Document Classification Method Based on Contents Using an Improved Multinomial Naïve Bayes Model

طريقة تصنيف الوثيقة استنادا إلى محتوياتها باستخدام تَحسِين نموذج متعدد الحدود نيف بايز

Done By

Junaina Jamil Najim Aldin AL-Bayati

Supervisor

Dr. Mohammed A. F. Al-Husainy

A thesis Submitted in Partial Fulfillment of the Requirements for the Master Degree in Computer Science

Department of Computer Science

Faculty of Information Technology

Middle East University

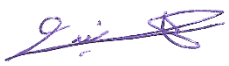
August/ 2015

Authorization

I Junaina Jamil Najim Aldin AL-Bayati authorize Middle East University (MEU) to provide libraries, organizations and even individuals with copies of my thesis when required.

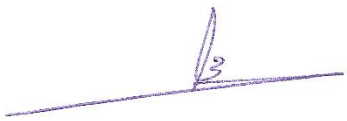
Name: Junaina Jamil Najim Aldin AL-Bayati

Date: Sat. 30 August, 2015

Signature: 

Examination Committee Decision

This is to certify that the thesis entitled **"Document Classification Method Based on Contents Using an Improved Multinomial Naïve Bayes Model"** was successfully defended and approved on 30/8/2015.

Examination Committee Member	**Signature**
1. Dr. Ghassan Ghazi Kanaan	
2. Dr. Mudhafar Al-Jarrah	
3. Dr. Mohammed Al-Husainy	

Acknowledgments

In the first, my thanks are hereby extended to my God, then to Dr. Mohammed A. F. Al-Husainy for his supportive and helpful supervision, as well as for assisting a student in every step of the project, and for providing important information and basics, which was very important for the successful implementation of the project. Further thanks are extended to everyone who helped me develop my understanding of the various nuances of the project and for everyone who believes that the knowledge is right for everyone.

Dedication

I would like to exploit this opportunity to dedicate this project to my father, mother, brothers and sister, without whose invaluable support. I would have not been able to have achieved this in my lifetime.

May God bless them.

Document Classification Method Based on Contents Using an Improved Multinomial Naïve Bayes Model

By

Junaina Jamil Najim Aldin AL-Bayati

Supervisor

Dr. Mohammed A. F. Al-Husainy

ABSTRACT

Currently, there are a lot of Arabic documents that are available in the most of applications in our lives, these Arabic documents have to be systematized and categorized according to a particular topic to be more expressive and more employed, the text classification was one of the approaches that used to arranged the Arabic documents, where the classifications of the Arabic documents were the technique to determine for which topic this text is related to, numerous studies were accompanied about this discipline to increase the performance of the document classification particularly the Arabic document, the Arabic linguistic is treasure and an actual complex inflectional language that changes the modest and normal approaches to difficult one . This research involved in improving and promoting the performance of the multinomial naive Bayes (MNB) classification by using three different approaches; at first by addition only the n-gram, the another one by applied the TF-IDF, and lastly by using both of n-gram and TF-IDF, then these improved classifiers had been evaluated based on the estimated values of the recall, precision and F-measure for each classifier next to apply it over the Arabic data set that covers six classes which

involved about 1500 arabic document dissimilar document. The average of F-measure for all classes when applying the bigram was (81.46%), while the average of F-measure for all classes when applying TF-IDF was (88.88%) and the average of F-measure for all classes when applying the combination of both bigram and TF-IDF was (89.70%). The variance F-measure between different suggested classifiers verified that the classifier which is enhanced by using both of the TF-IDF and bigram accomplished the highest values and it characterizes as the most effective classifier between the three suggested classifier. In the second stage of effectiveness, the classifier that enhanced by using only TF-IDF and finally the classifier which enhanced by using only the bigram.

Keywords: Multinomial Naïve Bayes, TF-IDF(Term Frequency-Inverse Document Frequency), N-gram , Data Set Arabic, Tokenization, Stemming, Remove Stop Words .

طريقة تصنيف الوثيقة استنادا إلى محتوياتها باستخدام تَحسِين نموذج متعدد الحدود نيف بايز

الطالبه

جنينه جميل نجم الدين البياتي

المشرف

الدكتور محمد عباس فاضل الحسيني

ملخص الرساله

في الوقت الحاضر، هناك الملايين من الوثائق التي تتوفر في معظم مجالات في حياتنا. يجب تنظيم هذه الوثائق وتصنيفها تحت موضوع معين ليكون أكثر وضوحا، وللاستفادة منها بشكل أفضل. تصنيف النصوص هي إحدى الأساليب التي تطبق بهدف تنظيم الوثائق، ويمكن تعريف تصنيف النصوص بأنه أسلوب يتم من خلاله اكتشاف المجموعة التي تنتمي إليها وثيقة معينة، لقد تم إجراء العديد من الأبحاث حول هذا العلم وما زالت تلك الأبحاث تجرى إلى الوقت الحالي بهدف التحسين من فعاليتها وخاصة عملية تصنيف الوثائق والنصوص العربية.وذلك لأن اللغة العربية غنية و معقدة للغاية، فالعمليات التي تجرى على النصوص العربية تكون أكثر تعقيدا من غيرها. في هذا البحث نهدف إلى زيادة وتعزيز أداء متعددة الحدود نيف بايز باستخدام ثلاث طرق. أول طريقة بإضافة فقط N-gram ، والثاني باستخدام (Term Frequency-Inverse Document Frequency)(TF-IDF)، وأخيرا عن طريق دمج (N-gram) و (TF-IDF)، ومن ثم تم تقييم هذه المصنفات اعتمادا على نتائج وقيم كل من Recall, Precision, and F-measure . تم تطبيق المصنفات المقترحة على قاعدة البيانات العربية والتي تحتوي على ستة مجموعات وعدد الوثائق في جميع المجموعات ما يقارب 1500 وثيقة عربيه مختلفة .

وكان متوسط F-measure لجميع classes عند تطبيق bigram (81.46٪)، في حين أن متوسط F-measure عند تطبيق (TF-IDF) (88.88٪) ومتوسط F-measure عند تطبيق مزيج من كل bigram و(TF-IDF) كان (89.70٪).

هذه النتائج أثبتت أن المصنف الذي يطبق كل من (TF-IDF) و (bigram) هو أفضل مصنف بين المصنفات الثلاثة المقترحه .

الكلمات المفتاحيه: متعدد الحدود نيف بايز , TF-IDF , ن– غرام , مجموعة بيانات عربيه , Stemming , Tokenization , حذف كلمات التوقف .

Abbreviations

MNB	Multinomial Naïve Bayes
TF-IDF	Term Frequency-Inverse Document Frequency
NB	Naïve Bayes
P	Precision
R	Recall
F	F-measure
EM	Expectation -Maximization
KNN	K Nearest Neighbor
IF	Information Filtering
NLTK	Natural Language Toolkit
SVM	Support Vector Machines
HMM	Hidden Markov Models
WSD	Word Sense Disambiguation

Table of Contents

AUTHORIZATION II
Examination Committee Decision III

ACKNOWLEDGMENTS IV
DEDICATION V
ABSTRACT IV
ملخص الرساله VI
ABBREVIATIONS VIII
TABLE OF CONTENTS X
LIST OF FIGURES XII
LIST OF TABLES XIV
CHAPTER ONE 1
INTRODUCTION 1
1.1. INTRODUCTION 1
1.2. BACKGROUND 1
1.3. TERMINOLOGY 2
1.4. PROBLEM STATEMENT 4
1.5. MOTIVATION 5
1.6. OBJECTIVE 5
1.7. METHODOLOGY 6
CHAPTER TWO 10
LITERATURE REVIEW 11
2.1. OVERVIEW 11
2.2. RELATED WORK 11
2.2.1. TYPES OF DOCUMENT CLASSIFICATION 12
2.2.2. THE DISADVANTAGES AND CHALLENGES OF DOCUMENT CLASSIFICATIO 13
2.2.3. DOCUMENT CLASSIFICATION APPLICATIONS 12
CHAPTER THREE 24
METHODOLOGY AND PROPOSED MODELS 24
3.1. DESIGN APPROACH 24
3.2. DATA SET 27
3.3. OVERVIEW OF ARABIC LANGUAGE 28
3.4. BAG OF WORDS MODEL 29
3.5. PREPROCESSING ISSUES 29
3.5.1. TOKENIZATION 30
3.5.2. STOP WORDS 30
3.5.3. STEMMING 31

3.6. Multinomial Naïve Bayes and Document Classification 34
3.7. Term Frequency - Inverse Document Frequency (TF-IDF) 36
3.8. N-Grams 37
3.9. The Proposed Classifiers 38
3.10. The Evaluation Criteria 39
3.11. The Implementation 40
3.11.1. The Graphical User Interface 40

Chapter Four 41
The Results 41
4.1. Multinomial Naïve Bayes without Enhancements 42
4.2. Enhanced Multinomial Naïve Bayes by Bi-gram Classifier 42
4.3. Enhanced Multinomial Naïve Bayes by Tri-gram Classifier 43
4.4. Enhanced Multinomial Naïve Bayes by 4-gram classifier 44
4.5. Enhanced Multinomial Naïve Bayes by TF-IDF Classifier 44
4.6. Enhanced Multinomial Naïve Bayes by both of Bi-gram and TF-IDF ... 45
Chapter Five 47
The Evaluation 47
5.1. Evaluation and Discussion 47
5.2. Comparison between Different Types of N-gram 47
5.3. Comparison between Proposed Classifier 50
Chapter Six 55
Conclusion and Future Work 55
6.1. Conclusion 55
6.2. Futhure Work 55
References 57
Appendices 63
A. The Implemented of Classes 63
B. The Results 69

List of Figures

Figure 3-1: the classifier model building.....26

Figure 3-2: the data set which used in expirements.....28

Figure 3-3: the charactersin Aarabic language 29

Figure 3-4: document preprocessing issues.....30

Figure 3-5: example of tokenization.....30

Figure 3-6: example of some stop words in the Arabic document 31

Figure 3-7: subset of grammatical suffixes and prefixes.....32

Figure 3-8: Viewing the results of light stemming and heavy stemming in terms of over-stemming and under-stemming percentages..... 34

Figure 3-9: Example of implement the bi-gram 37

Figure 3-10: The initial GUI of the system 41

Figure 5-1: Recall and precision and F-measure for bigram 48

Figure 5-2: Recall and precision and F-measure for trigram.....48

Figure 5-3: Recall and precision and F-measure for 4-gram.....49

Figure 5-4 : variation of f-measures between different n-gram.....50

Figure 5-5: Recall and precision and F-measure for bigram 51

Figure 5-6: Recall and precision and F-measure for TF-IDF 51

Figure 5-7: Recall and precision and F-measure for TF-IDF and bigram 52

Figure 5-8 : F-measure for different proposed classifiers 54

Figure A-1: the implemented classes…….....63

Figure A-2: the classifier classes…….....64

Figure A-3: Document class…… ……..... 64

Figure A-4: likelihood class ……..……. .65

Figure A-5: stop word handler class ……..…….…….65

Figure A-6: Term class ……..…… . 65

Figure A-7: Test document class ……..……...66

Figure A-8: Util class……...……66

Figure A-9: Naïve Bayes Classifier class…...……...……..67

Figure A-10: the class diagram of the stemmer name space……...……...…..67

Figure A-11: class diagram for the classification name space……..68

Figure B-1: the Recall and precision for the MNB classifier without enhancement……69

Figure B-2: The Recall and precision for the MNB classifier with bi-gram classifier……....70

Figure B-3: the Recall and precision for the MNB classifier with tri-gram classifier…….71

Figure B-4: the Recall and precision for the MNB classifier with 4-gram classifier……..…....72

Figure B-5: the Recall and precision for the MNB classifier with TF-IDF classifier……..…...73

Figure B-6: the Recall and precision for the MNB classifier with TF-IDF and bi-gram classifier
….....74

List of Tables

Table 3-1: the data set categories.....27

Table 3-2: words derived from the ground root (جمع "JM3") (Mustafa, 2012).....32

Table 4-1: the recall and precision for the MNB classifier without enhancement.....42

Table 4-2: the Recall and precision for the MNB classifier with bi-gram classifier.....43

Table 4-3: The Recall and precision for the MNB classifier with tri-gram classifier.....44

Table 4-4: the Recall and precision for the MNB classifier with 4-gram classifier.....44

Table 4-5: the Recall and precision for the MNB classifier with TF-IDF classifier.....45

Table 4-6: the Recall and precision for the MNB classifier with TF-IDF and bi-gram classifier46

Table 5-1: variation of Recall and precision and F-measures among different n-gram.....49

Table 5-2:variation of Recall and precision and F-measures among proposed classifiers.....52

Chapter One

Introduction

1.1. Introduction

A long time ago there were many problems in libraries in classification, document classification is a process to solve these problems also it helps to allocate the document to it specific subject. Many researches have been done about this topic of science but it has been still on progress, the target of this research is to enhance the document classification also to promote the accuracy of classification without affecting on the classification time.

