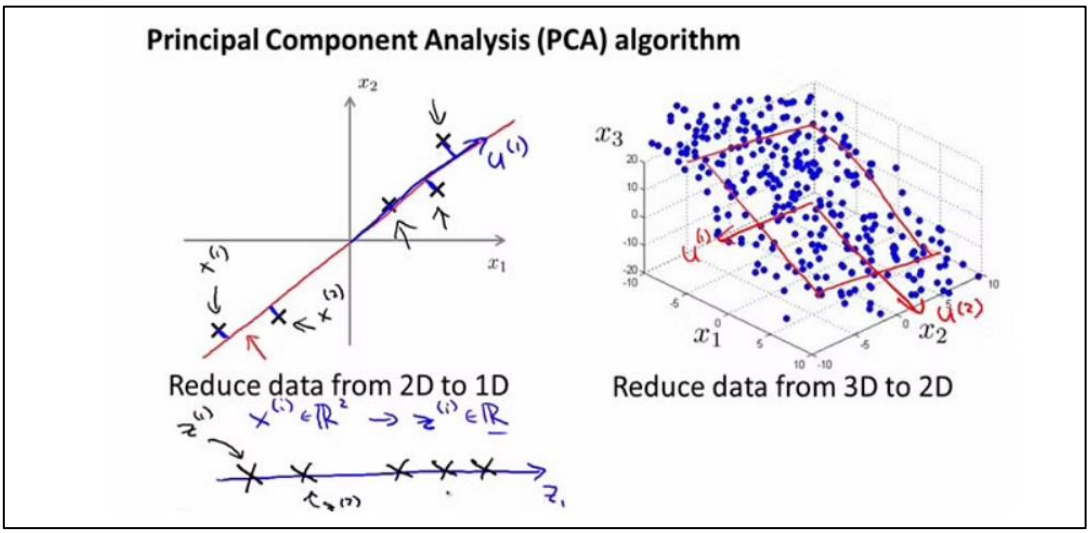


Figure (2.1) : principal component Analysis [25].

2.5Text mining

Text mining is a branch of computer science that combines approaches to text mining such as data mining, information retrieval, machine learning, statistics, natural language processing, and knowledge management. Customer complaints (or feedback)management, marketing focus group programs, competitive intelligence, market forecast, extraction of subject patterns in text streams, identifying semantic relationships between events, and customer satisfaction analysis are just a few examples of typical applications. Text mining entails tasks like text categorization, phrase extraction, document summarization (single or multi-document), clustering, association rules mining, and sentiment analysis.[1]

Machine learning algorithms create a model based on training data to make predictions or judgments without having to

be explicitly programmed to do so. as an example, the classification:

Different Types of Classification Algorithms [26] .

- Stochastic Gradient Descent.
- Logistic Regression.
- Nave Bayes.
- (K-Nearest).
- (SVM)
- Decision Tree.
- Random Forest.

2.5.1 K-Nearest

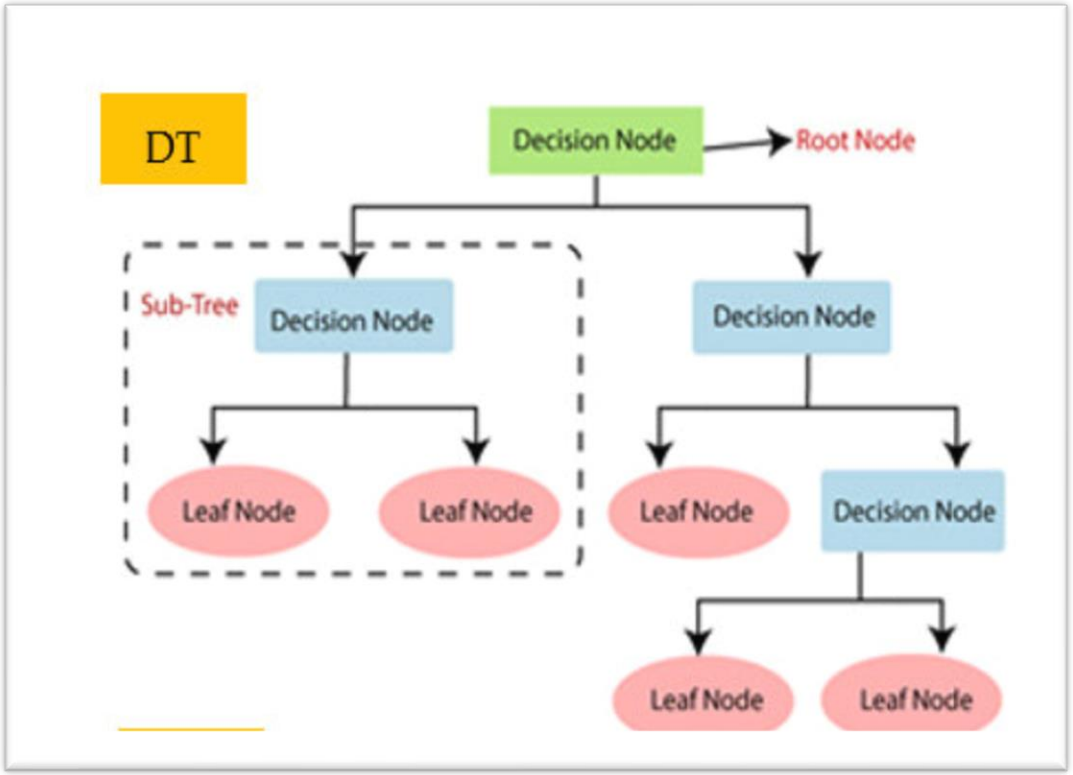
Nearest Neighbor learners are viewed as lazy learners since they put off modeling the training data until after a new document has been identified. The lazy learner known as the "rote classifier" memorizes all of the training data and only performs classification when a test document's characteristics (or attributes) exactly match one of the training documents. The instance-based learning technique, which includes the nearest-neighbor classification technique, essentially uses training documents to make predictions for tested documents without needing a model derived from data. To compare the similarity of the training documents and the classification function, which yields the predicted class of the document under test based on its closeness to previous training documents, instance-based learning techniques need a proximity measure The KNN classifier is chosen to construct the system because it is straightforward, has an acceptable similarity metric, and doesn't require any resources for training, despite certain drawbacks like an above-average classification time because no time was spent in the learning phase. [27].

2.5.2 SUPPORT VECTOR MACHINE

One of the most popular and well-known machine learning methods is called a support vector machine (SVM). SVMs are a binary classifier that Vladimir first suggested. SVMs are supervised learning models that are further used to analyze and classify text data. SVM is also frequently used for regression analysis and text categorization. Thorsten Joachims introduced and applied SVM in text classification and categorization in 1998. The basic purpose of the SVM training method is to create a model that classifies fresh documents into a number of predetermined categories. SVMs can also be employed as non-linear and linear classifiers.[28].

2.5.3 Decision Tree

A typical Decision Tree technique (DT) is useful for creating models with a tree topology. A dataset is divided into smaller and smaller portions using the Decision Tree method, as in DTin figure(2.5) .[11]



Figure(2.2): The Main Idea of Decision Tree [29].

. A comparable decision tree is being created progressively at the same time. This method creates a tree with nodes for decision and leaf. Both numerical and categorical data can be handled by decision trees. The main idea behind the decision tree algorithm is depicted in Figure (2.5). depicts the decision tree. [11].

2.5.2Random Forest Classifier

Random Forest Classifier (RF) is a classification algorithm that uses a random forest The selected tree has evolved into the irregular woodland calculation. Simply start by constructing a large number of selected trees with previous data, then fit his current data into one of the trees as a "random forest.". Irregular timberland models are accommodating as them

cure the choice tree's issue of "forcing" information to focus inside a category pointlessly. [30]. It can be seen in the figure (2.6).

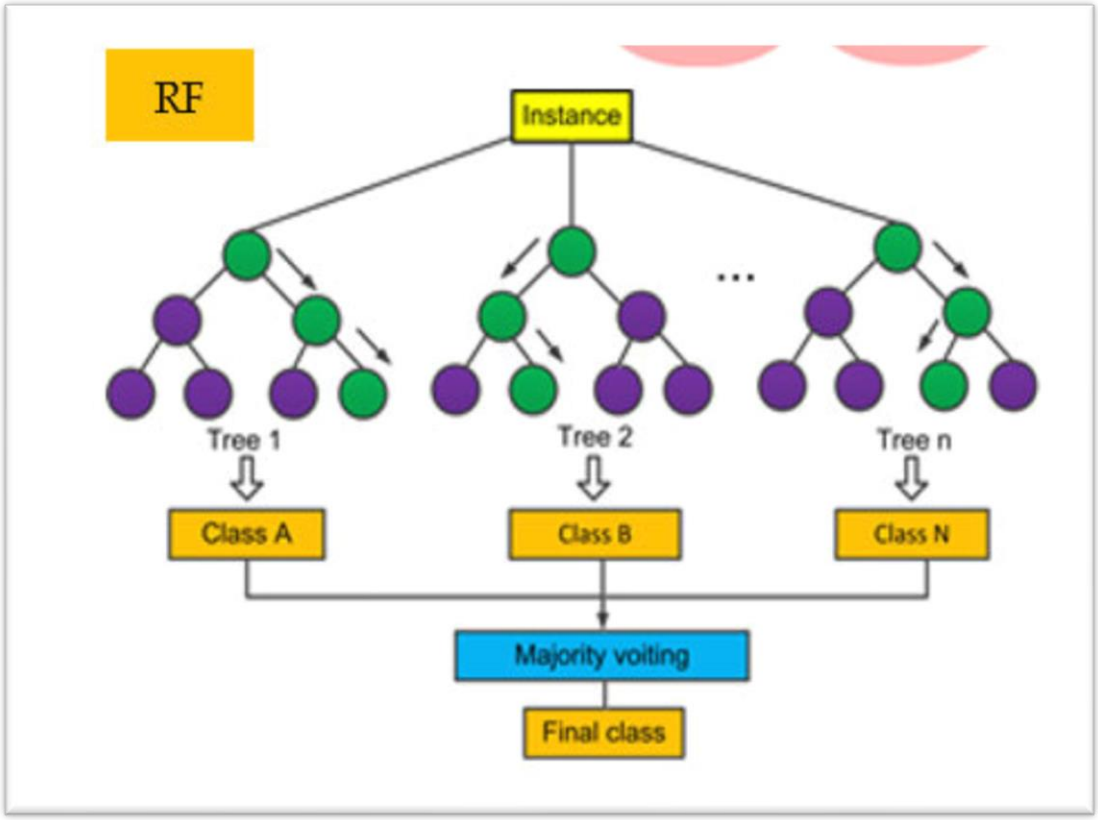


Figure (2.3): The Main Notion of Random Forest [29]

2.6Performance Metric for Classification Algorithms

Various measures can be used to assess the effectiveness of different classification methods. Accuracy, f1-score, precision, and recall are all factors. As shown in Table 2.4, these measurements are calculated using the computation confusion matrix, which is a matrix that summarizes the number of samples correctly or incorrectly predicted by a classification model. The table's most important values are listed below. [22]:

Table (0.4): Confusion Matrix [8]

		Predicted Class	
		Positive **+**	Negative **-**
Actual Class	Positive +	*f*++ (TP)	*f*+- (FN)
	Negative -	*f*-+ (FP)	*f*- - (TN)

1. True positive (T.P.): denotes the positive examples that are Sample classified.

2. False-negative (F.N.): denotes the positive examples that are incorrectly classified.

3. False-positive (F.P.): denotes the negative examples that are incorrectly predicted and classified.

4. True negative (T.N.): denotes the negative instances that the classification model adequately predicts.

The measures of accuracy, recall, precision, and f1-score are discussed:[22].

1. Accuracy is the number of correct predictions divided by the total number of predictions. The accuracy can be computed based on Equation 2.4.

$$\textbf{Accuracy} = \frac{\mathbf{TP + TN}}{\mathbf{TP+TN+FP + FN}} \qquad \textbf{(0.4)}$$

2. Precision is calculated by dividing the number of T.P. by the total number of T.P. and F.P. Equation 2.5 can be used to calculate the precision.

$$\textbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}} \qquad \textbf{(0.5)}$$

3. The recall is calculated by dividing the total number of T.P. by the total number of T.P. and F.N. Equation 2.6 can be used to calculate this measure.

$$\textbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \qquad \textbf{(0.6)}$$

4. The F1-score is calculated as 2*((precision*recall)/ (precision + recall). It's also known as the f1-score or f1-measure. Equation 2.7 can be used to calculate an equation for this measure.

$$\textbf{F1-score} = \frac{(\mathbf{2}*\mathbf{TP})}{(\mathbf{2}*\mathbf{TP+FN+FP})} \qquad \textbf{(0.7)}$$

CHAPTER THREE

PROPOSED SYSTEM

PROPOSED SYSTEM

3.1 Introducation

The Arabic text can be dealt with using many processing operations to obtain the desired purpose. Thus we guarantee to obtain an editable text through which we obtain information of better value than the previous one and can be used in future studies. Select One of the Datasets Akhbarona, Al-Arabiya, One of the Filtering thresholds 0, 2, 4,And one of the Stemmers Light, Complex,Apply PCA 80% Reduction or Ignore PCA,Apply Random Forest with 100 Trees,Calculate Confusion Matrix Calculate Precision, Recall and F-Measure,

The Goal: The Goal is to design an Arabic classifier while comparing the effect of traditional Arabic stemmers using basic and light rules against the more complicated grammar-based stemmers. To show the difference, a complete classification system is required.

3.2 System design

The proposed testing system is devised in a layer form; each layer can be substituted or replaced as required. Look at figure (1). The system layers are as follows:

1st Layer: Defining a dataset, the (SANAD Arabiya and SANAD Akhbarona) .

2nd Layer: Preprocessing and tokenization

3rd Layer: Stemming, two types of stemmers are used and compared (the complex and lite stemmers).

4th Layer and 5th Layer: Feature Extracting and feature reduction are applied; the document vector is created and filtered using a threshold.

6th Layer and 7th Layer: Feature Selection followed by feature reduction is applied, the TFIDDF matrix is created and then reduced using the PCA algorithm.

8th layer: the classification process using the Random Forest algorithm.

9th layer: this is the final layer where evaluation is implemented, the precision, recall, F1 measure and the confusion matrix are calculated.

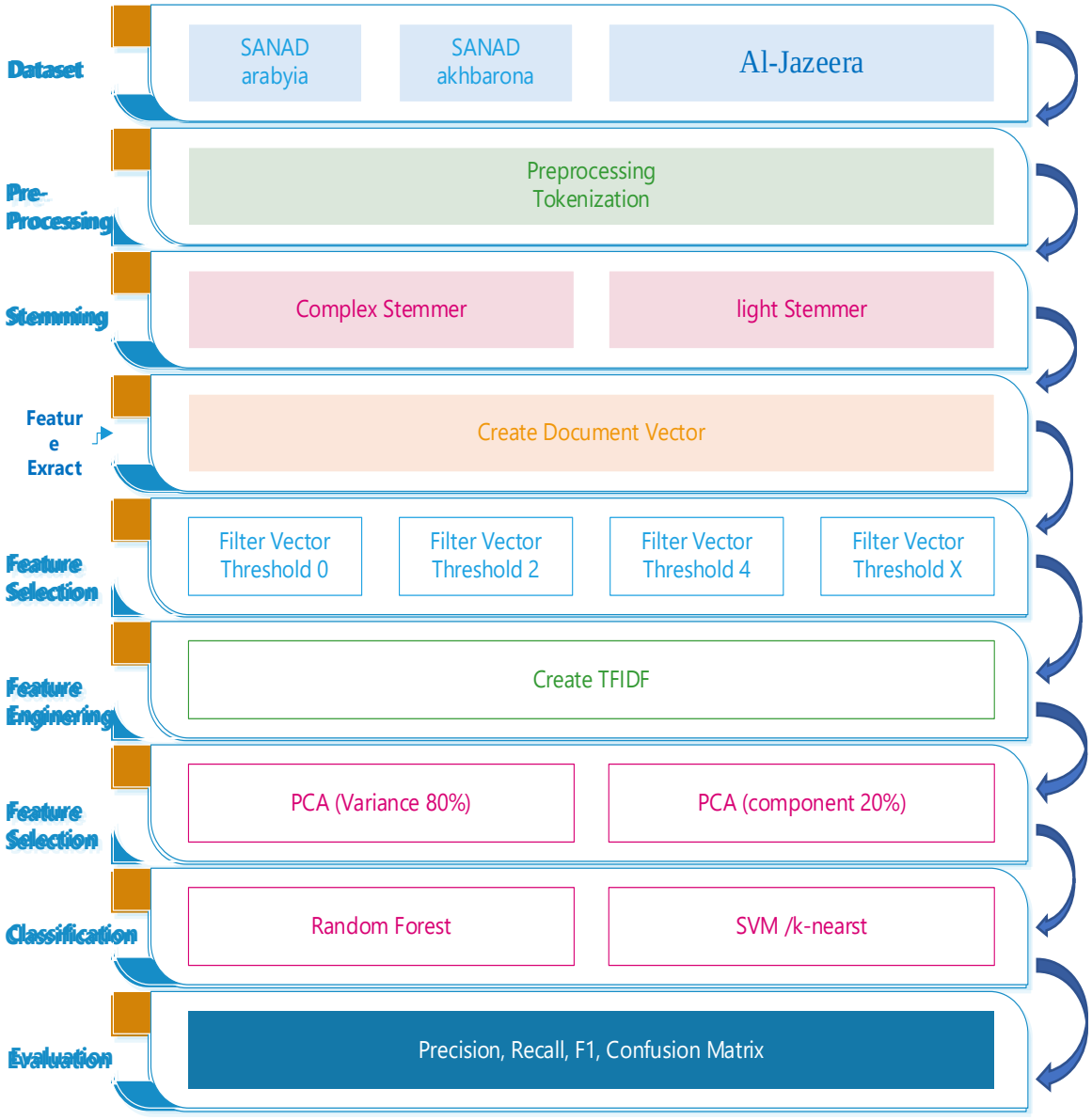


Figure (3.1) Showen the layers in the proposal system.

3.3Datasets

SANAD Dataset is an extensive collection of Arabic news articles used in different Arabic NLP tasks such as Text Classification and Word Embedding. Akhbarona has seven categories [Culture, Finance, Medical, Politics, Religion, Sports and Tech] while Al-Arabiya is missing the Religion category.

3.4The Preprocessing

Pre-processing is one of the essential processes in the treatment of Arabic text.

1. **Parsing dataset:** files are loaded to memory, then the dataset is separated into documents, documents are processed one by one, the title, body and tag are extracted and returned in a form usable by the following layers. As in Algorithm (3.1).

Algorithm (3.1): Parse SANAD Dataset		
Input SANAD Path		
Output ListOfDocuments		
1	ListOfFiles = Read Files from Disk to Memory (SANAD Path)	
2	ListOfDocuments = Empty List	
3	For Each File in ListOfFiles Do	
4		Extract Title, Body, Tag from File
5		Document = (Title, Body, Tag)
6		Add Document to ListOfDocuments
7	Next File	
8	Return Updated ListOfDocuments	

2. **Tokenization:** After separating dataset documents to title, body and tag, the body is processed, the text is split into tokens using the delimiters, then each token is passed to a set of rules, only then passed to the stemmer of choice. Each document has its stemmed tokens saved in a list called the ListOfTokens. As in Algorithm (3.2).

Algorithm (3.2): Dataset Preprocessing					
Input ListOfDocuments, TrimParameters, Delimiters, StopWords, StemmerType					
Output ListOfDocuments with tokens					
1	Foreach Document in ListOfDocuments Do				
2		UntokenizedWords = Split using Delimiters (Document. Body)			
3		ListOfTokens = Empty List			
4		Foreach token in UntokenizedWords Do			
5			If token length = 0, Go to the Next token		
6			token = Trim Both Ends using TrimParameters (token)		
7			If token length = 1, Go to the Next token		
8			If the rest of the word contains a non-Arabic letter, Go to the Next token		
9			Normalize Token		
10			If Stemmer Type is Light:		
11				Apply Light Stemming on (token)	
12			Else If Stemmer Type is Complex:		
13				Apply Complex Stemming on (token)	
14			End If		
15			If Token is in StopWords, Go to Next Token		
16			Add token to ListOfTokens		
17		Next token			
18		Add ListOfTokens to Document			
19	Next Document				
20	Return Update ListOfDocuments				

3. **Normalization** in step 9 in algorithm two is defined as:
   Normalization of ء with alef seat to bare alef.
   Normalization of ة to heh.
   Normalization of dotless yeh (alef maksura) to yeh.
   Removal of Arabic diacritics (the harakat).
   Removal of ت (stretching character).

3.5 The Stemming layer

Stemming is performed as part of algorithm 2 in step 10. Two types of stemmers are used in this thesis, where only one is chosen at each runtime; the lite stemmer is inspired from the work of and the complex stemmer is inspired from the work of Taghva K, Elkhoury R and Coombs in the 2005 paper titled “Arabic stemming without a root dictionary” [16].As in Algorithm(3.3).

Algorithm (3.3): Complex Stemmer.		
Input Dataset After Token		
Output Stemming A Word Token Using The Complex Stemmer		
1	Token	
2		1: Remove Diacritics Which Representing Arabic Short Vowels 2: Remove Length Three And Length Two Prefixes In This Order 3: Remove Length Three And Length Two Suffixes In This Order 4: Remove Connective ‘و’ If It Precedes A Word Beginning With ‘و’ 5: Normalize Initial Hamza To Bare Alif // If 4 <= Word Length <= 7, Then Stem; Otherwise, No Stemming1:
3		Helper Functions • word-4(W): 1. For first p ∈ Pro-W4 such that p ⊆ W return extract(W). 2. Do short-suffix(W). 3. If len(W) = 4 do short-prefix(W). • word-5(W) 1. For first p ∈ Pro-W53 such that p ⊆ W return extract(W). 2. Do short-suffix(W). 3. If len(W) = 4 do word-4(W) 4. else (a) do short-prefix(W) (b) if len(W) = 4 do word-4(W) (c) otherwise, if len(W) = 5 , for first p ∈Pro-W54 s.t. p ⊆ W return extract(W). • word-6(W) 1. For first p ∈ Pro-W63 s.t. p ⊆ W, return extract(W). 2. Do short-suffix(W). 3. If len(W) = 5 then do word-5(W) 4. otherwise (a) do short-prefix(W). (b) if len(W) = 5 do word-5(W). (c) otherwise, if len(W) = 6 for any p ∈ Pro-W64s.t. p ⊆ W return extract(W). • short-suffix(W) 1. For first s ∈ S1 s.t. s ⊆ W, W = W − s. 2. Return W. • short-prefix(W)

		1. For first p ∈ P1 s.t. p ⊆ W, W = W − p. 2. Return W.
4		Main algorithm For every word W 1. Remove all d ∈ D from W 2. Normalize ئ,ء, ؤ to ا 3. If len(W) ≥ 6, for first p ∈ P3 s.t. p ⊆ W, W =W−p, else if len(W) ≥ 5, for first p ∈ P2 s.t. p ⊆ W,W = W − p. 4. If len(W) ≥ 6, for first s ∈ S3 s.t. s ⊆ W, W =W − S3. else if len(W) ≥ 5, for first s ∈ S2 s.t.s ⊆ W, W = W − s. 5. If len(W) ≥ 4 and initial characters of W are ××, remove initial ×. 6. Normalize initial charaacter آ،أ to ا if necessary. 7. if len(W) ≤ 3 return W. 8. (a) if len(W) = 4 then do word-4(W) (b) else if len(W) = 5 do word-5(W) (c) else if len(W) = 6 do word-6(W) (d) else if len(W) = 7 i. do short-suffix(W). ii. if len(W) = 6 do word-6(W) else A. Do short-prefix(W). B. If len(W) = 6 do word-6(W). 9. Return W

3.6 Weight generation layers and feature reduction layers:

Feature Extract is divided into four layers. Starting at the 4th layer and ending in the 7th layer.