The algorithms which will be carried out in this research are the Naive Bayes theorem, for the reason that it improves the classification accuracy and the speed of the classification. To have this specification we will be used the Multinomial Naive Bayes model "MNB" which enhanced by using some modification like the bi-gram and TF-IDF.

The document classification would be outlined as a content-based task of one or more predefined subjects of documents. In reality the classifications of documents have many applications like the articles filtering for expert workers; spam emails filtering for customers moreover to use it in the government sector (Goller, Löning, Will and Wolff, 2000).

1.2. Background

The traditional classifications were the most effective way and the most accurate way to manage and organize the various kinds of the text document due to the dependency on the users experiences, and on the schematic meaning of the document, then it classifies under the suitable

classes, every class goes back to a particular topic, sometimes one text document may be go back to more than one topic with dissimilar ratios. That is happened when the content of it integrates among a lot of topics, but this technique when utilizing a huge amount of data become worthless and ineffective process, because it requires a lot of effort and time (Anagnostopoulos, Broder and Punera, 2006).

A lot of classification functions have been resolved manually, as an example books in libraries are categorized by the librarian, but it's a very costly process. The ("handcrafted rules") is a second classification process that utilizing the "standing queries", each query consists of a large combination of all keywords that affined to a certain subject, generally, these queries were written by an experienced person who has a high practice in such affined topics, also has the capability to write this rules in a way that could match the computer classification tool writing, but finding an expert persons is not an easy mission (Manning, Raghavan and Schütze, 2008).Machine learning-based document classification, is an another process in the text classification, in this method, the text classification tool is learned from training data automatically (Mitchell, 1997). Also in this process of classification there were two types of classification (Padhye, 2006), which take part in the observed document classification; where in the observed classification, there is the training group which outlined by a set of classes. While in the non-observed document classification, there is a training group of the predefined classes, where the classes have been generated based on the content of documents (Patrick, 2009).

1.3. Terminology

TheDocument classification: defined as a content-based task of one or more predefined subjects of documents (Goller,Löning , Will and Wolff, 2000).

Stemming: reducing the amount of the initial features, by ignoring the features which have the similar stem also by ignoring the misspelled (Ikonomakis,Kotsiantis and Tampakas ,2005).

Tokenization: defined as the procedure of exchanging the full text document to an assortment of separated components by breaking down the text corpus (Alkafije and Ajam, 2013).

Naïve Bayes: None complex probabilistic classifier that based on utilizing the Bayes theorem, it is an simple and influential theorem (Alsaleem, 2011) .

Term Frequency-Inverse Document Frequency: a numerical statistic which is designed to reflect the importance of a word to a document in an assortment or corpus (Rajaraman and Ullman, 2012).

Precision: the classifiers capability to classify or categorize the tested document as existence under the valid class as opposed to all classified documents in that class, both valid or invalid (Steffen, 2004).

Recall: the classifier capability to classify the texted document to a class that should be chosen (Steffen, 2004).

F-measure: conjoining the precision and recall measured, in order to get a big image about the performance with taking in mind that the recall and precision have the same significance in measuring the performance (Steffen, 2004).

Multinomial Naïve Bayes: A model have been utilized as an algorithm of classification, it is a model of Naïve Bayes In that concerning with the frequency of features (McCallum and Nigam, 1998).

1.4. Problem Statement

These days there are a massive number of the text files in the storage unit in the computer should be classified. usually the users don't know the files content, like the news that need to classify the articles and text into various classes, every class talk about a particular subject, the classification mission of these files to their related subject is a difficult task, especially when there is a huge number of documents, also when the subjects are convergent, and when the accuracy of classification is essential, so when the users want to find a particular document he/ she will spend much more time to find the wanted subject.

- In this study, the suggested document classification tool that depending on the Multinomial Naïve Bayes Model would be improved, by utilizing three approaches in the first situation, the model would be improved by utilizing the n-gram, in the second way by utilizing the TF-IDF and finally would be improved by integrate the TF-IDF and n-gram. In the three suitcases, the measurements of certain evaluation criteria would be assessed, to estimate the performance of classification tool, so as to obtain the greatest enhanced model.
- Since the previous-mentioned drawbacks or disadvantages in the pre-existing document classification tool, and due to the enthusiasm to increase and improve the performance of the document classification tool, the concept of this study is emanated. The thought emphases on utilizing an enhancement to MNB classification tool, so that to advance the performance of categorizing document that correlated to an exact subjects based on the files content even these files are not branded. In this research, the planned classifier would be applied over an Arabic documents, because of the Arabic documents are containing

more complex matters than the English documents, and these matters have to be taken in mined while the preprocessing matters that have to be implemented before the classifying.

1.5. Motivation

In the text classification discipline, the performance of the classification tool depends on numerous aspects, like the accuracy level and the classification procedure time. Several applications are motivating with time more than accuracy, and the other application concerned in the opposite way, and other applications attracted in both of the features. All of this encourages to design a document classification tool that is able to accomplish a better accuracy to deploy the classifies in an critical classification like the classified texts that are related to the similar field, but must be categorized to sub topics without effect on the performance of the classification time as probable as can. In addition to the rarity of the previous studies that is deal with the Arabic documents; encourage us to appliance the classification process over the Arabic document.

1.6. Objective

Because of the increase in document files that related to the same topic under of more subjects, the necessary of document classification and the management of the files became a must with an effective method which it let the easy to find the wanted files which is related to the wanted subject. Consequently, the automated classification of the files depends on documents of the predefined training has been witnessed a high increase in interest over the previous years. The aim of this study is to apply the technique of classification of the files monitoring and management to have a higher level of organized and high level of retrieval of the flies. This study aims to apply this classification technique for files management to raise the level of organization and retrieval of

files. The task of estimating different suggested enhanced MNB classifier aims to get a good classification tool which able to accomplish the next major targets:

- Maximize the accuracy of classification, especially when the files related to a similar field, the classification would be more complex and the accuracy would be critical.
- Advance the performance of the document classification to be more effective and well-organized to decrease the classification time which spent by the users to find the wanted files that related to a particular subject.

1.7. Methodology

In this study procedure, the Naive Bayes classifier would be applied as an algorithm of classification. The Naive Bayes theorem is the greatest common text classification technique that is dealing with the documents as a bag words and chooses if a particular keyword exists in an exact document or not (Shimodaira, 2014). The classifiers depending on Naïve Bayes have outdone the stronger alternatives, due to the easiness to implement, and the fastest, and the accuracy of it (Rish, 2001; Domingos and Pazzani, 1997). In this scheme, the Multinomial Naïve Bayes (MNB) model have been utilized as an algorithm of classification, In MNB model concerning to the number of keywords in every class of a data set. It is given that the error reduction in excess of the Bernoulli model (McCallum and Nigam, 1998).

In this study, the suggested document classifier that depends on the Multinomial Naïve Bayes Model will be improved, by utilizing three approaches in the first approach, the model will be boosted by utilizing the n-gram and in the second approach by utilizing the TF-IDF and finally will be boosted by integration the TF-IDF and n-gram, and in the three cases, the values of some

performance measurements like the **Recall, precision** and **F-measure** for every model of the improvement will be assessed, so as to measure the performance of the classification then obtaining the greatest boosted model. The GUI of the system will present the three methods in a drop list, then the percentages of **Recall, precision** and **F-measure** of the suggested classifiers will be showed, the perfectness of the organism improved while the percentage of **Recall ,precision** and **F-measure** enlarged. Lastly the methodology of this investigation could be abridged by the subsequent points:

1. An open source data set Arabic will be utilized, it covers a number of classes related to the similar field, each class covers a massive number of documents, the documents in every class will be separated into two sections, the first part of the text documents will be designated as training documents and the second part will be utilized as a test documents.

2. In this classification, the predefined classes may be distinct as (C) such as every $c \in C$, then the MNB model categorizes the tested document (X) to the class that has the top value of the possibility that means the distance among the tested document and the top possibility class is the smallest, the utmost possibility could be measured by the Bayes' rule (see in chapter three).

3. The sets of the words model would be utilized to exemplify the texted document in a modest way. In this model, the texted document is exemplified as a group of string anyway to the sentence grammar or to the place of string in the sentence (Sivic and Zisserman, 2009).

4. While extract the distinct terms for all the documents that exist in all the classes some processing Issues apply to the text document, in order get just the good words, the following points represent such processing Issues:

    - Tokenization: is the process of converting the full text document to a collection of individual components by breaking the text corpus down (Alkafije and Ajam, 2013).

    - Removing stop words: The stop words defined by the words that occurs commonly in the text document, and which are not meaningful alone (Alkafije and Ajam, 2013).

    - Stemming: is used to lessen the number of initial features, by omitting the features which have the same stem and omitting the misspelled. The algorithm that used for stemming is called a stemmer (Ikonomakis,Kotsiantis and Tampakas, 2005).

5. Three suggested classification tools that enhanced the MNB classifiers will be involved, the next points represent it:

    - Enhanced MNB by using N-gram: some of words are given a completely different meaning when it is combined with another word. Text data can be split ever as sequential pairs of keywords which called (bi-gram), or as three sequential keywords which called (tri-gram), or as four or sequential keywords which called (4-gram), using

the N-grams can increase the performance of the classification, in this system, the bi-grams, tri-grams, 4-gram have been implemented and evaluated in order finding the best performance n-gram which can be the MNB.

- Enhanced MNB by using TF-IDF: term frequency–inverse document frequency (TF-IDF) can improve the classification by correcting some of the hypothesis of multinomial data by using the MNB. When a keyword occurs more times the TF-IDF value increases the probability of this keyword (Rennie,Shih ,Teevan and Karger, 2003).

- Enhanced MNB by the merge of n-gram and TF-IDF: in this enhanced classifier, the most effectiveness n-gram which have been implemented and evaluated will be merged with the classifier which has been enhanced by (TF-IDF).

6. There are different methods to measure the performance, in this system the recall, precision, and F-measure will be used in order to measure the performance for every proposed classifier. The related classes of the test documents are known before classifying, and after the classifying, the system checks if the test documents are classified in their related classes or not, based on the results, the recall, precision and F-measure for every class are estimated.

7. The MNB model and the enhanced techniques will be implemented by using the c# language programming because it has different libraries and functions which can support our system in addition to design the GUI by using the c# windows form. In the GUI, the user have to specify only the path of the folder which contains the training and test

documents, then the system will be measure the Recall, precision and F-measure for the test document after the system classified it. The related classes of text documents must be known before classifying in order measuring the recall, precision and F-measure.

Chapter Two

Literature Review

2.1. Overview

According to Wijewickrema and Gamage ,(2013) due to the high growing of the digital purport and the small efficiency of the manual text classification and the semi-automatic text classification in officialdoms, the automatic classification of text grew excessive significance, additionally to that the manual and the semi-automatic classification need a lot of time and effort moreover, the manual and semi-automatic classification may be manufactured misclassification in order of the ambiguity in these category of classification.

According to Goller,Löning , Willand Wolff,(2000) there are two levels or stages in the document classification especially the automatic one; in the first stage is the learning stage while the second is the subsequent classification. In the learning stage, the users have to choose the subject as they want or according to the system need, and the chosen of a documents that related to the wanted subjects. Most of the operations of the automatic classification for the document are require a counter examples document to all subjects, the counter examples documents should not refer to such subject. the documents can be related to more of one subject, it could be had various subject in the hierarchy mode, but during the classification stage it's a must and responsible on the classifier to present the rake related to every document for every subject, in the classification stage it should have a high performance level, due to the number of documents which are have to be classified.