1. **4th layer** in the proposal system is for creating document vectors. Each document is represented by a vector that contains the statical distribution of the terms used in the document in relation to the whole dataset. The process starts with finding the frequency of every term in the document in relation to its total tokens. As in Algorithm (3.4).

Algorithm (3.4): Finding Term Frequency T.F	
Input ListOfDocuments	
Output ListOfDocuments with term frequencies	
1	Foreach Document in ListOfDocuments Do
2	ListOfTermFrequencies = Empty List
3	Foreach token in Document. ListOfTokens Do
4	if the token is already in the ListOfTermFrequencies
5	Increase its count by 1
6	Update the token frequency
7	Else
8	Add a new item to the ListOfTermFrequencies with a count of 1
9	Update the token frequency
10	End If
11	Next token
12	Add ListOfTermFrequencies to Document
13	Next Document
14	Return Updated ListOfDocuments

The vector shape is created by finding every term used throughout the dataset and sorting them alphabetically; this vector is called the term dictionary. As in Algorithm (3.5).

Algorithm (3.5): Compose Total Term Dictionary	
Input ListOfDocuments	
Output TotalTermDictionary	
1	TotalTermDictionary = Empty Dictionary
1	Foreach Document in ListOfDocuments Do
2	Foreach TermFrequency in Document. ListOfTermFrequencies Do
4	if the term is found in TotalTermDictionary:
5	Update the total count of this term
7	Else
8	Add a copy of the term to the dictionary
10	End If
13	Next TermFrequency
15	Next Document
16	Return TotalTermDictionary

2. **In the 5th layer** in the proposal system, after creating the TotalTermDictionary, the vector can be reduced by removing terms with a total number of occurrences below the termIgnoreThreshold.

Algorithm (3.6): Filter Total Term Dictionary			
Input TotalTermDictionary, termIgnoreThreshold			
Output Filtered TotalTermDictionary			
1	Filtered TotalTermDictionary = Empty Dictionary		
1	Foreach Term in TotalTermDictionary Do		
4		if the term count > termIgnoreThreshold:	
5			Add term to Filtered TotalTermDictionary
10		End If	
15	Next Term		
16	Return Filtered TotalTermDictionary		

3. **In the 6th layer** of the system, the Embedding matrix is created using the TFIDF algorithm; the TFIDF is explained in chapter 2. Parallel processing is used to speed up the Matrix creation; each row is processed in a different thread, using as many threads as available. Also, some files are saved to the disk for debugging and diagnostics purposes if needed; the saved output includes:
    a. The documents extracted from the dataset in the 1st layer are saved each as a separate file.
    b. After applying the pre-processing and stemming in the 2nd and 3rd layers, the documents were found.
    c. The document term frequencies are calculated in the 4th layer.
    d. And finally, the Filtered TotalTermDictionary with the calculated counts and frequencies of all the terms available in the dataset.

Algorithm (3.7): Apply TFIDF		
Input ListOfDocuments, Filtered TotalTermDictionary, Debug flag		
Output Embedding Matrix		
1	Check available threads in the System	
2	Embedding matrix = Apply TFIDF Algorithm Using available threads on (ListOfDocuments, Filtered TotalTermDictionary)	
3	If Debug flag = True:	
3		Save Original Documents as separated files on disk
4		Save TotalTermDictionary to Disk
5		Save Pre-processed Documents to Disk
6		Save Extracted Document Frequencies to Disk
6	End If	
7	Save Embedding Matrix to Disk	
8	Return Embedding Matrix	

4. **In the 7th layer**, the PCA algorithm reduces the TFIDF matrix's size further. This layer can be skipped entirely if no further compression is required. The PCA algorithm is explained in detail in chapter 2. Two implementations are available. The first uses a preset number of components for the PCA to use, and the PCA keeps the topmost components having the highest variant and discards the rest. In contrast, the sound implementation uses a preset retained variant. The PCA chooses the topmost component that gives the preset retained variant value and discards the rest when their variant is combined.

Algorithm (3.8): Apply PCA		
Input Embedding Matrix, Use Variant Flag, RetainedVariant, NoOfComponents		
Output New Embedding Matrix		
1	New Embedding Matrix = Empty Matrix	
2	If Use Variant flag = True:	
3		Set PCA to use RetainedVariant Mode
4		Fit PCA on Embedding Matrix using RetainedVariant
5		Apply PCA Transformation on Embeddings Matrix
6		New Embedding Matrix = Principal Components From PCA
7		Show the number of components retained
8	Else	
9		Set PCA to use NoOfComponents Mode
10		Fit PCA on Embedding Matrix using NoOfComponents
11		Apply PCA Transformation on Embeddings Matrix
12		New Embedding Matrix = Principal Components From PCA
13		Show the variant retained
14	End If	
15	Return New Embedding Matrix	

3.7 Classification layer

Any classification algorithm can be used in this step, But the random forest algorithm is chosen for this layer; algorithm details are in chapter 2. The feature used in this step is the Embedding matrix extracted from the previous steps with or without applying PCA to it. The labels are the categories predefined for every document in the dataset used. The random forest is initialized with a predefined number of trees. It is trained using a portion of the dataset called the training set. Both the features of the documents and their labels are used for the training. The resulting classifier

is used to predict the labels of the other portion of the dataset, called the testing set, and the classifier is saved to disk to be used later if required.

Algorithm (3.9): Apply Random Forest	
Input Features of Training Data, Features Matrix of Testing Data, Labels of Training Data, NoOfTrees	
Output Predicted Labels, Forest Classifier Object	
1	Initialize the Random Forest Classifier using (NoOfTrees)
2	Train forest on (Features of Training Data, Labels of Training Data)
3	Use the Forest to Predict the labels of (Features of Testing Data)
4	Save the Forest classifier object to disk.
5	Save the Predicted Labels to Disk.
6	Return Predicted Labels, Forest Classifier Object

3.8 Evaluation Layer

In this final layer, the model accuracy is measured by comparing the resulting predicter labels for the test set against the original predefined labels of the same set. This is done using the precision, recall, and f1-score algorithm, generating a confusion matrix.

The System preference is measured by changing the algorithms and their parameters in every layer. It is possible to get the best possible combination that produces the most accurate results while also using less memory and computation power.

CHAPTER FOUR

RESULTS AND DISCUSSION

RESULTS AND DISCUSSION

4.1 System

All testing is done using the system described in table (4.1). Pre-processing and stemming are all done in java, while the random forest and PCA are performed in python.

Table (4.1) Testing System Specifications

O.S. Version	Microsoft Windows 10 21H1
CPU	AMD Ryzen 9 3900X
RAM	64 GB DDR4
Storage	NVME 512 GB
Python Version	Python 3.8.11
Scikit-learn Library	Version 0.24.2
Python IDE	JupyterLab 3.2.9
Java JRE Version	8u291 x64bit
Java JDK Version	16.0.2 x64bit
Java IDE	Apache NetBeans 12.3 Windows x64bit

4.2 Dataset:

The dataset used in this thesis is the SANAD dataset referenced in chapter 3. This dataset IS A collection of documents chosen from Akhbarona and Al-Arabiya.

Akhbarona contains 78050 Arabic news articles organized into seven categories [Culture, Finance, Medical, Politics, Religion, Sports and Tech] from www.Akhbarona.com; the data is split into 42209 articles in the training set and 4690 articles in the testing set.

Al-Arabiya contains 71247 Arabic news articles organized into six categories [missing the Religion category] from www.alarabiya.net. The data is split into 16650 articles in the training set and 1850 in the testing set.

4.3 The light stemming test

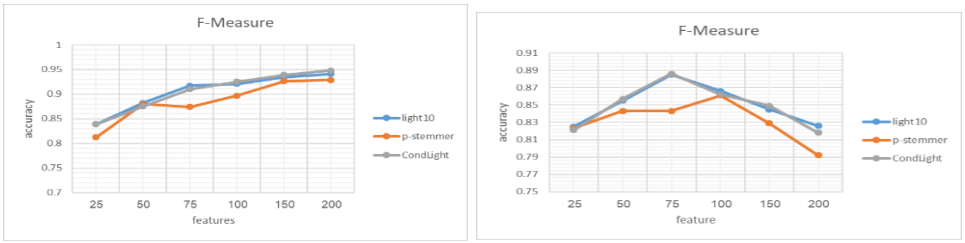
First experiment is light stem. Three algorithm uses light10, P-stemmer, and conditional stemmer. The Al-Jazeera-news dataset is used in all three stemmer tests. They use statistical analysis to compare it in two ways: the accuracy of each stemmer and the degree of similarity among us. The efficiency of each stemmer was calculated using F-measure. NB, SVM, and KNN are the three classifiers used. Each attribute's weight is generated using information gain, then sorted in descending order and given high weight in each experiment.

The result of the NB classifier is demonstrates that the difference between stemmers is not in their appearance but in the number of traits they have. All stemmers produced the same result, however, the difference between them did not surpass 0.2. the light 10 and CondLight stemmers are better than P-stemmer.

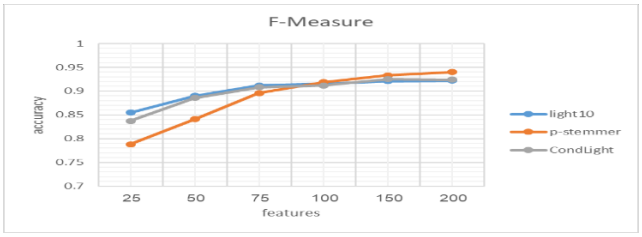
The result of the KNN classifier is demonstrates that the CondLight stemmer is more stable than the others. Changes in attribute values have an impact on both light10 and P-stemmer.

The result of the SVM classifier is demonstrates that the P-stemmer has the highest accuracy when rising features. With minimal qualities, both light10 and CondLight stemmers reach a stable state.

SVM outperforms NB and KNN in terms of accuracy. In Figure(3.1), the result of the F-Measure shows that light10 and CondLight stemmer have the same behavior when rising of attributes number. They are identical when the number of features is increased or decreased. The P-stemmer improves more than other classifiers when the number of attributes increases with SVM. KNN has strange behavior with attribute change.



(a)-NB (b)- KNN



(c)-SVM

Figure(4.1): Result Of F-Measure (A)-NB, (B) - KNN (C) - SVM.

Table (4.2): Jaccard Similarity Between Stemmer

	light10	p-stemmer	CondLight
light10	1	0.3475	0.7904
p-stemmer	0.3475	1	0.3422
candlelight	0.7904	0.3422	1

The second aspect is the similarity and difference among different stemmer methods. The Jaccard is used measurement. The weight of each attribute is computed and arranged in descending order. Selected the best 200 attributes and compute Jaccard similarity among stemmer methods.

The result is shown in Table(4.2). Both Light10 and CondLight have more in common than the P-stemmer.

4.4 Testing Steps:

The testing steps are performed as the following

Select One of the Datasets Akhbarona, Al-Arabiya

Select One of the Stemmers Light, Complex

Select One of the Filtering thresholds 0, 2, 4
(for the approximation of the repetition ratios between light and complex derivation).

Apply PCA 80% Reduction or Ignore PCA

Apply Random Forest with 100 Trees sample.

Calculate Confusion Matrix

Calculate Precision, Recall and F-Measure

In total, 24 tests are performed. As a table (4.3).

Table (4.3): Tests to perform

Test Number	Dataset	Filtering Threshold	Stemmer	Apply PCA
1	Akhbarona	0	Light	None
2				80%
3			Complex	None
4				80%
5		2	Light	None
6				80%
7			Complex	None
8				80%
9		4	Light	None
10				80%

11			Complex	None
12				80%
13	Al-Arabiya	0	Light	None
14				80%
15			Complex	None
16				80%
17		2	Light	None
18				80%
19			Complex	None
20				80%
21		4	Light	None
22				80%
23			Complex	None
24				80%

Sense Dataset and Pre-processing layers (layer 1 and 2) steps are the same across all the tests. The results are collected and shown in table (4.4) without duplication, the time needed to read the files from disk to memory, the number of documents in reach subset of the datasets, the total size of files represented in bytes, the total untokenized and tokenized words in the whole subset.

Table(4.4): Preprocessing Results

Dataset	Subset	Number Of Documents	Reading Time in ms	Total Size of Files in Bytes	Total Untokenized Words	Total Tokens
Akhbarona	Training	42,209	51,730	15,290,554	13,312,496	7,883,967
Akhbarona	Testing	4,690	11,660	1,757,536	1,505,181	892,419
AlArabiya	Training	16,650	44,050	8,596,767	4,156,673	2,604,839
AlArabiya	Testing	1,850	1,007	935,188	451,883	283,649

4.5 **Stemmer Light and Complex**

After applying the stemming layer (layer 3), the selected stemmer affects the pre-processing time and the remaining terms in the total term

dictionary. The results are collected and shown in table(4.5). Note that the total term dictionary is created only once using the training subset and used when creating feature vectors for training and testing sets.

Table (4.5) : Stemmer results				
Dataset	Chosen Stemmer	Pre-processing Time in ms	Creating Term Dictionary Time in ms	Total Terms
Akhbarona / Train	Light	7,021	1,133	152,749
Akhbarona / Test	Light	869		
Akhbarona / Train	Complex	11,070	875	47,814
Akhbarona / Test	Complex	1,258		
AlArabiya / Train	Light	2,635	472	65,811
AlArabiya / Test	Light	366		
AlArabiya / Train	Complex	3,774	349	23,213
AlArabiya / Test	Complex	408		

4.6 **Feature Extracting And Selection**

The stemming layer is flowed by the feature Extracting and selection layers (Layers 4, 5 and 6), where filtering thresholds are chosen, and their effect on the remainder of terms in the total term dictionary after filtering, the size of the feature vector and the TFIDF embedding matrix are all measured. The results from Akhbarona are collected and shown in table (4.4), while AlArabiya results are in table (4.6).R is row and C is column.

Table (4.6): Feature Extracting Results					
Chosen parameters					
Dataset	Threshold	Stemmer	Filtered Terms / Feature Vector	Training TFIDF Size (R x C)	Testing TFIDF Size (R x C)
Akhbarona	0	Light	152,749	42209 * 152749	4690 * 152749
		Complex	47,814	42209 * 47814	4690 * 47814

Akhbarona	2	Light	68,522	42209 * 68522	4690 * 68522
		Complex	20,000	42209 * 20000	4690 * 20000
Akhbarona	4	Light	50,776	42209 * 50776	4690 * 50776
		Complex	15,628	42209 * 15628	4690 * 15628
AlArabiya	0	Light	65,811	16650 * 65811	1850 * 65811
		Complex	23,213	16650 * 23213	1850 * 23213
AlArabiya	2	Light	30,934	16650 * 30934	1850 * 30934
		Complex	11,788	16650 * 11788	1850 * 11788
AlArabiya	4	Light	23,248	16650 * 23248	1850 * 23248
		Complex	9,487	16650 * 9487	1850487

4.7 PCA reduction

After the TFIDF matrix (Embedding matrix) is created, the PCA reduction (Layer 7) is performed by setting the variant percentage to 80%. The results are collected and shown in table (4.7). It is important to note that PCA failed in some of the tests because the matrix is too big to be fitted into memory and processed for the reduction algorithm.

Table (4.7): PCA Reduction Results

Dataset	Number Of Docs	Components		Embedding Size in M.B.	
		Before	After	Before	After
Akhbarona \ Th 0 \ Light \ Train	42209	152,749	Failed	25,790	-
Akhbarona \ Th 0 \ Light \Test	4690			2,866	
Akhbarona \ Th 0 \ Complex \Train	42209	47,814	Failed	8,073	-
Akhbarona \ Th 0 \ Complex \Test	4690			897	
Akhbarona \ Th 2 \ Light \ Train	42209	68,522	Failed	11,569	-

Akhbarona \ Th 2 \ Light \Test	4690			1,285	
Akhbarona \ Th 2 \ Complex \Train	42209	20,000	2,908	3,377	491
Akhbarona \ Th 2 \ Complex \Test	4690			375	55
Akhbarona \ Th 4 \ Light \ Train	42209	50,776	Failed	8,573	-
Akhbarona \ Th 4 \ Light \Test	4690			953	
Akhbarona \ Th 4 \ Complex \Train	42209	15,628	2,787	2,639	471
Akhbarona \ Th 4 \ Complex \Test	4690			293	52
AlArabiya \ Th 0 \ Light \ Train	16650	65,811	4,133	4,383	275
AlArabiya \ Th 0 \ Light \Test	1850			487	31
AlArabiya \ Th 0 \ Complex \Train	16650	23,213	2,165	1,546	144
AlArabiya \ Th 0 \ Complex \Test	1850			172	16
AlArabiya \ Th 2 \ Light \ Train	16650	30,934	3,758	2,060	250
AlArabiya \ Th 2 \ Light \Test	1850			229	28
AlArabiya \ Th 2 \ Complex \Train	16650	11,788	2,020	785	135
AlArabiya \ Th 2 \ Complex \Test	1850			87	15
AlArabiya \ Th 4 \ Light \ Train	16650	23,248	3,510	1,548	234
AlArabiya \ Th 4 \ Light \Test	1850			172	26
AlArabiya \ Th 4 \ Complex \Train	16650	9,487	1,932	632	129
AlArabiya \ Th 4 \ Complex \Test	1850			70	14

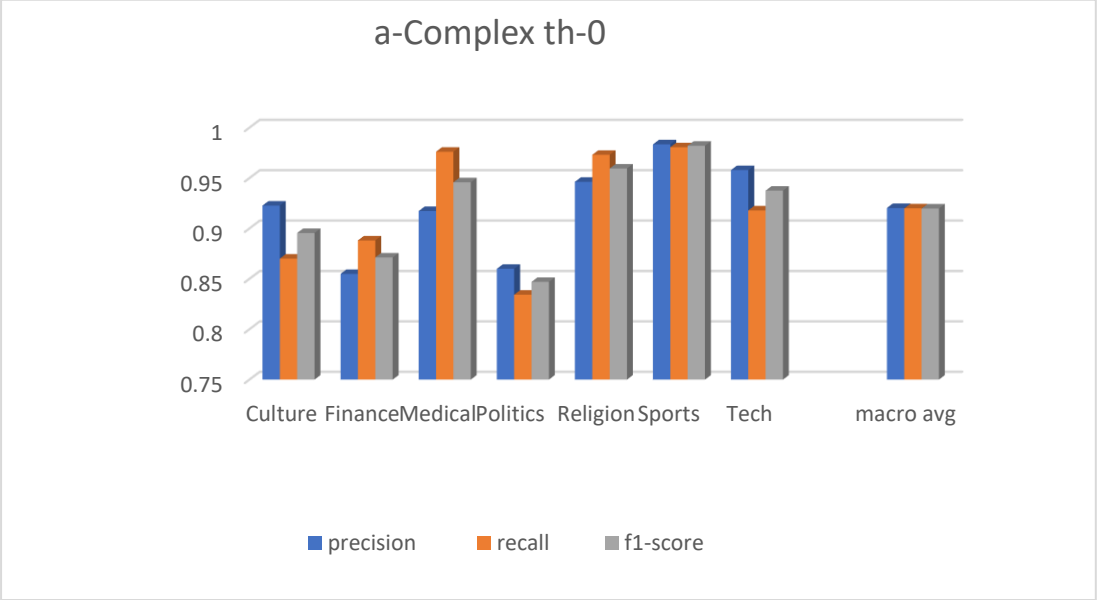
4.8 **Classification and Evaluation**

The random forest classification (layer 8) is applied using 100 trees, and the process is accelerated using multi-threading provided in the Scikit-learn Library. Then the evaluation process (layer 9) is applied, and this is done to all the data of every test. The results are in table (4.8).

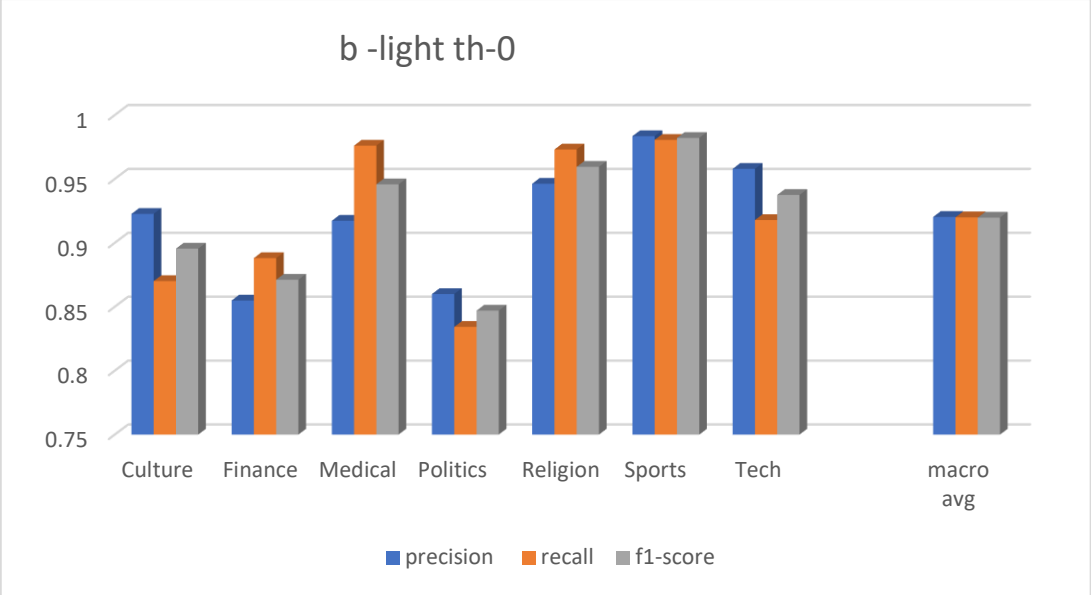
Table (4.8): Evaluation							
Test Number	Dataset	Stemmer Type	Filter Threshold	Reduction Method	Precision	Recall	F1
1	Akhbarona	Complex	0	No	0.9144	0.9145	0.9141
2				PCA (Variance 80%)			
3			2	No	0.9200	0.9198	0.9197
4				PCA (Variance 80%)	0.9117	0.9113	0.9113
5			4	No	0.9161	0.9156	0.9154
6				PCA (Variance 80%)	0.9101	0.9094	0.9095
7		light	0	No	0.9203	0.9200	0.9198
8				PCA (Variance 80%)			
9			2	No	0.9250	0.9247	0.9245
10				PCA (Variance 80%)			
11			4	No	0.9243	0.9243	0.9240
12				PCA (Variance 80%)			
13	AlArabiya	Complex	0	No	0.9624	0.9622	0.9621
14				PCA (Variance 80%)	0.9555	0.9551	0.9552
15			2	No	0.9613	0.9611	0.9611
16				PCA (Variance 80%)	0.9535	0.9530	0.9531
17			4	No	0.9652	0.9649	0.9648
18				PCA (Variance 80%)	0.9576	0.9573	0.9574

19		light	0	No	0.9674	0.9670	0.9670
20				PCA (Variance 80%)	0.9603	0.9600	0.9601
21			2	No	0.9662	0.9659	0.9659
22				PCA (Variance 80%)	0.9620	0.9616	0.9617
23			4	No	0.9662	0.9659	0.9660
24				PCA (Variance 80%)	0.9642	0.9638	0.9639

The results of the proposed system applies with different threshod.In the First Uses Dataset akhbarona , threshold equal to zero. The result are shown in figure (4.2) and figure (4,3) without PCA .