2.2. Related Work

2.2.1.Types of Document Classification

According to Padhye ,(2006) there were two forms of automatic text classification that consist of the supervised document classification approaches and the non-supervised document classification approaches.

Supervised document classification:

In supervised classification, there is a training set and a defined by collection of set classes (Patrick, 2009).The supervised methods most of times used in the standards of multi label classification, however, the number of labeled document is very small in comparison with the non-labeled documents, and because of that the supervised classification has limitations in the multi label text classifiers because the supervised algorithms need the labeled training documents (Dharmadhikari ,Ingle and Kulkarni ,2012).

Unsupervised document classification:

The addresses of unsupervised document classification is the problem of defined classes to documents which related to these classes by its content without using predefined classes or training set, the efficient of this to assign every document to a class or set of classes, a lot of real- world application depends on the unsupervised classification, like using it in the classify the e-mails to spam or not by its content, or using it to reduce the time of processing the queries by providing more pertinent results, in addition to filtering the news by its content to several subjects (Patrick, 2009).

Semi-supervised document classification:

In Semi-supervised classification, there is a training set and a defined by collection of set classes (Patrick, 2009).Also the semi supervised achieves more performance when the number of unlabeled document more the number of labeled documents, the main aim of using the semi-supervised classification reduce the classification errors when using a mix of labeled and non-labeled documents (Dharmadhikari, Ingle and Kulkarni , 2012).

2.2.2. The Disadvantages and Challenges of Document Classification

The text classification that is a term-based have a certain disadvantages: they have need of a stage of a linguistic preprocessing which in smallest, identify the terms, also they have a problematic in the data (sparse data), even with the massive training corpora, there is a great number of terms that standing in the testing data and inattentive in the training data, and this problematic get up when comprehensive vocabulary field of the book advertising domain. The sparse data can be reduced by increasing the linguistic preprocessing which very cost, additionally, there is the problematic of the spelling mistakes which can't be abridged (Steffen, 2004).

Affording to Power, Chen, Kuppusamy and Subramanian,(2010) the text classification influenced by two core features which make two major barriers can be summarized by:

- The extraction process: in all of the algorithms of document classification, the procedure of extracting keywords acting a vital portion in the percentage of the classification precision.
- The vagueness of topics: the procedure of categorizing the text with vagueness subjects is not an easy task. The vagueness subjects could be having dissimilar meaning and vagueness subject with its associated keywords may be going on another subject.

2.2.3. Document Classification Applications

The Automated classification of document is an essential step in the task of text mining, particularly while the quick growth of online documents amount that is written in Arabic language. The main objective of text classification allocates the document to a predefined category in an automatic way depending on the linguistic features. The document classification process has many influential applications, detect the spam e-mail, filtering of web page content, and routing the automatic message. In their study they presented the experimental results that classified document to seven Arabic classes, the classifier achieved by using statistical methodology (Al-Harbi, Almuhareb, Al-Thubaity, Khorsheed and Al-Rajeh, 2008).

El kourdi, Bensaid and Rachidi, (2004) accompanied a study that discussed the automatic classification of the web documents that is written in Arabic language. They presented that the classification is very essential for affording the guide to the search operations that has been used by many portals of webs and search engines to overcome the huge increasing in the amount of the documents that appears on the web. In this paper they used the Naive Bayes (NB) as a "statistical machine learning algorithm" for classifying the Arabic web documents that non-vocalized to a specific predefined category, before classification the words that is inside documents has been transformed to the roots, they used the experiments of cross validation in order to evaluate the NB classifier, a data set that have been used contains 300 document per class, the results show that, the average of accuracy of all classes around of 68.78. Furthermore, the most effectiveness categorization by classify using experiments of cross validation rises up to 92.8%.

Duwairi, (2007) conducted a study about classification text documents for documents which written in Arabic language, it has been the distance-based classifier. Distance -based classifier

depending on collecting the information about the classes in the learning phase, every class described by a set of keywords which present vectors, these vectors are extracted from predefined documents. The documents which classified are pre-processed by omitting the stop words, symbols, etc. Then the remaining keywords are stemmed and stored. The pre-processing issues are very important because the Arabic language a very richness, and the pre-processing issues reduce the number of features, in addition of that the stemmer is filtering the features, the results presented by the Recall, Precision, F-measure.

Al-Shalabi and Obeidat ,(2008) conducted a study about classifying the Arabic documents, they proposed two classifiers that based on the KNN algorithm, at the first classifier they used the n-gram in document indexing, and in the second, they used the single term indexing technique which is the traditional method(bag of words). Results showed that using n-gram in the classifier achieve more performance than using the single terms.

According to Li and Jain , (1998) examine engines were realized to drop the effort, power and time of the employers and make the data recovery easier, then to rise the utilization of the World Wide Web. Utmost of commercial examine engines just as example InfoSeek, Yahoo, HotBot, etc., based on the document classification in all its processes just like document recovery, document direction-finding, document cataloguing and filtering organism. The classification of document technique depends on the predefined groups of labels examples that referred to two or more classes, then categorizing the new text to the class which has the maximum resemblance. The document would be labeled for every class as relevant or non-relevant file, when the employer examine about the files, the relevant document will be existing. The text classification met some challenges just such as it is too hard to imprisonment the improvement semantic of the natural

languages from the keywords of the text. There are dissimilar forms of classifier techniques could be applied in the document classification, just as Naive Bayes classifier, decision trees and nearest neighbor.

Referring to Lam, Ruiz and Srinivasan, (1999) they have been improved a modern process in the automatic document classification also they have been utilized the automatic classification in the document retrieval. This modern process of classification is derived from a learning sample called the "instance-based learning" also depends on the new document retrieval technique called the "retrieval feedback. The performance of the modern process of classification is too high due to use of set of a two real-world text exported from the MEDLINE database. Moreover they established and achieved a high performance for the document retrieval outputs by utilizing of the manual classification of documents with equal performance level in the automatic one.

Nigam, McCallum, Thrun and Mitchell, (2000) conducted a study about the document classification and he set up that the number of categorized text in compare with the uncategorized text is too small. And the manual labeled document to the unlabeled document, it is a hard work intensive effort. The automatic classification is based on the training texts to obtain the keywords and establish the rules (classes).Consequently, they utilizing an algorithm depends on both of the Expectation-Maximization (EM) and the Naive Bayes classifier that work to train a classifier dependent to the document that are even now labeled. At that time probabilistically label the portion of unlabeled documents, then utilizing the innovative assembly of labeled texts to probabilistically label to another part, and repeat this process up until labeled all the texts.

According to Moulinier and Jackson, (2002) conducted a study about the text or the document classification and got out that the maximum common text classification problematic is filtering the

spam e-mail letters to two classify spam or non-spam. The essayist finds the solution of this problem by creation of two classes. First class consists of all the term of expected terms of the spam e-mail while the other class consists of the terms which expected to be in the non-spam emails. By applying the Bayes Statement with some document classification models, evenhanded like the Bernoulli document model.

Du, Safavi-Naini and Susilo, (2003) propose a new technique on the web filtering depending on the text classification. The web filtering objective is denied the accessing of the non-useful web pages. And by using the text classification, they make two classes first class contains samples of non-useful web pages which must be blocked. The system should have the ability to prevent access the web pages which similar to the forbidden class and allowing the access of the web pages which dissimilar to the forbidden class.

Frank and Bouckaert, (2006) accompanied a research about applied the naive Bayes classification tool with the multinomial model in document classification. The blend of them generates the Multinomial naive Bayes (MNB) which is an actually well-known technique in classification document that have a high performance in forecasting furthermore .It has a computational efficiency. As well of that this performance may be improved by utilizing the transforming the data.

Nanas,Domingue , Watt , and Motta, (2001) conducted a research about the text classification in addition to information filtering and the information retrieval, and they found that because the unexpected increment in the availability of digital documents, particularly because of the World Wide Web (WWW) has imparted a new issue in access the information in sharp concentrate. The text classification in addition to information filtering and the information retrieval are using to deal

with this issue which called by the researchers the "Information Overload". Furthermore, the information filtering (IF) can be considered as a specific case of text classification where every user corresponds to two parts, pertinent and not pertinent documents to the particular user.

Ramdass and Seshasai,(2009) conducted a study about document classification for newspaper articles and they found there is several scenarios in our real-lives are desired the process of classifying the different documents to different classes, and that's can be achieved by using the automatic classification, one of these scenarios the newspaper articles, the newspaper articles have to be classified to several classes such as the 'sports' or 'news', etc. They applied many algorithms to achieve the best accuracy classification, finally they used the Natural Language Toolkit (NLTK) package to implement their classifier in python, and NLTK involved a naive Bayes classifier.

Pop, (2007) conducted a study about using the Naive Bayes classifier for the document classification, he shows how the Naive Bayes improves the accuracy of the web mining process, also how the accuracy is a significant part in the real-world applications such as the email spamming, the mining of log files for the system management computing, search queries, semantic web in machine learning, in addition to a lot of fields of web mining.

Unstructured information sources have drawn recently more attention mostly because of a rising number of electronic documents accessible through different sources like e-mails, huge digital libraries, local networks, but most significantly via WWW. Researchers from many different fields try to use their own techniques to automatically organize these data collections and enable users to access data in some informed way, i.e. users know how to navigate through these data sources and understand the organizational structure without a priori organizing those data. One of the

techniques usually employed is a classification, which enables automatic routing of a particular document into some pre-specified sub-collection (Alkafije and Ajam, 2013).

This paper suggests some challenging research problems that can be found in the area of text classification and then concentrates on the feature set reduction methodology as one of the key topics. Different existing methods for feature set reduction have been developed in the areas of information retrieval and further in text classification. Although these techniques have been independently developed over many years, they have a strong relationship with methods from pattern recognition area where the methodology seems to have reached more complex theoretical results. The paper therefore aims to put special text-oriented techniques into the context and terminology developed in pattern recognition. Experimental results compare different feature set reduction methods and illustrate how the use of some well-known pattern recognition methods can improve classification accuracy (Anagnostopoulos, Broder and Punera, 2006).

According to Godbole , Harpale , Sarawagi and Chakrabarti, (2004) they conducted a study about classifying documents by using the terms label and the interactive supervision of document, they discussed that , one of the most crucial issue of real-life document classification applications, is the high depending on the human expertise, these real life applications not addressed enough by the learners of "batch-supervised high accuracy", there is one way to supervised the standard text document classifiers, which is determining labels to all the documents, this prevent the humans from taking the advantages of phrases and words in the context of the text.

They proposed a HI-Class that is a package of labeling, which is an exploratory and interactive package; this approach gathers the opinions of users on the representations and choices of features, in addition to the all document labels. They suggested at first, starting with basically an unlabeled

text document, with little cognitive labor to constitute a labeled collection that applying the standard classifieds in a good way, they also presented an overview of HI-Class. The first layer remains the entities of main data in addition to the fundamental processing units, there is also a small amount of labeled documents, on otherwise a huge amount of unlabeled documents There is a small pool of labeled documents, the documents are converted to feature vectors in the feature extraction, their proposed system able to access and store by using a specific classifiers (Godbole, Harpale, Sarawagi and Chakrabarti, 2004).

Anagnostopoulos ,Broder and Punera, (2006) presented that the Frequency -inverse document frequency (TF-IDF) which is another alternative that characterizing text documents in many cases. It's understood to be the weighted term frequency that is especially useful if stop words have not yet been removed from text corpora. The TF- IDF is an approach that assumes the importance of a word is inversely proportional to how often it occurs across all documents. Although TF-IDF is most commonly used in ranking the documents by relevance for different task mining of text, like a page that ranks through search engines, it can also be applied to text classification via naive Bayes.

Han and Karypis,(2000) conducted a study about a simple algorithm for "linear-time centroid-based document classification", they present that, the most significant of this system is arising from the function which it implements in order to calculate the similarity between a centroid vectors of the class in the text document. In addition, to find the correlations between the terms that existed in the documents, they also discussed that there is many ways can be improved the such classifier, such as generating a new form that able to deal with the multimodal classes, the multi-modality can be supported in an easy way by using the algorithm of clustering which defined

by divided the documents of every class into many subclasses, and improving the classifier performance by using techniques which cares about the significance of various terms in a supervised setting.