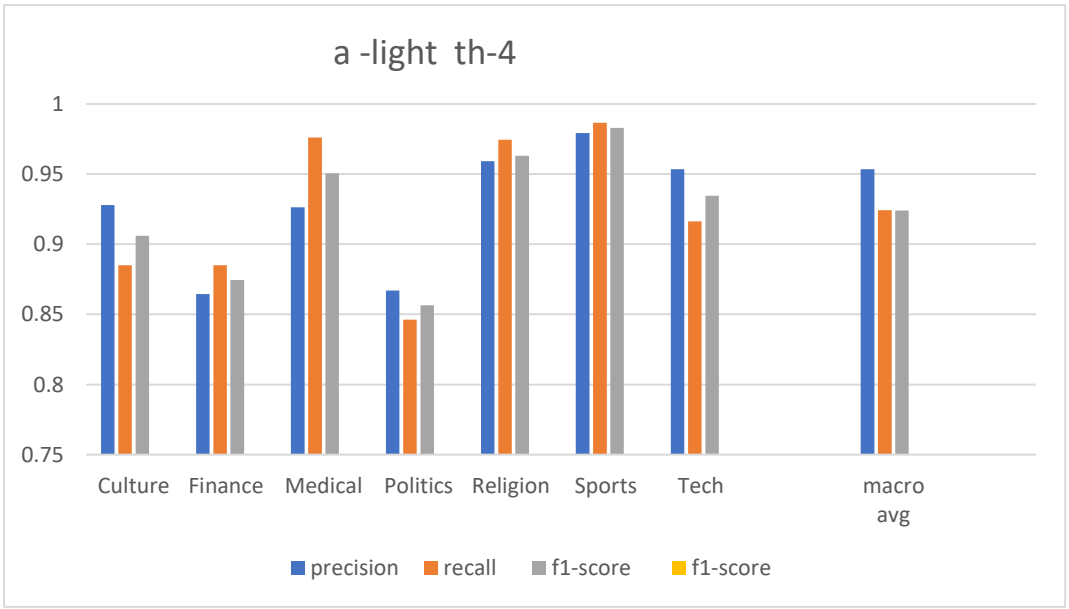


Figure(4.2): Result Of Complex Stemming Uses Dataset akhbarona, Th_0, PCA=None, And Number Of Trees=100,



figure(4.3): akhbarona,th_0,b -light, USE PCA=None, Number Of Trees=100,

The results of proposed system applies with different threshod. First, threshold equal 4. The result are shown in figure (4.4) without PCA figure (4.5) with PCAfor both complex and light10 stemmers. Both are same.



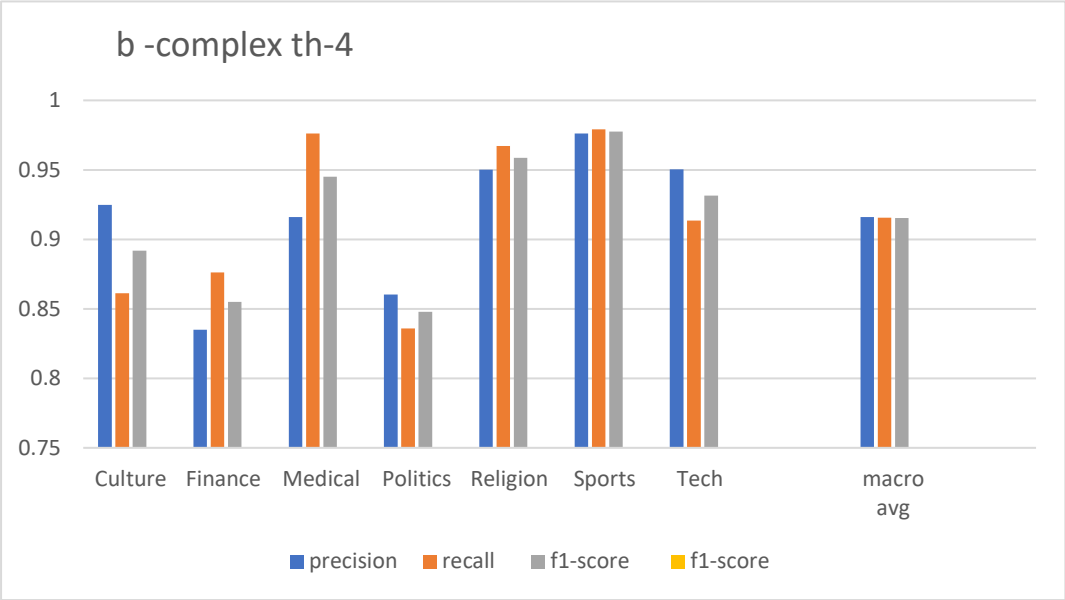
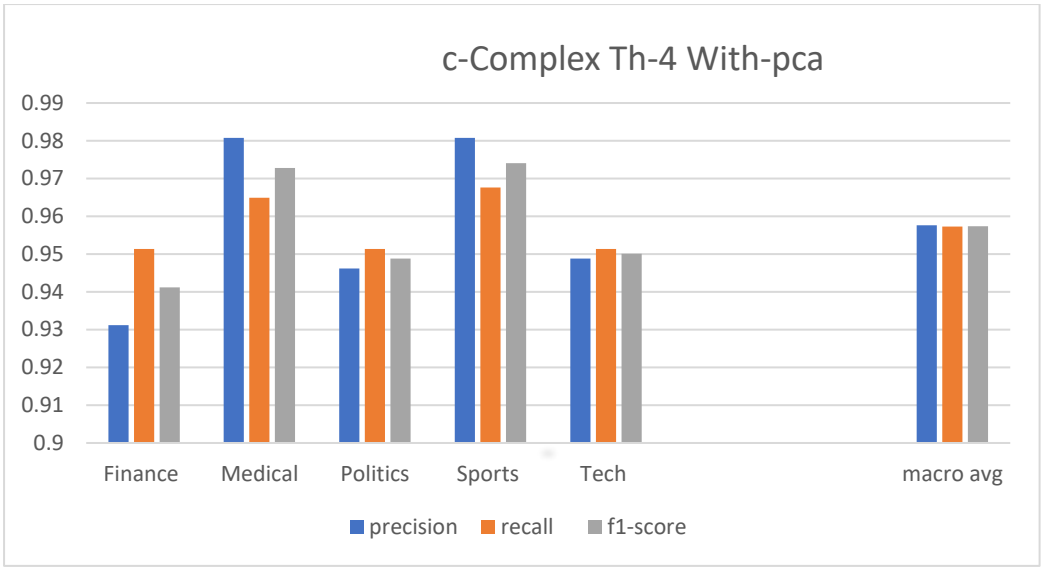
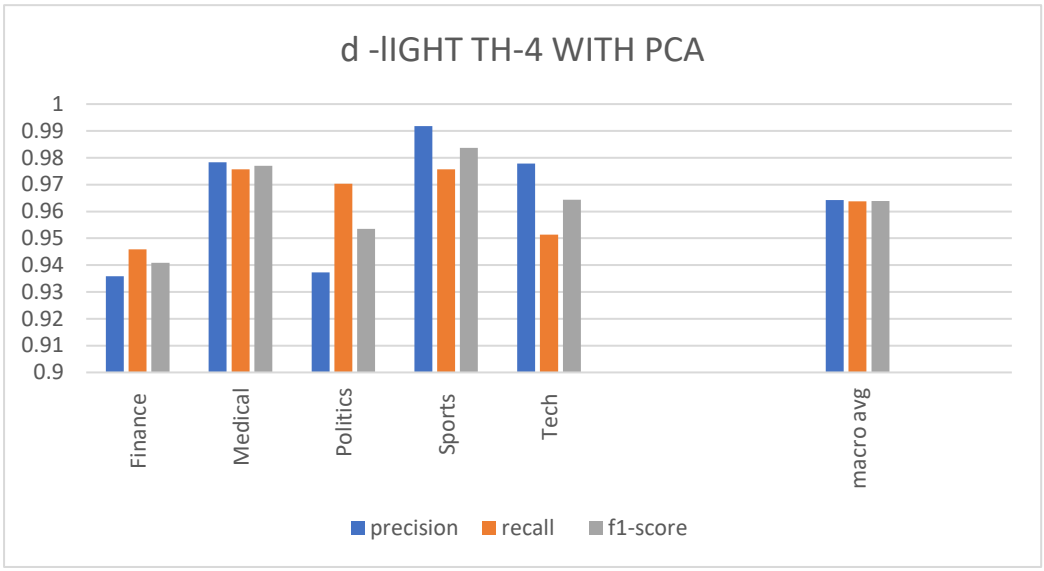


Figure (4.4): Dataset,output,akhbarona,th_4,a -light b -complex, USE PCA=None, Number Of Trees=100.





Figure(4.5): Dataset,output,Arabiya,th_4,c -complex,d -light, USE PCA=0.80, Number Of Trees=100

Second with threshold equal 2. The result are shown in figure (4.6) without PCA and Figure(4.7) with PCAfor both complex and light10 stemmers. Light stemming is best than complex stemming with less 0.01.

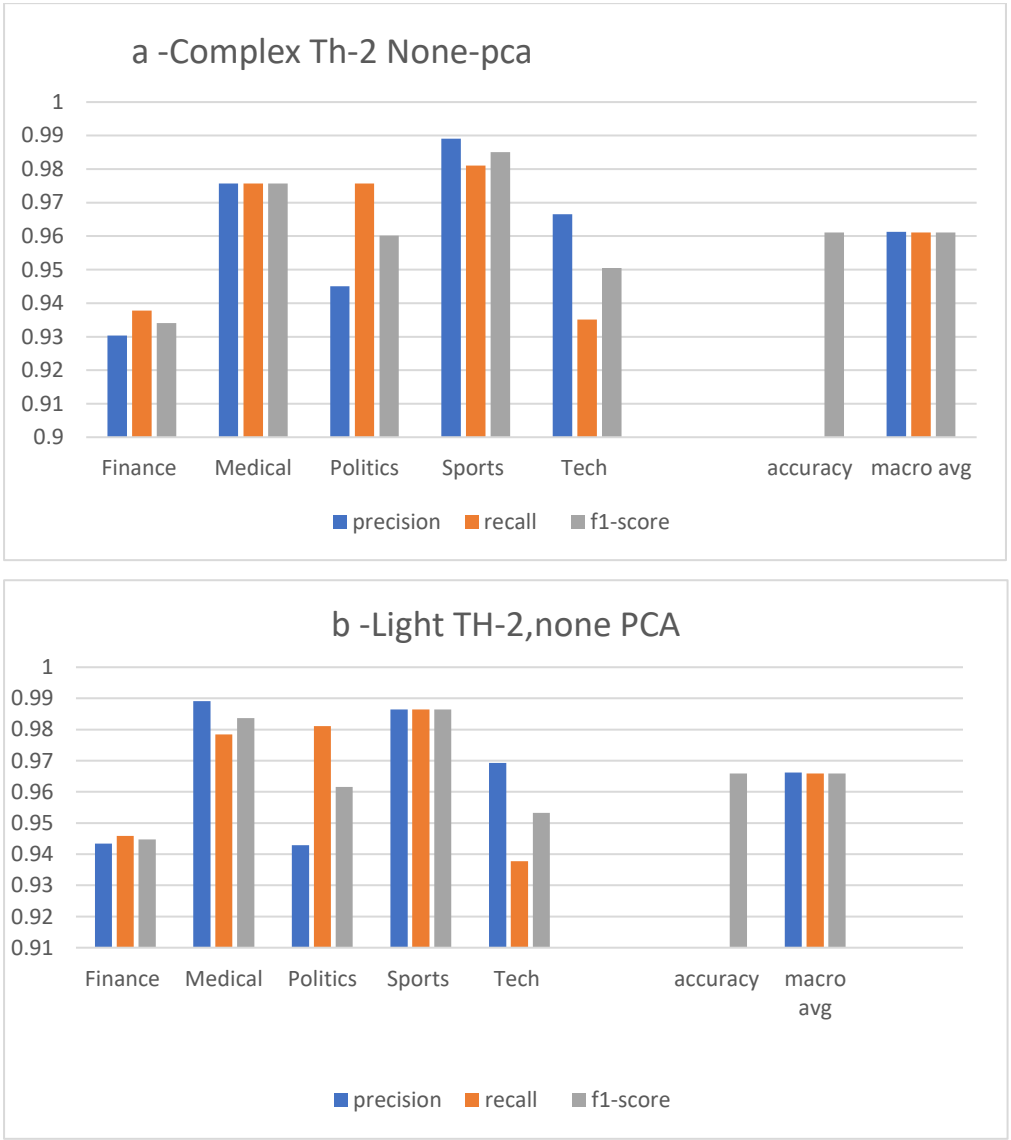
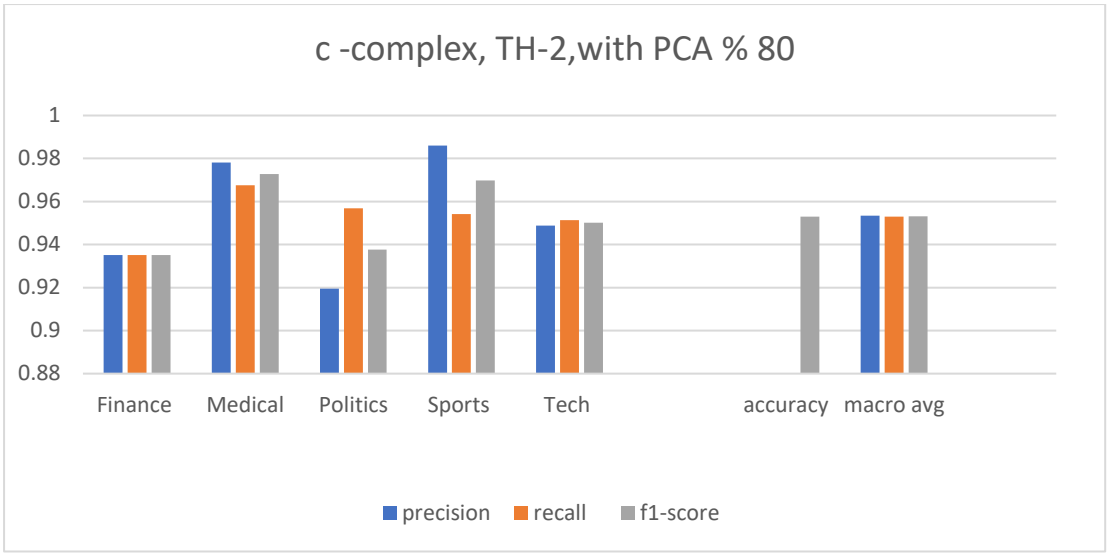


Figure (4.6):Arabiya,th_2,a -complex,b -light, USE PCA=None, Number Of Trees=100



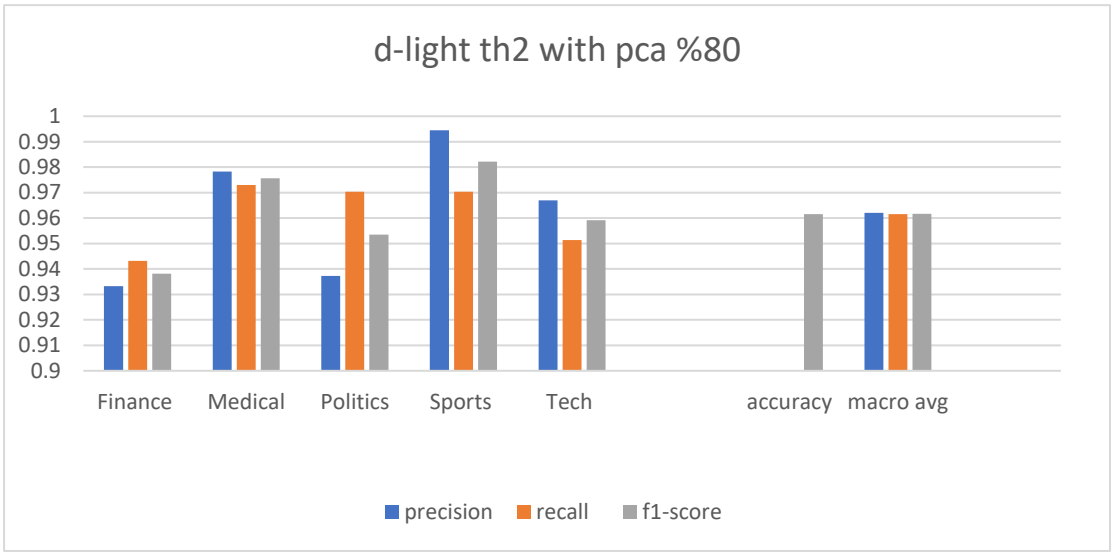


Figure (4.7):Arabiya,th_2,c -complex,d -light, USE PCA=%80, Number Of Trees=100

last with threshold equal 0. The result are shown in figure (4.8) without PCA and Figure(4.9) with PCAfor both complex and light10 stemmers. Light stemming is best than complex stemming with less 0.03 expected precision is more than 0.04 in light10 stemming but with PCA all is same with 0.02 complex stemming is best of light stemming .

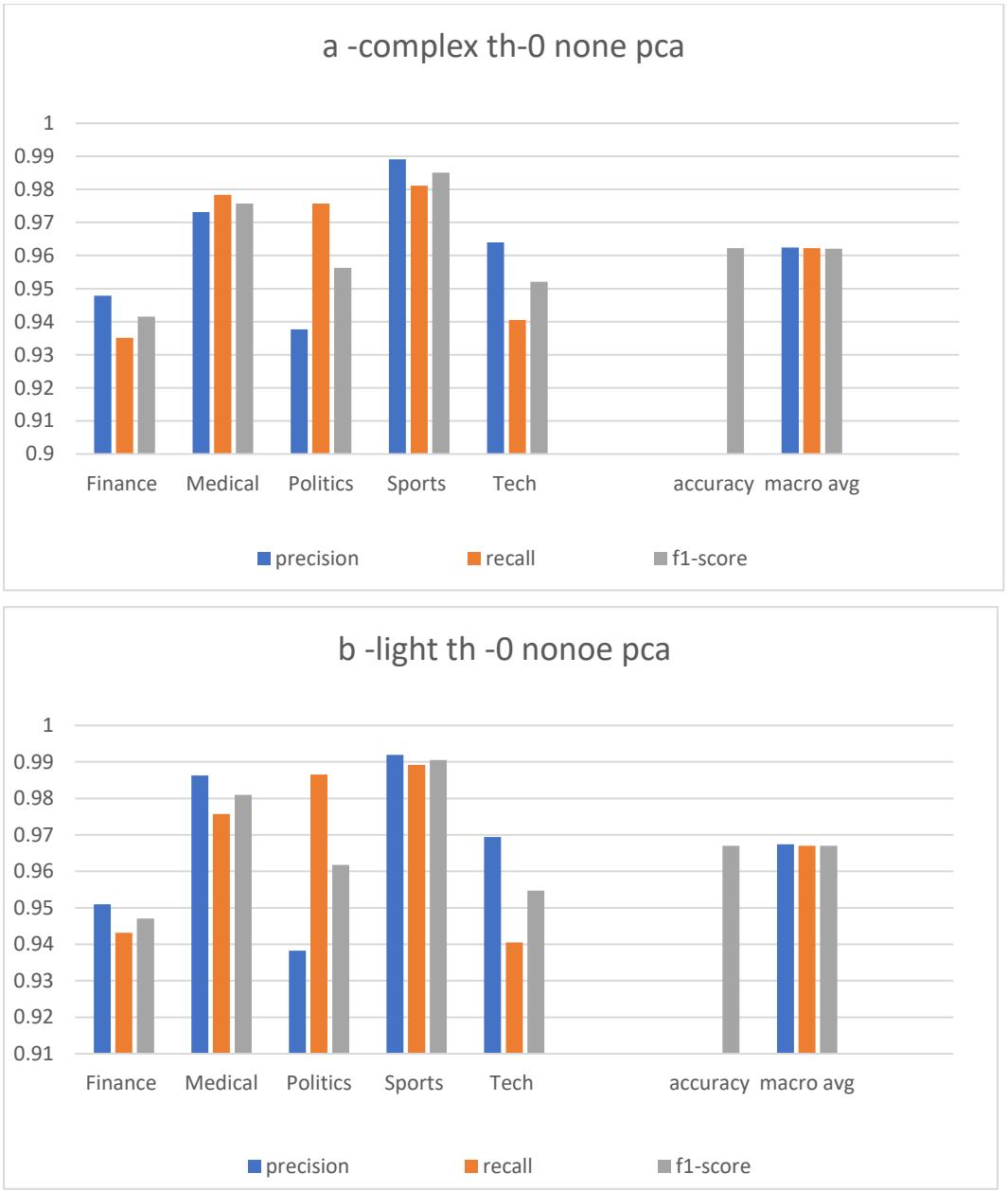
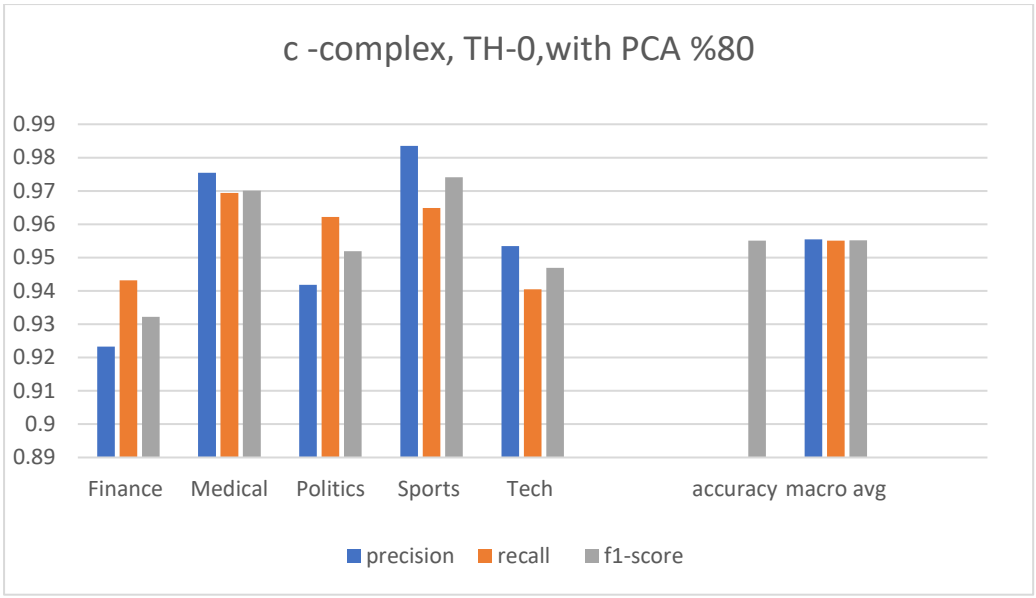


Figure (4.8):Arabiya,th_0,a -complex,b -light, USE PCA= none, Number Of Trees=100.



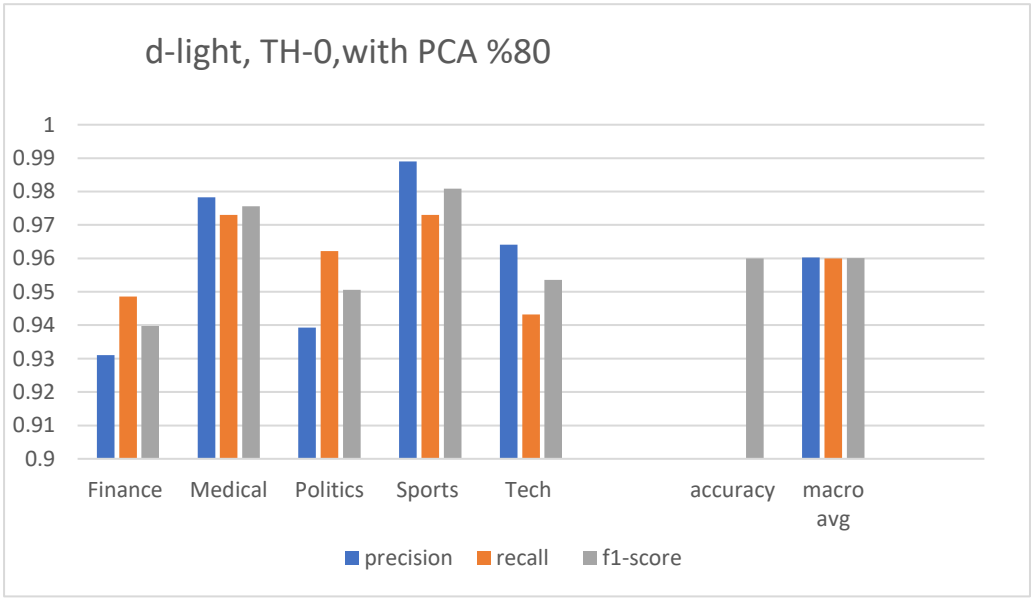


Figure (4.9):Arabiya,th_0,c -complex,b -light,USE PCA=%80, Number Of Trees=100.

after applied different experiment, t clear that light stemming is not better than complex in both stats with PCA or without a threshold is not effect on result. PCA and filter PCA are effect on the size of feature only not on accuracy of classification and when threshold increase accuracy of complex is little effect (it decreases with 0.01).