Alkafije and Ajam, (2013) discussed that to be able to classify documents, one must find a way how to reasonably simply represent documents in a way that this representation preserves as much of the original information as possible and also is simple enough from a computational point of view. Different ways of representing documents that reflect the different needs of their users have been proposed. The simplest method called bag of words used in the vast majority of current applications is based on the application of basic terms (either all of them or a subset like nouns). It is also used in this paper. Many other representations have been found, which behave better for some special purposes. For example, conceptual features (represent meaning of the original documents), contextual features (contain contextual information of terms, e.g. bigrams, trigrams, or more sophisticated noun-bigrams, mechanically extracted features (extracted from documents without using any knowledge about its content or language structure, possibly based even on a compressed version of the original document.

According to Anagnostopoulos,Broder and Punera, (2006) being an eager learner, naive Bayes classifiers are known to be relatively fast in classifying instances that are new. These eager learners learn algorithms which learn a model from a training dataset when the data is available. Once the model is learned, the training data does not have to be re-evaluated for one to predict newly. Eager learner's computationally most expensive step is the model building step, whereas the classification of new instances is relatively fast. Lazy learners, however, memorize and re-evaluate the training dataset for predicting the class label of new instances. The advantage of lazy learning

is that the model building (training) phase is relatively fast. On the other hand, the actual prediction is typically slower compared to eager learners due to the re-evaluation the data of the training. The great disadvantage for lazy learners involves training data has to be retained, which can also be expensive storage space as it requires a lot of space. A good example is lazy learner would be a k-nearest neighbor algorithm: Every time a new instance is encountered, the algorithm would evaluate the k-nearest neighbors in order to decide upon a class label for the new instance, e.g., via the majority rule (i.e., the assignment of the class label that occurs most frequently amongst the k-nearest neighbors).

Ting, Ip and Tsang, (2011) they discussed that among many applications, the simplicity in both the classifying and training procedures, increasing the utilized of the "Naive Bayes classifier". Many researchers proved that the "Naive Bayes classifier" is effective enough to utilize in many real-life domains. In this study, they proposed to employ the "Naïve Bayes" as an algorithm in a document classifier, and evaluated it by comparing the effectiveness of it with other classifiers, (such as SVM), they depicted the structure of methodology in their proposed classifier starting with the preprocessing and finishing with the evaluation .

Denoyer , Zaragoza and Gallinari , (2001) they talk over a fresh version of "generative models" for "information retrieval", dependent on "probabilistic sequence models", they encouraged such models with some increasing in the difficulty of missions and textual data, they offered how classical notion of document retrieval could be extended, then they advanced a HMM carrying out for accomplishing the document classification and ranking, and discussed numerous potential variations that can be an advantage to prove the accuracy of their suppositions. Moreover, they presented the suggested models that popularized the traditional "multinomial Naive Bayes"

models, with bearing in mind the existence of non-pertinent passages in the midst of pertinent passages. Then they estimate the model based on the Reuters data set and used the MNB model as a baseline in the comparison.

Wijewickrema and Gamage, (2013) they suggested a study which worries with the automatic system that able to categorize the documents via declining the vocabulary opacity, their suggested system is a lengthy to a previous study that has been technologically advanced a "semi-automatic system" for document classification. subsequently in this study they attempted to advance the document classification that depend on a hybrid of a TF-IDF depend semi-automatic that they termed it by HTCS, and to drop the vocabulary ambiguities, they utilized a domain-ontology, although HTCS has given better results than the manual technique; , but they founded that , this technique has a lot of challenges, due to the nature of semi-automatic process, they discussed that even the manual intervention have been removed , but the final classification until now depends on the decisions of humans, on the other hand the fully automatic method saves the time and saves the labor of a particular mission, they presented the key steps of the of the methodology that have been suggested to fully automatic classification system .

Liu ,(2008) he suggested a research about the document Classification with the Word sense, in this research he utilized the WSD in the procedure of the scheme, with the intention of generate new geographies to represent the documents for the mission of classification, the WSD assisted them to distinguish a particular group of the selected mysterious words by word sense clustering, all of that to rise the performance of the text classification and he defines the design of WSD for document classification experiment, in this research he suggested two assumptions, the first assumption; taking in mind the effectiveness of WSD to distinguish word senses which can find a

solution to the "polysemy dilemma", utilizing the word sense as features able to growth the document classification, the second assumption; it is not significant to disambiguate each word in the document to usage its senses as terms, that is since the WSD is expensive and not very exact, in addition of that the massive amount of word sense disambiguation fit to make the system's risk more hazard. hence they suggested that, the finest way is to describe a group of words which are distinctive as ambiguous words, he presented a sample that how the "java" word that occurs in more one classes ant not a stop words carefully chosen with the ambiguous words, due to the training document, the words which are occurs habitually in the classes and that is able to rise the confuse in the classification should be disambiguated.

Chapter Three

Methodology and Proposed Models

3.1. Design Approach

In this design approach which used the Arabic data set which is contain six classes, each class contains 250 documents. At first divided the documents into training documents that each class contains 160 document and the other 90 documents as test documents, in order to implement the three proposed classifier. All the training documents will be processed by implement a number of processes. At first we tokenized all the training documents after that all the predefined stop words

will be removed, and finally an Arabic stemmer predesigned and evaluated will be implemented in order to reduce the number of features for each document. After the preprocessing implemented, each training document will be have a vector, and each class also will be.

In the test documents part, the test documents which have to be classified for a class will be processed by implementing number of processes just like the process that implemented over the training document and after generated the test document vector, the three proposed classifier will be implemented. At first the N-gram then the MNB will be implement, Then another proposed classifier will be implemented which state on applying the TF-IDF then the MNB. Finally the last proposed classifier state on applying the N-gram then TF-IDF and MNB. After classifying a test document using three classifiers, three results will be obtained. The obtained results will be evaluated in order to find a good classifier, by comparing the classified classes that the classifiers choose will be compared with the real class of the test document with included in the data set before dividing it to training document and test document.

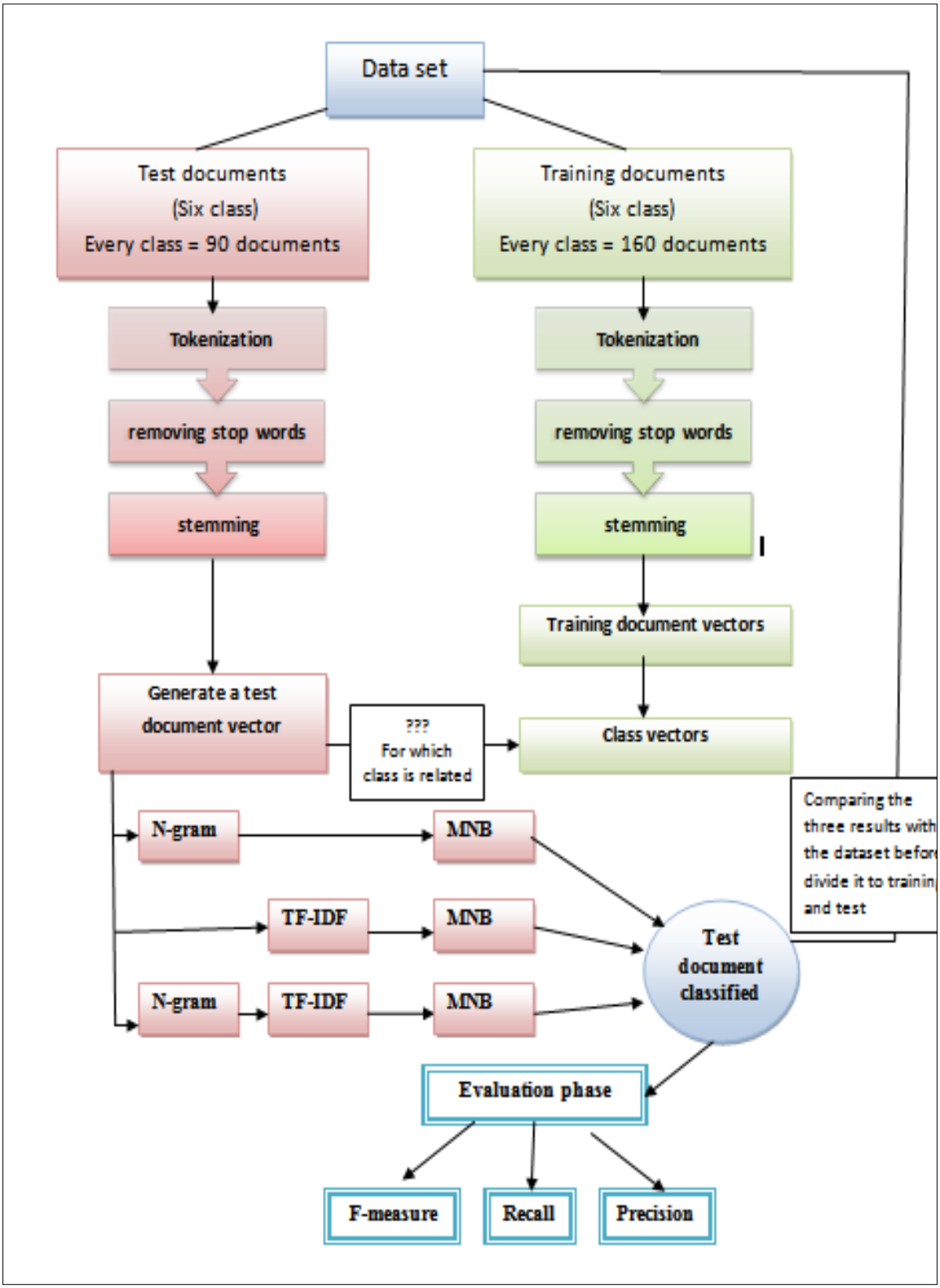


Figure 3-1: the classifier model building

3.2. Data Set

The data set contains1500 Arabic texts have been utilized; the data set which has been utilized is available on the new data set cited at (Saad and Ashour, 2010) and published on http://en.osdn.jp/projects/sfnet_ar-text-mining/downloads/Arabic-Corpora/cnn-arabic-utf8.7z/ each class contains 250 documents, these documents are disseminated as training and test documents as shown in table (3-1), where the classified documents have been utilized as test documents, so as to measure the performance of the classifier in every case of the enhancement which would be carry out in this research.

The Arabic data set that has been utilized covers six classes, each class contains 250 documents. At first divided the documents into training documents that each class contains 160 document and the other 90 documents as test documents, in order to implement the three proposed classifier.