CHAPTER FIVE

CONCLUSION AND FUTURE WORKS

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

Based on the research findings, many conclusions can be drawn:

1. The proposed system's pre-processing procedures were carried out as a crucial stage since they can lead to the removal of extraneous information, undesirable letters, simples, and stop words from all texts. As a result, after deleting redundant characteristics, extracting vital features will be more accurate.
2. In spite of the fact that the Arabic dialect is exceptionally wide and cannot be constrained, in spite of all the challenges that this dialect has, we discover through the establishing of words that the method of light stemming enters the content and evacuates prefixes and suffixes and produces an awfully expansive content and a tall precision of up to 92%. The moment strategy of establishing is the complex strategy that employments the roots of words It comes about in a content much littler than light stemming, but with a tall precision of up to 91%, which is 1% less than light stemming,, meaning that the complex stemming, handle has less information, less preparing time, and less exactness than light stemming.

5.2 Future Work

Based on the findings of this thesis, a number of directions might be identified. Among them include, but are not limited to:

1. Analyzing material written in a different language. based on the perspectives of text users from many regions and cultures.
2. To use Root_based stemmer algorithm with two algorithm for to extraction best feature selection.
3. The proposed system opens the door for further research on analyzing Arabic texts .
4. Other data mining techniques like Nave Bayes (NB), Stochastic Gradient Descent and Logistic Regression may be utilized in the future to demonstrate their accuracy in classifying ArabicTexts..
5. Can be used, Clustering techniques as unsupervised learning algorithms .

5.4 REFERENCES

[1] https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

[2] Rafal Ali Sameer," Modified Light Stemming Algorithm for Arabic Language",Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq, 2016.

[3] Mohamed I. El-Disooqi, Waleed M. Arafa & Kareem M. Darwish, Stemming techniques of Arabic Language: Comparative Study from the Information Retrieval Perspective, The Egyptian Computer Journal, Vol. 36 No. 1, June 2009.

[4] Alshammari, Riyad,"Arabic Text categorization using machine learning approaches," Journal : International Journal of Advanced Computer Science and Applications, 2018.

[5]Khoja. "Stemming Arabic Text". Lancaster, U.K., Computing Department, Lancaster University, 1999.

[6] Waed Zaid, et al" Arabic Stemmer System based on Rules of Roots",Aplications International Journal of Information Technology and Language Studies (IJITLS), 2018.

[7] Y. Alhanini and M. Aziz, "The Enhancement of Arabic Stemming by Using Light Stemming and Dictionary-Based Stemming," Journal of Software Engineering and Applications, Vol. 4 No. 9, 2011, pp. 522-526. doi: 10.4236/jsea.2011.49060.

[8] Rouhia M. Sallam Yemen ,Hamdy M. Mousa ,Mahmoud Hussein ," Improving Arabic Text Categorization using Normalization and

Stemming Techniques" International Journal of Computer Applications (0975 – 8887) Volume 135 – No.2, February 2016 .

[9] Ali Alnaieda,Mosa Elbendakb,Abdullah Bulbulc ,"An intelligent use of stemmer and morphology analysis for Arabic information Retrieval",Egyptian Informatics Journal,2020. Volume 21, Issue 4, Pages 209-217Egyptian Informatics Journal, December 2020.

[10] Rafea Mohammed , " New Arabic Stemming based on Arabic Patterns", College of Islamic Science, The Iraqi University Baghdad, Iraq, 2016.

[11] Essam Kazem Mohammed Al-Yasir, "Arabic Sentiment Analysis for Identifying Terrorism Supporters on Twitter Using Data Mining Techniques",2019.

[12] Huda Abdulrahman Almuzaini, Aqil M. Azmi, "Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization", 2020.

[13] Sabria Mohammed Hussien,and Hazim J. Aburagheef, "Arabic light-based stemming: a comparative study among ligh10 stemmer, P-stemmer, and Conditional light stemmer", 4th International Conference on Engineering Technology and its Applications 2021- (4thiiceta2021).

[14] Arafat Awajan ,Arabic Text Preprocessing for the Natural Language Processing Applications ,Computer Science Department Princess Sumaya University for Technology, Amman, Jordan awajan@psut.edu.jo,2014.

[15] Jaffar Atwan, Mohammad Wedyan & Hadeel Al-Zoubi,Arabic Text Light Stemmer,Conference: Fourth International Conference on Trends in Computing and Information Technology (ICTCIT 2019).

[16] K. Taghva, R. Elkhoury, and J. Coombs, “Arabic stemming without a root dictionary,” *Int. Conf. Inf. Technol. Coding Comput. ITCC*, vol. 1, pp. 152–157, 2005, doi: 10.1109/itcc. 2005. 90.

[17]https://www.sciencedirect.com/science/article/pii/S1319157815000166#! . https://www.sciencedirect.com/science/journal/13191578 .

[18] https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

[19]https://www.google.com/search?q=types+of+stemming+algorithms&oq=&aqs=chrome.7.35i39i362l7j69i59i450.794762084j0j15&sourceid=chrome&ie=UTF-8.

[20] Khawla Ahmad AL Matarneh ,Dr. Mohammad Hassan “Conditional light Stemming For enhanced Arabic Information Retrieval “, Graduate Studies Zarqa University Zarqa – Jordan, 2017.

[21].https://en.wikipedia.org/wiki/Feature_extraction.

[22] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Pre-processing techniques for text mining-an overview," Int. J. Comput. Sci. Commun. Networks, vol. 5, no. 1, pp. 7–16, 2015.

[23] https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089.

[24] Aman Gupta, Feature Selection Techniques in Machine Learning,October 10, 2020.

https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning.

[25] https://www.projectpro.io/data-science-in-python-tutorial/principal-component-analysis-tutorial .

[26] https://analyticsindiamag.com/7-types-classification-algorithm.

[27] Amer Al-Badarenah, Emad Al-Shawakfa, Khaleel Al-Rababah, Safwan Shatnawi ,Basel Bani-Ismail,"Classifying Arabic Text Using KNN Classifier", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016 259 | P a g e (IJACSA)

www.ijacsa.thesai.org.

[28] Adel Hamdan Mohammad, Tariq Alwada'n, Omar Al-Momani, "Arabic Text Categorization Using Support vector machine, Naïve Bayes and Neural Network", GSTF Journal on Computing (JOC) DOI: 10.5176/2251-3043_4.4.360 ISSN:2251 - 3043 ; Volume 5, Issue 1; 2016 pp. 108-115 © The Author(s) 2016. This article is published with open access by the GSTF.

[29] https://www.researchgate.net/figure/The-structure-of-DT-and-RF_fig1_356431860.

[30] https://monkeylearn.com/blog/classification-algorithms/.

Appendix A

The Published Paper

Formal Acceptance

2nd International Conference of Information Technology to enhance E-learning and other Application-2021

[IT-ELA 2021]

To/ Sabria Mohammed Hussien, Hazim J. Aburagheef

Dear respected author(s):

With heartiest congratulations. We are pleased to inform you that based on the recommendations of the reviewers and the Technical Committees, your paper entitled:

"Arabic light-based stemming: a comparative study among ligh10 stemmer, P-stemmer, and Conditional light stemmer"

It has been accepted for oral presentation at the 2nd International Conference of Information Technology to enhance E-learning and other application-2021[IT-ELA 2021], technically sponsored by IEEE, to be held on Dec 28-29, 2021, Baghdad, Iraq.

Your paper will be submitted (**within the conference proceeding**) to the IEEE Xplore digital library for (**final acceptance of uploading to the Digital library**).

IT-ELA 2021 Baghdad – Iraq

Email:it-ela2021@baghdadcollege.edu.iq

Website: https://baghdadcollege.edu.iq/it-ela2021

Baghdad College of Economic Sciences University

Computer Sciences Department

ID: 1570778435

4th International Conference on Engineering Technology and its Applications 2021- (4thiiceta2021)

Arabic light-based stemming: a comparative study among ligh10 stemmer, P-stemmer, and Conditional light stemmer

Sabria Mohammed Hussien
Dept. of software, IT college
University of Babylon
Karbala, Iraq
sabriya.mohammed@student.uobabylon.edu.iq

Hazim J. Aburagheef
Dept. of software, IT collage
University of Babylon
Babil, Iraq
hazim.aburagheef@uobabylon.edu.iq

***Abstract*— Arabic stemming is a key stage in natural language processing's preprocessing (NLP). It takes affixes out of words. It improves text classification (TC) as well as information retrieval (IR). Light-based stemming and root-based stemming are the two types of stem. When compared to root-based stemming, light-based stemming consumes more energy. Only suffixes and prefixes are removed from the words. The light10 stemmer, the p-stemmer, and conditional light stemming (CondLight) are three well-known methods of light stemming. Prefixes and suffixes are removed by Light10 stemmers under a few conditions. Only prefixes are removed by the P-stemmer, while the CondLight stemmer is the same as the Light10 stemmer but with eight conditions. We measured the extent of improvement in Arabic TC by evaluating the stemmers. Three classifiers employ the Support Vector Machine (SVM), the k-nearest neighbor algorithm (KNN), Nave Bays (NB), and statistical similarity measurement. With stemming, the outcome indicates a small improvement (about 2 percent improvement).**

Keywords — Arabic language, natural language processing, text classification, Arabic stemming.

I. INTRODUCTION

Basic stemming consists of two types: light-based stemming and root-based stemming. Light stemming is also known as affix stripping, which involves removing a small number of prefixes and/or suffixes without attempting to discover roots, deal with infixes, or recognize patterns. Larkey [1] devised the basic approach, which generates a sequence of light streams like light1, light 2, light 3, light 8, and light 10 [2]. The light stemmer has the significant drawback of generating novel words (ambiguity), not working with irregular plural, and removing affixes without prior knowledge of linguistic rules [3]. The root-based stemmer approach relies mainly on the morphological analysis of words to extract the roots of the Arabic words. Kohja stemmer is most popular root-based stemmer. It removes suffix, infix, and prefix, before matches with the pattern. It has over-stemming and miss-stemming problems. Over-stemming is root generation from different words and miss-stemming is removing actual affixes from words [4]. The light stemmer is better than the root-based stemmer in IR and TC [5].

The fundamental light stemmer just eliminates suffixes and prefixes, as in light10, or only prefixes, as in P-stemmer [6]. The light stemmer improves by adding a new rule to the basic light stemmer. The rule is to enhance the basic method and reduction the error of the stemmer [7]. There are three well-known light-based stemming methods: light10, P-stemmer, and CondLight stemmer. Table I list the suffixes and prefixes of well-known light stemmers.

TABLE I. LIST OF REMOVED AFFIXES FOR KNOWN LIGHT STEMMERS [6].

Methods	Remove prefixes	Remove suffix
Light1	ال, و, ال, بال, كال, فال	None
Light2	ال, و, ال, بال, كال, فال, و	none
Light3	ال, و, ال, بال, كال, فال, و	ه, ه
Light8	ال, و, ال, بال, كال, فال, و	هاتان, ات, ون, ين, يه, يه, ه, ه, ي
Light10	ال, و, ال, بال, كال, فال, ل, ل, و, ل, و, ل	هاتان, ات, ون, ين, يه, يه, ه, ه, ي
CondLight	ال, و, ال, بال, كال, فال, ل, ل, ب, ب, ك, ك, ب, ب, ف, س, و	هاتان, ات, ون, ين, يه, يه, ه, ه, ي, ه, هن, هن, و
P-stemmer	ال, و, ال, بال, كال, فال, ل, و, ل, و, ل	None

CondLight stemmer is an improved Larkey light stemmer. It introduces new prefixes (بـ, سـ, لـ, فـ, كـ, بـ) and new suffixes (ها, هن, هم, ون). It creates a new condition for each suffix and prefix. It creates pre-defined conditions for nouns and verbs. The purpose of setting conditions is to distinguish the prefix and suffix from the original parts of the word. The prefix and suffix remove only when the condition is satisfied [7].