Table 3-1: the data set categories

Category name	**Number of documents**
Business	Training = 160, test= 90
Entertainment	Training = 160, test= 90
Middle east	Training = 160, test= 90
Scitech	Training = 160, test= 90
Sport	Training = 160, test= 90
World	Training = 160, test= 90

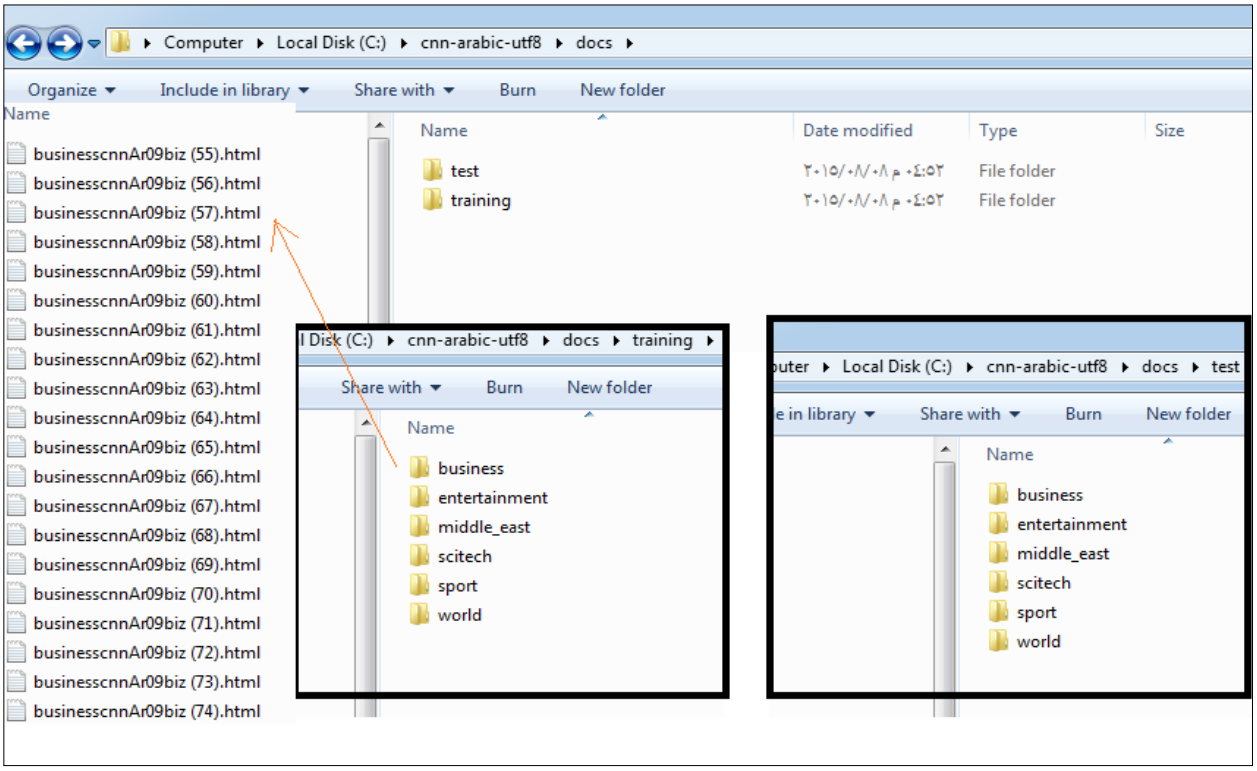


Figure 3-2: the data set which used in expirements

3.3. Overview of Arabic Language

Arabic language is a widespread between 250 million Arabs, additionally of that, about more billions of Muslims is able to realize this language, since it is the language of the Holy Koran (the holy book for Muslims). Arabic language contains 28 characters that are in the mention in the following figure (3-3). furthermore, the Arabic language is written from right to left , the letters of Arabic may be have several styles based on the position of letters in the word, and if it joined the neighbor letters or not, also there is the diacritics which distinct by the signals that be real under or above the letters to give the letter stronger pronunciation or to read it as the short vowel (vowel letters are three letters "اوي "), the Arabic shada , Arabic dama, Arabic fathah are some examples of diacritics in Arabic language, in the computer representation, the letter with a diacritic, is

considered as two characters: one represents the letter itself and another represents the diacritic. Most of the time the diacritic are omitted while writing the Arabic documents (Duwairi, 2006).

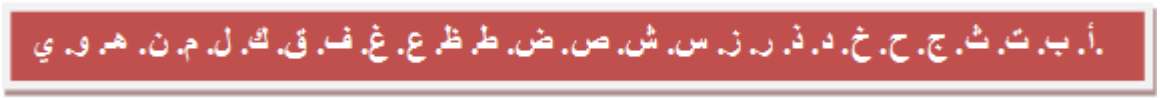


Figure 3-3: the charactersin Aarabic language

3.4. Bag of Words Model

The bags of words model is utilized to symbolize the document in a humble technique, it is applied in the information intermission and in the normal language handing out. In this model, the text or document is characterized as a set of string irrespective to the grammar of the sentence or the place of it in the sentence. The significance of bag of words model in the text classification stand up from saving the frequency of happening in each term in the document, By the bags of words model, so we can epitomize a feature vector for every distinct word in the document which are existed in the training set classes, every vector represents the whole words associated with their frequency (Sivic and Zisserman, 2009).

3.5. Preprocessing Issues

In Arabic document pre-processing, the digits, special symbols and strings of characters, most of times exist in the text document. In the document pre-processing, only the features which that depict the document are extracted, so the formatting tags, punctuation marks, English characters, prepositions, conjunctions, and stop words most of the times are removed. In English text document any word consisting of two letters is removed, but in Arabic language, there is

meaningful words consisting just two letters, so if these words do not stop words will be not removed figure(3-4) below describes the document preprocessing in this classifier.

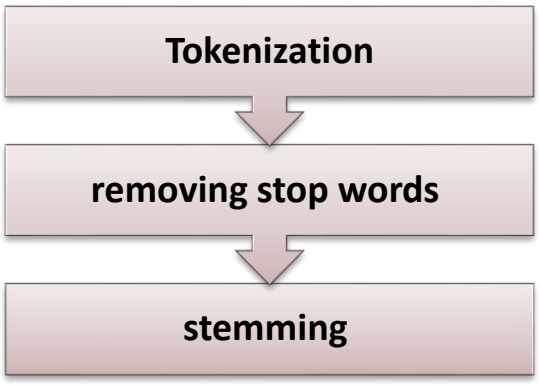


Figure 3-4: document preprocessing issues

3.5.1 Tokenization

Tokenization well-defined by the procedure of transformed the full text or document to a group of singular components by breaking the text corpus down. Most of times the tokenization procedure united with another procedure such as removing stop words and punctuation characters, stemming, Unicode conversion, removing diacritics, numbers, etc., (Alkafije and Ajam, 2013). The figure (3-5) is an example of the tokenization procedure.

Smoking, especially smoking a pipe, is a factor of cancer incidence gums, tongue, mouth surface factors															
("التدخين لاسيما تدخين الغليون،هوعامل من عوامل الإصابة بسرطان اللثة واللسان وسطح الفم")															
الفم	سطح	و	اللسان	و	اللثة	بسرطان	الإصابة	عوامل	من	عامل	هو	الغليون	تدخين	سيما	لا	التدخين

Figure 3-5: example of tokenization

3.5.2 Stop Words

Stop words well-defined by the words that happens frequently in the text or document, and which are not significant alone, the mission of it is attaching the meaningful words to products a sentence with a full meaning (Alkafije and Ajam, 2013).The figure below shows an example of the stop words in the Arabic text or documents that have been removed from the text or the documents (Alkafije and Ajam, 2013).

اللى	او	ذلك	فقد
اللي	اول	ريتويت	فقط
الى	اى	زى	فكان
الي	اي	زي	فهو
اليه	ايه	ستكون	فى
اليها	ب	صار	في
اليوم	بإن	ضحى	فيه
امسى	بات	ضد	فيها
ان	بد	ضمن	فيهم
انا	بدلا	ظل	قال
انت	بعد	ع	قبل
انتا	بعض	على	قد
انتي	بل	علي	كأن

Figure 3-6: example of some stop words in the Arabic document

3.5.3 Stemming

One of the most significant matters in the pre-processing is the stemming, where the stemming is utilized to reduce the number of initial features, by removed this features which have the same stem also by omitted the misspelled. The algorithm that is utilized for the stemming is called a stemmer (Ikonomakis, Kotsiantis and Tampakas ,2005).

According to Mustafa, (2012) all of the stemming algorithm or it's set of rules that realized over the Arabic language must base on the assumption which states that, all the words in Arabic

language have the same root are associated semantically. This supposition comes from the environment of the Arabic language that definite as the derivative language. The key objective of the stemming process is to give the employer more ability to retrieve the related terms morphologically which may have the same semantic, see table (3-2).

Table 3-2: words derived from the ground root (جمع "JM3") (Mustafa, 2012)

Word	meaning	word	meaning
جمع	crowd	جمعية	association
جماعة	group	جامعة	university
جماع	mating	مجتمع	society
مجموع	sum	اجتماع	meeting
جامع	mosque	جمعة	Friday

An ordinary light stemming approach was advanced in the stemmer which has been implemented in this research. This process is well thought-out with only a minor number of grammatical suffixes and prefixes, which is appear in the normal text or documents more than the others. The figure (3-7) below presents a list of suffixes and prefixes:

```
pref = { "أ", "ال", "بال", "ت", "ست", "فت", "في", "ل", "لل", "و", "وال", "وبال", "ولل", "ون", "وي", "ي" };
suff = { "ا", "ت", "كم", "نا", "ه", "ها", "هم", "هما", "وا", };
```

Figure 3-7: subset of grammatical suffixes and prefixes

In the advanced scheme, the applied of Suleiman Mustafa stemmer have been utilized, this stemmer depending on applying a light stemming algorithms, (Mustafa, 2012) accompanied a research about a stemmer that dealing with the Arabic document, and he deliberated the benefits

of light stemming which realized on the Arabic text document, and he obtainable the light stemming algorithm that has been evolved on the logical source over the real happening of prefixes and suffixes in the documents. As well, he associated the efficiency of this stemming algorithm with other intensive stemming algorithms which is very worries with the most grammatical suffixes and prefixes .The results of his research showed that just a little amount of the prefixes and suffixes can affect the correctness of stems produced. To conclude, he concluded that the performance of light stemming has outperformed the performance of the heavy stemming based on the measures of over-stemming and under-stemming, the figure (3-8)show these measures (Mustafa, 2012).

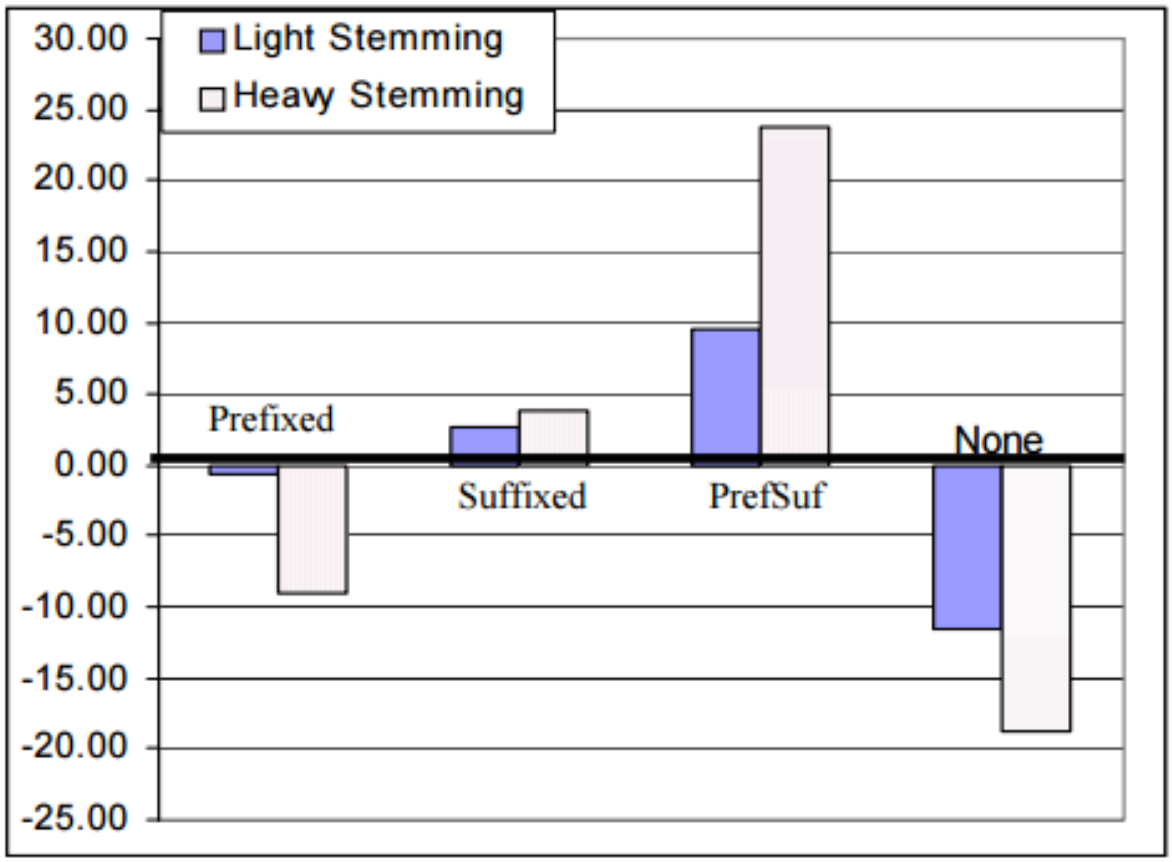


Figure 3-8: Viewing the results of light stemming and heavy stemming in terms of over-stemming and under-stemming percentages

3.6. Multinomial Naïve Bayes and Document Classification

The Naïve Bayes is a non-difficult probabilistic classification tool which is based on applying the theorem of Bayes, it is easy and powerful theorem (Alsaleem, 2011), the features in the documents supposed as commonly independent from this characteristic the adjective "naïve" came. The classifiers depends on the Naïve Bayes outperform the stronger alternatives, due to it is easiness to implement, and the fastest, and the accuracy of it (Rish, 2001 ;Domingos and Pazzani, 1997). In this system, the MNB model has been utilized as an algorithm in the classification, In MNB

model on the subject of to the number of keywords in the classes of a data set. It is providing the error drop over the Bernoulli model (McCallum and Nigam, 1998).The term (feature) frequency is a technique to classify the test or documents, the term frequency ($tf(t,d)$) is well-defined by the number of happening each term (t) in such document(d), it also sometimes clear by the row term frequency. In the MNB model, the feature vectors represent how many a specific feature occurs in the document, every feature vector (X_i) represents every distinct term (feature) with the frequency of occurring it in a specific document.

Now the strategy of classifying a test document for a specific document using the multinomial document classifier will be discussed, let the predefined classes will be defined as (C) such as every $c \in C$, then the multinomial naïve Bayes model classify the test document (X) to the class which has the highest value of probability that means the distance between the test document and the highest probability class is the smallest, the highest probability can measures by the following Bayes' rule which presented in equation (1) .

$$P(c\backslash X) = \frac{P(c)P(X\backslash c)}{P(X)} \qquad Bays'rule\ (to\ obtain\ the\ highest\ probability \quad (1)$$ (Alsaleem, 2011)

The prior class $P(c)$ can be measured by the equation (2);

$$P(c) = \frac{the\ number\ of\ document\ that\ belongs\ to\ class\ c}{the\ number\ of\ the\ whole\ documents\ in\ the\ whole\ classes} \qquad (2)$$(Alsaleem, 2011)

The probability of obtaining the test document (X) in class c measured by the equation (3) ;

$$P(X\backslash c) = \prod_{i=1}^{m} P\ (X_i\ \backslash c) =\ P\ (X_1\ \backslash c) *\ P\ (X_2\ \backslash c) * \dots\dots\dots P\ (X_m \backslash c) \quad (3)$$ (Alsaleem, 2011)

$P\ (X_i\ \backslash c)$ Is the probability of every feature vector (X_i) that occurred in test document and can be measured by the equation (4) (Alsaleem, 2011);

$$P\ (X_i\ \backslash c) =\ \frac{\sum(tf(X_i, d\ \in\ c)) +\ \alpha}{\sum N_{d\in c} +\ \alpha\ .V} \quad (4)$$

- $(\mathrm{tf}(\mathrm{X_i}, \mathrm{d}\ \in\ \mathrm{c}))$ Is the sum of raw term frequencies of feature xi from all the training documents in the class (c).
- $\sum \mathrm{N_{d\in c}}$ Is the summation of all the frequencies of terms in the training documents that related to class (c).
- α Is an additive smoothing parameter (in Laplace smoothing is equal one).
- V represents the size of all the distinct words in all the documents in all the classes

3.7. Term Frequency - Inverse Document Frequency (TF-IDF)

Term frequency–inverse document frequency (TF-IDF) can advance the classification by modifying some of the hypothesis of multinomial data by utilizing the MNB. The (TF-IDF) is an method for symbolizing documents. It is well-thought-out as a technique that gives the term frequency a weight, The (TF-IDF) method rises probability of term when it happens more than one time (Rennie, Shih , Teevan and Karger, 2003). The (TF-IDF) could be measured by the equation (5).

$$TF - IDF = tf_i(t,d) * idf(t) \quad (5)$$ (Rajaraman and Ullman, 2012).

$tf_i(t, d)$ is the count of term t in document d .

idf(t) is the inverse document frequency that can be measured by the following equation;

$$idf(t) = log \frac{\text{The total number of documents}}{\text{The number of documents that contain the term t.}} \quad (6)$$ (Rajaraman and Ullman, 2012)

3.8. N-Grams

In the n-gram, an arrangement of n strings could be defined as a token in the default case, the uni-gram is utilized which is consist just one string with letters, symbols, etc., and it is the simplest form of n-gram, but when a sequential pairs of strings combined it is called (bi-gram), and the three sequential strings it is called (tri-gram), and the four sequential keywords it is called (4-gram), by using the n-grams the performance of the classification can be improved(Zečević, 2011), the figure(3-9)below is an example of different n-gram.

Smoking, especially smoking a pipe, is a factor of cancer incidence gums, tongue, mouth factors							
("التدخين لا سيما تدخين الغليون،هوعامل من عوامل الإصابة بسرطان اللثة واللسان والفم")							
التدخين لا	سيما تدخين	الغليون هو	عامل من	عوامل الاصابة	بسرطان اللثة	و اللسان	و الفم

Figure 3-9: Example of implement the bi-gram

3.9. The Proposed Classifiers

In the enhanced system, limited methods have been utilized to raise the performance of the MNB classifier, then the enhanced MNB classifiers have been evaluated in order to obtain the best enhancement technique, in the following points the enhanced MNB classifiers which have been realized and assessed in this research.

- Enhanced MNB by utilizing the N-gram: some of words are given a completely dissimilar meaning when it is combined with another word. Text data can be split ever as sequential pairs of keywords which called (bi-gram), or as three sequential keywords which called (tri-gram), or as four or sequential keywords which called (4-gram), utilizing the N-grams can growth the performance of the classification, in this system, the bi-grams, tri-grams, 4-gram have been implemented and evaluated in order finding the best performance n-gram which can be the MNB.

- Enhanced MNB by using TF-IDF: term frequency–inverse document frequency (TF-IDF) can advance the classification by adjusting some of the hypothesis of multinomial data by using the MNB. When a keyword occurs more times the TF-IDF value increases the probability of this keyword (Rennie, Shih, Teevan and Karger, 2003).

- Enhanced MNB by the combine of n-gram and TF-IDF: in this enhanced classifier, the most effectiveness n-gram which have been implemented and evaluated will be merged with the classifier which has been enhanced by (TF-IDF).

3.10. The Evaluation Criteria

There are dissimilar approaches to evaluate the performance of classifier like the recall, precision and F-measure. When a text is categorized, the result of test document categorizing to such class will be either a true positive (TP), false positive (FP), true negative (TN), or false negative (FN), see following points;

- True positive (TP): when the test document classified in the true class.
- False positive (FP): when the test document classified in the incorrect class.
- False negative (FN): when the document should be related to a specific class and it is not marked to such class.
- True negative (TN); when document should not be related to a specific class and it is not marked to such class.

In the advanced system to measure the performance, the connected class of the test documents are known before categorizing, and after the categorizing, the system checks if the test documents are categorized in their related classes or not, based on these results the TP, FP, FN, TN values can be measured, then the recall and precision for every class in every proposed classifier are estimated by using the following equation;

- Recall R_i is defined by the ability of the classifier to classifying a test document to the class which should be chosen (Steffen, 2004).

$$R_i = \frac{TP_i}{TP_i + FN_i} * 100\% \qquad (7)$$ (Steffen, 2004).

- Precision ρ_i describes the ability of the classifiers to classify the test document as being under the valid class as opposed to all documents classify in that class, both valid and invalid (Steffen, 2004).

$$P_i = \frac{TP_i}{TP_i + FP_i} * 100\% \qquad (8)$$ (Steffen, 2004).

- Combining the precision and recall measured, in order getting a big picture of performance with considering that the recall and precision have the same importance in measuring the performance, the following equation present how to estimated it (Steffen, 2004).

$$F = 2 * \frac{P_i * R_i}{P_i + R_i} * 100\% \qquad (9)$$ (Steffen, 2004).

3.11. The Implementation

The MNB model and the enhanced procedures have been implemented by utilizing the c# language programming for the reason that it has different libraries and functions that can support this system additionally to design the GUI by using the c# windows form.

3.11.1. The Graphical User Interface

In the suggested GUI, the employer have to lay down only the track of the folder which contains the training and test documents, then the system will be measured the recall, precision and F-measure for the test document after the system categorized it. The related classes of test documents must be recognized before categorizing in order measuring the recall, precision and F-measure. Figure (3-10) presents the initial GUI of the system. In the text box folder, the path of the training and test document will be added, and the form of the drop list, the suggested enhanced classifier can be chosen, the button start, will be disabled after the user pressed it, until the system finished, so the system included multithreads, finally in the GUI there is a progress bar represented the status of the classifying process.

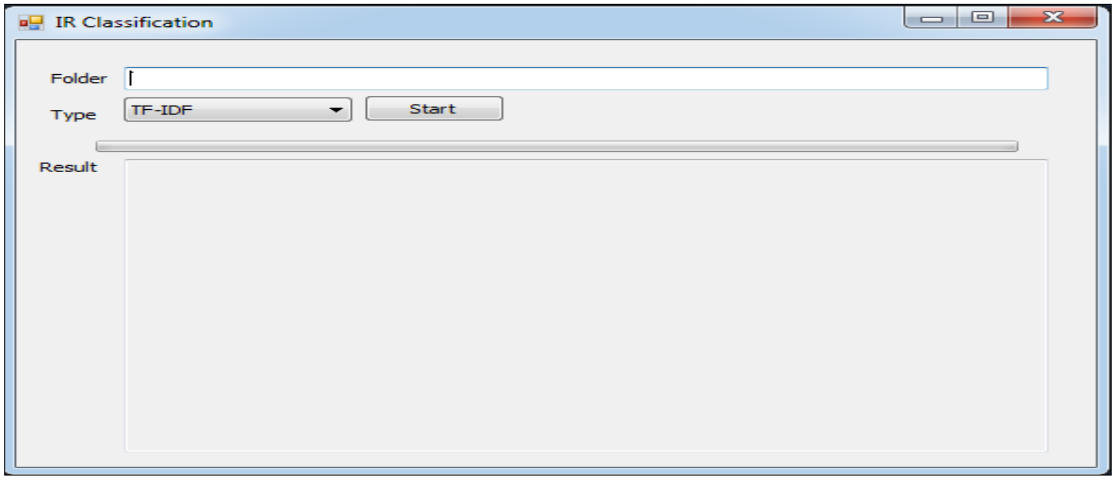


Figure 3-10: The initial GUI of the system

Chapter Four

The Results

In this chapter the recall, precision and F-measure will be measure for every class has been assessed for every suggested classifier. Firstly, the measurements of the recall, precision and F-measure for the (2, 3, 4) gram have been estimated so as to select the most effectiveness N-gram method.

4.1. Multinomial Naïve Bayes without Enhancements

As discussed in the design approach, an Arabic data set that contains six classes has been used. The documents have been divided into training documents and test documents, in order to implement several proposed classifiers, in this case a MNB classifier without enhancing has been be implemented. The obtained results will be evaluated in order to find a good classifier, by comparing the classified classes that the classifiers choose with the real class of the test document that included in the data set before dividing it to training document and test document. The Table (4-1) presented the Recall, precision and F-measure for the MNB classifier without enhance, the effectiveness of classifier increased while these measures increased, the average results are quit low as the table presented.

Table 4-1: the recall and precision for the MNB classifier without enhancement

Class name	Precision	Recall	F-measure
Business	76.67%	78.41%	77.52%
Entertainment	65.56%	49.58%	56.45%
middle_east	64.44%	74.36%	69.04%
Scitech	62.22%	61.54%	61.87%
Sport	82.22%	94.87%	88.09%
World	61.11%	63.95%	62.50%
Average	68.70%	70.45%	69.25%

4.2. Enhanced Multinomial Naïve Bayes by Bi-gram Classifier

In this case a MNB classifier enhancing by using bi-gram has been be implemented. The obtained results will be evaluated in order to find the best N-gram can be used, by comparing the classified classes that the classifiers choose with the real class of the test document that included in the data set before dividing it to training document and test document. The Table (4-2) presented the Recall, precision and F-measure for MNB classifier enhancing by using bi-gram, the effectiveness of classifier increased while these measures increased, the average results are quit high as the table presented.