TABLE II. LIGHT STEMMERS APPLIES ON LIST WORD.

word	Light 10	P-stemmer	candlelight
يدرسها	يدرس	درسها	درس
لمهرجانك	مهرجان	مهرجانك	مهرجان
بهجمات	بهجم	بهجمات	هجم
لسببها	لسبب	لسببها	سبب
باضافة	باضاف	باضافة	اضاف
المباحثات	مباحث	مباحثات	مباحث
بقتل	بقتل	بقتل	بقتل
بلدان	بلد	بلدان	لد
تعديلات	تعديل	تعديلات	عديل

TABLE III. P-STEMMER DISADVANTAGE.

word	والمعلم	فالمعلم	كالمعلم	المعلمون	المعلمات
Light 10	معلم	معلم	معلم	معلم	معلم
P-stemmer	معلم	معلم	معلم	**معلمون**	**معلمات**
candlelight	معلم	معلم	معلم	معلم	معلم

Appendix B

The confusion matrix is also provided bellow for every test

Test 1: Akhbarona\th_0\complex, USE PCA=None

	precision	recall	f1-score	support
Culture	0.9227	0.8552	0.8877	670
Finance	0.8553	0.8731	0.8641	670
Medical	0.9212	0.9776	0.9486	670
Politics	0.8486	0.8284	0.8384	670
Religion	0.9414	0.9597	0.9505	670
Sports	0.9719	0.9806	0.9762	670
Tech	0.9395	0.9269	0.9331	670

Test 2: Akhbarona\th_0\complex, USE PCA=80%

Not Available

Test 3: Akhbarona\th_0\lite, USE PCA=None

	precision	recall	f1-score	support
Culture	0.9225	0.8701	0.8955	670
Finance	0.8549	0.8881	0.8712	670
Medical	0.9173	0.9761	0.9458	670
Politics	0.8600	0.8343	0.8470	670
Religion	0.9463	0.9731	0.9595	670
Sports	0.9835	0.9806	0.9821	670
Tech	0.9579	0.9179	0.9375	670

Test 4: Akhbarona\th_0\lite, USE PCA=80%

Not Available

Test 5: Akhbarona\th_2\complex, USE PCA=None

	precision	recall	f1-score	support
Culture	0.9276	0.8791	0.9027	670
Finance	0.8412	0.8776	0.8590	670
Medical	0.9315	0.9746	0.9526	670
Politics	0.8675	0.8403	0.8537	670
Religion	0.9541	0.9627	0.9584	670
Sports	0.9777	0.9821	0.9799	670
Tech	0.9406	0.9224	0.9314	670

Test 6: Akhbarona\th_2\complex, USE PCA=80%

	precision	recall	f1-score	support
Culture	0.9098	0.8881	0.8988	670
Finance	0.8253	0.8672	0.8457	670
Medical	0.9351	0.9672	0.9508	670
Politics	0.8580	0.8299	0.8437	670
Religion	0.9520	0.9478	0.9499	670
Sports	0.9685	0.9642	0.9663	670
Tech	0.9330	0.9149	0.9239	670

Test 7: Akhbarona\th_2\lite, USE PCA=None

	precision	recall	f1-score	support
Culture	0.9274	0.8776	0.9018	670
Finance	0.8543	0.8925	0.8730	670
Medical	0.9263	0.9761	0.9506	670
Politics	0.8727	0.8493	0.8608	670
Religion	0.9576	0.9776	0.9675	670
Sports	0.9807	0.9866	0.9836	670
Tech	0.9563	0.9134	0.9344	670

Test 8: Akhbarona\th_2\lite, USE PCA=80%

Not Available

Test 9: Akhbarona\th_4\complex, USE PCA=None

	precision	recall	f1-score	support
Culture	0.9247	0.8612	0.8918	670
Finance	0.8350	0.8761	0.8551	670
Medical	0.9160	0.9761	0.9451	670
Politics	0.8602	0.8358	0.8478	670
Religion	0.9501	0.9672	0.9586	670
Sports	0.9762	0.9791	0.9776	670
Tech	0.9503	0.9134	0.9315	670

Test 10: Akhbarona\th_4\complex, USE PCA=80%

	precision	recall	f1-score	support
Culture	0.8897	0.8910	0.8904	670
Finance	0.8225	0.8716	0.8464	670
Medical	0.9298	0.9687	0.9488	670
Politics	0.8618	0.8284	0.8447	670
Religion	0.9531	0.9403	0.9467	670
Sports	0.9655	0.9612	0.9634	670
Tech	0.9484	0.9045	0.9259	670

Test 11: Akhbarona\th_4\lite, USE PCA=None

	precision	recall	f1-score	support
Culture	0.9280	0.8851	0.9060	670
Finance	0.8644	0.8851	0.8746	670
Medical	0.9263	0.9761	0.9506	670
Politics	0.8670	0.8463	0.8565	670
Religion	0.9519	0.9746	0.9631	670
Sports	0.9793	0.9866	0.9829	670
Tech	0.9534	0.9164	0.9346	670

Test 12: Akhbarona\th_4\lite, USE PCA=80%

Not Available

Test 13: AlArabiya\th_0\complex, USE PCA=None

	precision	recall	f1-score	support
Finance	0.9479	0.9351	0.9415	370
Medical	0.9731	0.9784	0.9757	370
Politics	0.9377	0.9757	0.9563	370
Sports	0.9891	0.9811	0.9851	370
Tech	0.9640	0.9405	0.9521	370

Test 14: AlArabiya\th_0\complex, USE PCA=80%

	precision	recall	f1-score	support
Finance	0.9233	0.9432	0.9332	370
Medical	0.9754	0.9649	0.9701	370
Politics	0.9418	0.9622	0.9519	370
Sports	0.9835	0.9649	0.9741	370
Tech	0.9534	0.9405	0.9469	370

Test 15: AlArabiya\th_0\lite, USE PCA=None

	precision	recall	f1-score	support
Finance	0.9510	0.9432	0.9471	370
Medical	0.9863	0.9757	0.9810	370
Politics	0.9383	0.9865	0.9618	370
Sports	0.9919	0.9892	0.9905	370
Tech	0.9694	0.9405	0.9547	370

Test 16: AlArabiya\th_0\lite, USE PCA=80%

	precision	recall	f1-score	support

Finance	0.9310	0.9486	0.9398	370
Medical	0.9783	0.9730	0.9756	370
Politics	0.9393	0.9622	0.9506	370
Sports	0.9890	0.9730	0.9809	370
Tech	0.9641	0.9432	0.9536	370

Test 17: AlArabiya\th_2\complex, USE PCA=None

	precision	recall	f1-score	support
Finance	0.9303	0.9378	0.9341	370
Medical	0.9757	0.9757	0.9757	370
Politics	0.9450	0.9757	0.9601	370
Sports	0.9891	0.9811	0.9851	370
Tech	0.9665	0.9351	0.9505	370

Test 18: AlArabiya\th_2\complex, USE PCA=80%

	precision	recall	f1-score	support
Finance	0.9351	0.9351	0.9351	370
Medical	0.9781	0.9676	0.9728	370
Politics	0.9195	0.9568	0.9377	370
Sports	0.9860	0.9541	0.9698	370
Tech	0.9488	0.9514	0.9501	370

Test 19: AlArabiya\th_2\lite, USE PCA=None

	precision	recall	f1-score	support
Finance	0.9434	0.9459	0.9447	370
Medical	0.9891	0.9784	0.9837	370
Politics	0.9429	0.9811	0.9616	370
Sports	0.9865	0.9865	0.9865	370
Tech	0.9693	0.9378	0.9533	370

Test 20: AlArabiya\th_2\lite, USE PCA=80%

	precision	recall	f1-score	support
Finance	0.9332	0.9432	0.9382	370
Medical	0.9783	0.9730	0.9756	370
Politics	0.9373	0.9703	0.9535	370
Sports	0.9945	0.9703	0.9822	370
Tech	0.9670	0.9514	0.9591	370

Test 21: AlArabiya\th_4\complex, USE PCA=Non

	precision	recall	f1-score	support

Finance	0.9482	0.9405	0.9444	370
Medical	0.9810	0.9757	0.9783	370
Politics	0.9381	0.9838	0.9604	370
Sports	0.9892	0.9865	0.9878	370
Tech	0.9693	0.9378	0.9533	370

Test 22: AlArabiya\th_4\complex, USE PCA=80%

	precision	recall	f1-score	support
Finance	0.9312	0.9514	0.9412	370
Medical	0.9808	0.9649	0.9728	370
Politics	0.9462	0.9514	0.9488	370
Sports	0.9808	0.9676	0.9741	370
Tech	0.9488	0.9514	0.9501	370

Test 23: AlArabiya\th_4\lite, USE PCA=None

	precision	recall	f1-score	support
Finance	0.9409	0.9459	0.9434	370
Medical	0.9837	0.9757	0.9796	370
Politics	0.9452	0.9784	0.9615	370
Sports	0.9919	0.9892	0.9905	370
Tech	0.9694	0.9405	0.9547	370

Test 24: AlArabiya\th_4\lite, USE PCA=80%

	precision	recall	f1-score	support
Finance	0.9358	0.9459	0.9409	370
Medical	0.9783	0.9757	0.9770	370
Politics	0.9373	0.9703	0.9535	370
Sports	0.9918	0.9757	0.9837	370
Tech	0.9778	0.9514	0.9644	370

جمهورية العراق
وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية تكنولوجيا المعلومات
قسم البرمجيات

تعزيزاختيار الصفات في تصنيف النص للغه العربيه

أطروحة

مقدمة إلى مجلس كلية تكنولوجيا المعلومات في جامعة بابل والتي هي جزء من متطلبات نيل درجة الماجستير في تكنولوجيا المعلومات – البرمجيات

من قبل

صبريه محمد حسين عبعوب

باشراف

الدكتورحازم جليل حسن ابورغيف

هـ 1444 م 2022

الخلاصة

عمليةالاشتقاق هي خطوة أساسية في خطوة ما قبل المعالجة. يتم استخدام العمليه لتقليل ابعاد عن طريق تقليل عدد الميزات . تقوم بإرجاع كل كلمة إلى جذر الكلمات

. إنها عملية إزالة مجموعة البادئات واللواحق في طريقة الاشتقاق الخفيفه وتجميع الكلمة لإنتاج الجذر في طريقة الاشتقاق المعقدة, يحول الجمع إلى مفردات على وزن فعل.

يتكون النظام المقترح من خمس مراحل رئيسية. الأول هو الحصول على البيانات من مجموعة سند الفرعية. والثاني هو المعالجة المسبقة للبيانات التي تم إجراؤها من خلال تطبيق طرق مختلفة. تتضمن المرحلة الثالثة استخراج الميزات والاختيار باستخدام نهج تردد المستند المعكوس للتردد(TF-IDF).

. (Random Forest)رابعًا ، تم تطبيق طرق التصنيف باستخدام الغابات العشوائية علاوة على ذلك ، تم تقييم نتائج خوارزميات التصنيف استنادًا إلى f1- وrecall و precision وaccuracyقياس الأداء باستخدام مقاييس مع طريقة الاشتقاق . أظهرت النتائج أنه تم تحقيق أعلى دقة من خلال تطبيقscore f1- score وrecallو precision وacuracy حيث بلغت في الاشتقاق الخفيف الى حوالي:

٠,٩٦٦٢٪ و ٠,٩٦٥٩٪ و ٠,٩٦٥٩٪ والدقة في التعقيد و هي ٠,٩٦١٣٪ و ٠,٩٦١١٪ و ٠,٩٦١١.

تظهر نتيجة التقييم أن المشتق المعقد أفضل من المشتق الخفيف. إنه يقلل حجم المتجه من الحجم الإجمالي للمصطلح ١٥٢٧٤٩ إلى ١٥٦٢٨في المعقد و ٥٠٧٧٦

في الاشتقاق الخفيف، ولكن الدقة في الاشتقاق الخفيف ليس أفضل من المشتق المعقد. وهو أكثر من دقة المشتق المعقدة ب ٠,٠١ فقط .