Table 4-2: the recall and precision for the MNB classifier with bi-gram classifier

Class name	Precision	Recall	F-measure
Business	85.56%	91.67%	88.50%
Entertainment	82.22%	57.36%	67.57%
Middle_east	84.44%	76.77%	80.42%
Scitech	80.00%	91.14%	85.20%
Sport	87.78%	96.34%	91.86%
World	65.56%	88.06%	75.15%
Average	80.93%	83.56%	81.46%

4.3. Enhanced Multinomial Naïve Bayes by Tri-gram Classifier

In this case a MNB classifier enhancing by using N-gram still implemented, but here a tri-gram has been be implemented. The obtained results will be evaluated in order to find the best N-gram can be used, by comparing the classified classes that the classifiers choose with the real class of the test document that included in the data set before dividing it to training document and test document. The Table (4-3) presented the Recall, precision and F-measure for MNB classifier

enhancing by using Tri-gram, the average results are lower than the results of MNB that enhance with bi-gram as the table presented.

Table 4-3: The recall and precision for the MNB classifier with tri-gram classifier

Class name	Precision	Recall	F-measure
Business	84.44%	73.79%	78.75%
Entertainment	33.33%	56.60%	41.95%
middle_east	81.11%	73.00%	76.84%
Scitech	76.67%	56.10%	64.78%
Sport	91.11%	91.11%	91.11%
World	65.56%	83.10%	73.29%
Average	72.04%	72.28%	71.12%

4.4. Enhanced Multinomial Naïve Bayes by 4-gram classifier

Finally a MNB classifier enhancing by 4-gram has been be implemented. The obtained results will be evaluated in order to find the best N-gram can be used. The Table (4-4) presented the Recall, precision and F-measure for MNB classifier enhancing by using 4-gram, the average results are very coverage with the results of MNB that enhance with Tri-gram.

Table 4-4: the recall and precision for the MNB classifier with 4-gram classifier

Class name	Precision	Recall	F-measure
Business	84.44%	73.79%	78.75%
Entertainment	33.33%	55.56%	41.66%
Middle_east	81.11%	73.00%	76.84%
Scitech	76.67%	56.56%	65.09%
Sport	91.11%	91.11%	91.11%
World	65.56%	83.10%	73.29%
Average	72.04%	72.18%	71.12%

4.5. Enhanced Multinomial Naïve Bayes by TF-IDF Classifier

In this section the second proposed classifier has been implemented that state on using the TF-IDF with the MNB classifier also this classifier implemented on an Arabic data set that contains six classes, and the results obtained for all the classes, and then take the average of all classes. The Table (4-5) presented the Recall, precision and F-measure for MNB classifier enhancing by using TF-IDF, the average results are high comparing with MNB classifier enhancing by using N gram it seems as a good classifier.

Table 4-5: the recall and precision for the MNB classifier with TF-IDF classifier

Class name	Precision	Recall	F-measure
Business	95.56%	94.51%	95.02%
Entertainment	91.11%	72.57%	80.78%
Middle_east	88.89%	86.02%	87.43%
Scitech	80.00%	96.00%	87.27%
Sport	93.33%	100.0%	96.55%
World	83.33%	89.29%	86.20%
Average	88.70%	89.73%	88.88%

4.6. Enhanced Multinomial Naïve Bayes by both of Bi-gram and TF-IDF

The final proposed classifier has been implemented, states on using both of TF-IDF and N-gram with the MNB classifier, also this classifier implemented on an Arabic data set that contains six classes, and the results obtained for all the classes, and then take the average of all classes. The Table (4-5) presented the Recall, precision and F-measure for MNB classifier enhancing by using both TF-IDF and N-gram, the average results are higher than the results of MNB with TF-IDF.

Table 4-6: the recall and precision for the MNB classifier with TF-IDF and bi-gram classifier

Class name	Precision	Recall	F-measure
Business	94.44%	90.43%	92.39%
Entertainment	76.67%	90.79%	83.13%
Middle_east	96.67%	75.00%	84.46%
Scitech	87.78%	96.34%	91.86%
Sport	95.56%	100.0%	97.72%
World	86.67%	90.70%	88.63%
Average	89.63%	90.54%	89.70%

Chapter Five

The Evaluation

5.1. Evaluation and Discussion

At first, a judgment stuck between the effectiveness of utilizing different type of n-gram should be accompanied, including of the bigram and trigram besides to the 4-gram, so as to take a decision about the utmost effectiveness n-gram approach, the judgment will be choose depending on the measurements of the F-measure, when the F-measure values rise up, also the effectiveness of the n-gram will be improved. Consequently the n-gram that records the uppermost values of the F-measure will be carry out in the third enhanced classifier (where the third enhanced classifier is the enhanced by the TF-IDF and the most effective type of n-gram).

5.2. Comparison between Different Types of N-gram

Figures from (5-1) to (5-3) below show the recall, precision and F-measures for the different n-gram types (bigram, trigram, 4-gram).

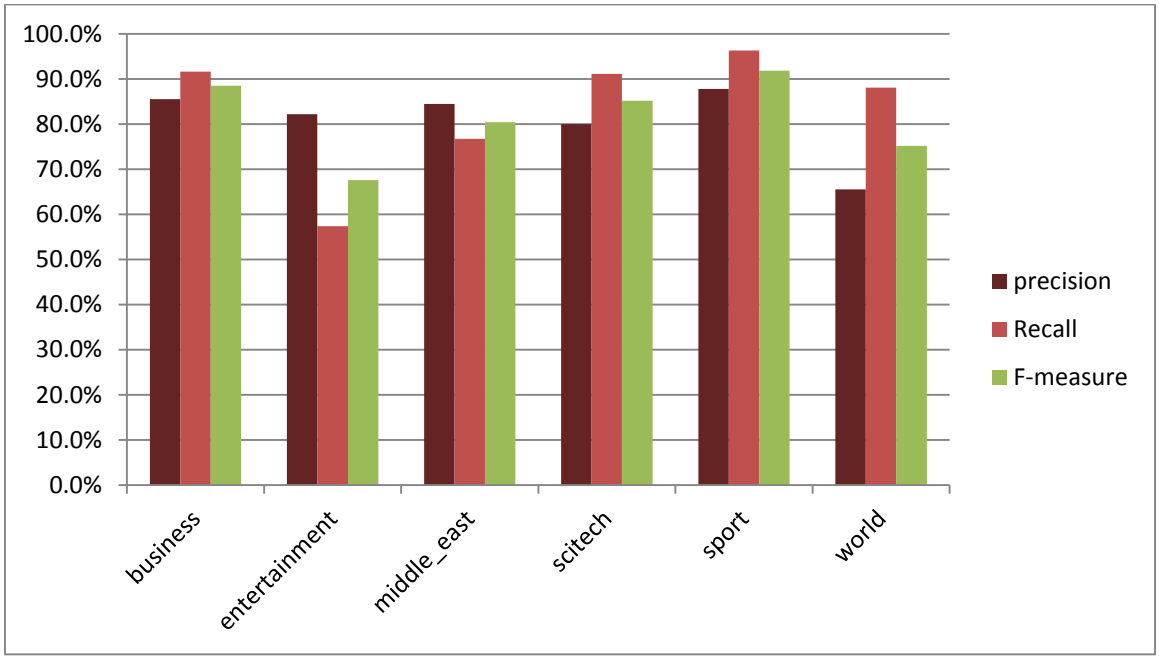


Figure 5-1: Recall and precision and F-measure for bigram

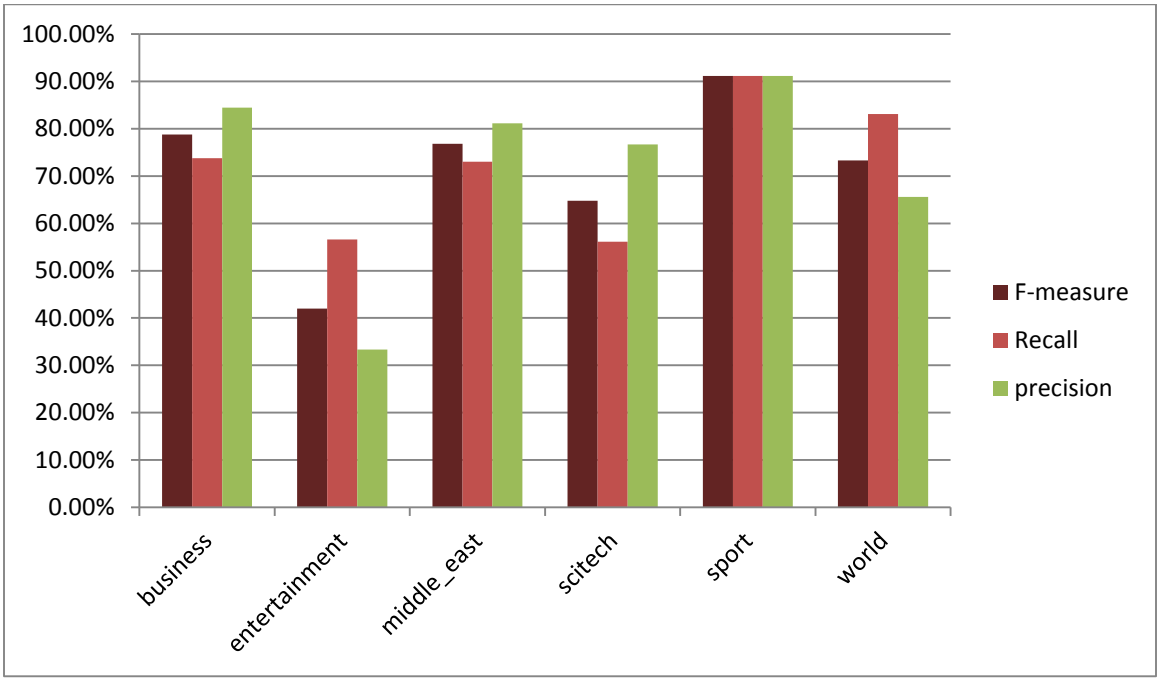


Figure 5-2: Recall and precision and F-measure for trigram

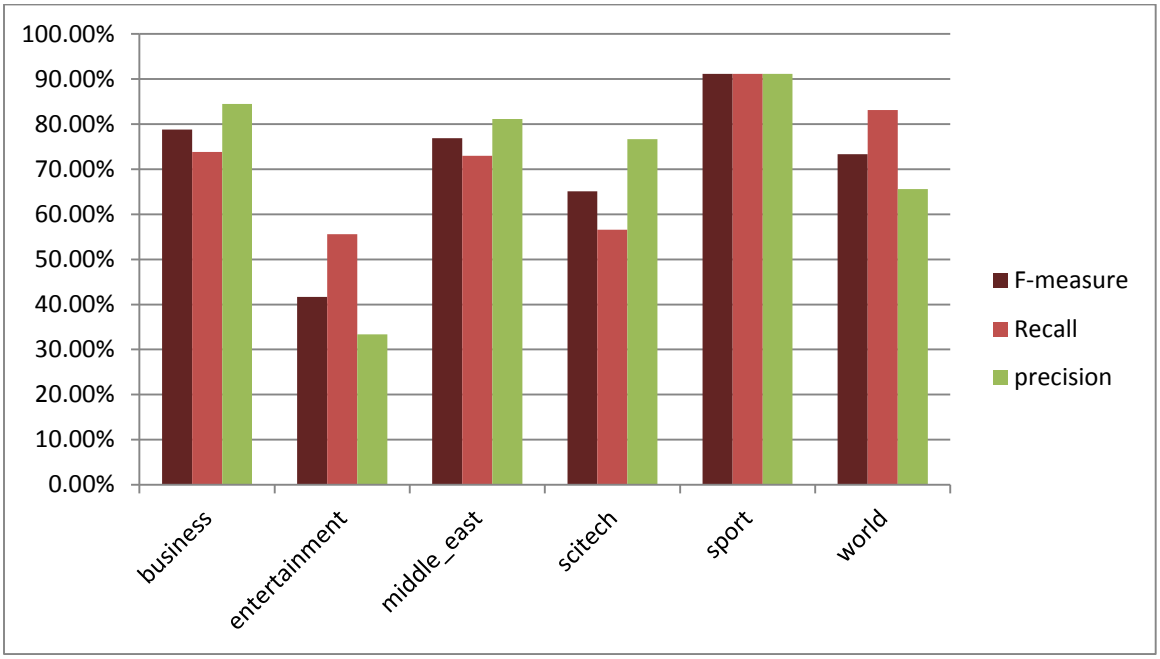


Figure 5-3: Recall and precision and F-measure for 4-gram

Table 5-1: variation of recall and precision and F-measures among different n-gram

Class name	Bi-gram			Trigram			4-gram		
	F-measure	Recall	Precision	F-measure	Recall	Precision	F-measure	Recall	Precision
Business	88.50%	91.67%	85.56%	78.75%	73.79%	84.44%	78.75%	73.79%	84.44%
Entertainment	67.57%	57.36%	82.22%	41.95%	56.60%	33.33%	41.66%	55.56%	33.33%
Middle_east	80.42%	76.77%	84.44%	76.84%	73.00%	81.11%	76.84%	73.00%	81.11%
Scitech	85.20%	91.14%	80.00%	64.78%	56.10%	76.67%	65.09%	56.56%	76.67%
Sport	91.86%	96.34%	87.78%	91.11%	91.11%	91.11%	91.11%	91.11%	91.11%
World	75.15%	88.06%	65.56%	73.29%	83.10%	65.56%	73.29%	83.10%	65.56%
Average	81.46%	83.56%	80.93%	71.12%	72.28%	72.04%	71.12%	72.18%	72.04%

Based on the former figures and table, the difference among the recall, precision and F-measure between different n-gram types is observed, it's clear to show that the average of F-measure. when appliance the bigram is (81.46%), while when appliance the tri-gram is (71.12%), and while the applying the 4-gram (71.12%). It's clear to show that the bi-gram accomplished the highest values, so verify that the utmost effectiveness type of n-grams types is the bigram, while the trigram and 4-gram realized low and very convergent values. Figure (5-4) presents the F-measures for all the classes, and the difference among them very clear.

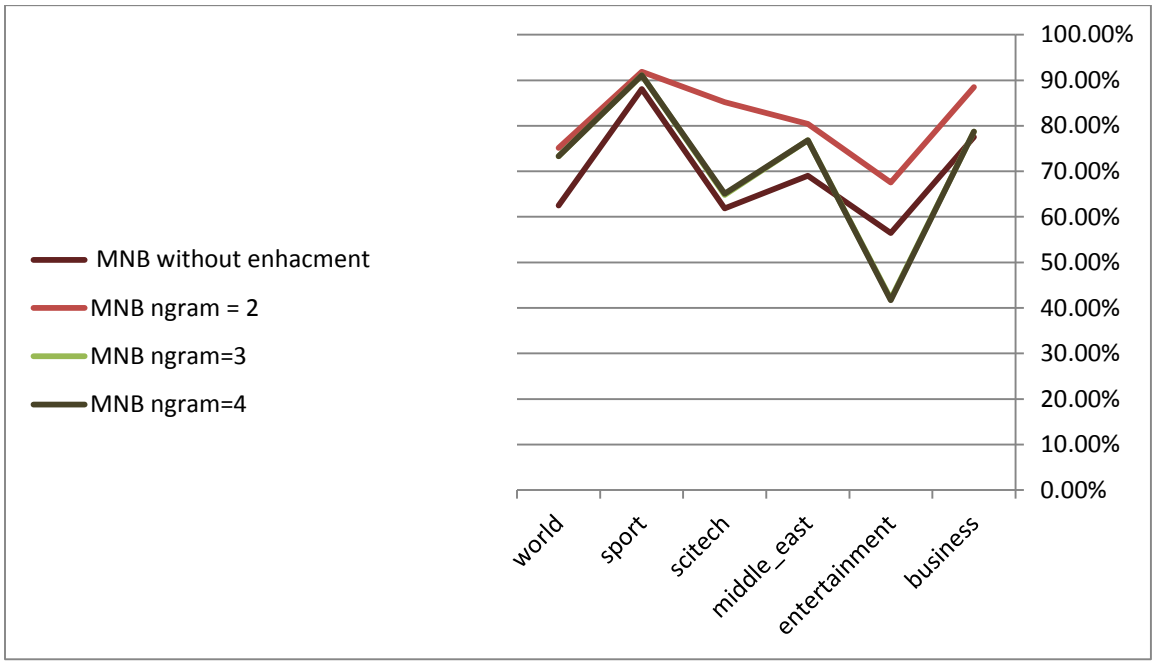


Figure 5-4 : variation of f-measures between different n-gram

5.3. Comparison between Proposed Classifier

Figures from (5-5) to (5-7) show the variability of the recall, precision and F-measures values for the different suggested classifiers.

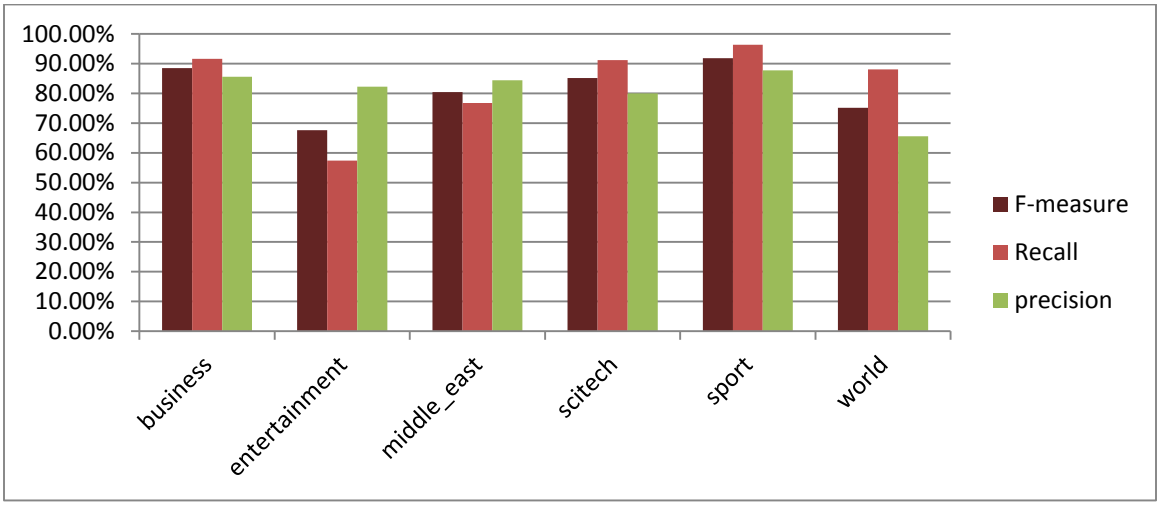


Figure 5-5: Recall and precision and F-measure for bigram

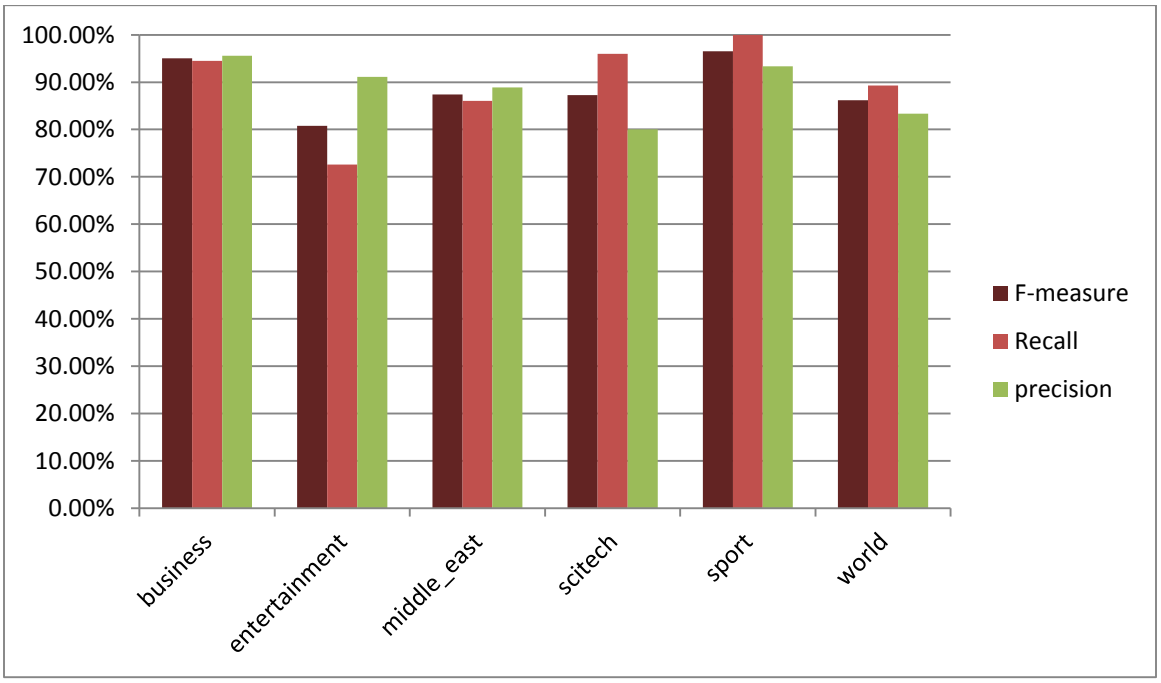


Figure 5-6: Recall and precision and F-measure for TF-IDF

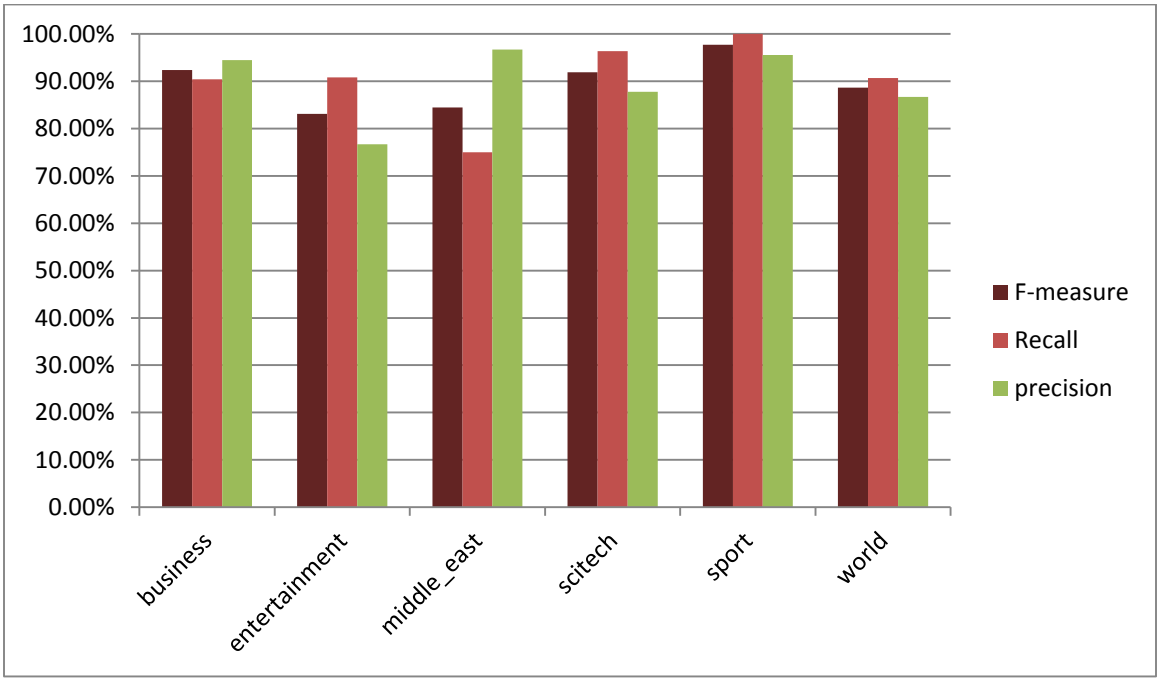


Figure 5-7: Recall and precision and F-measure for TF-IDF and bigram

Table 5-2:variation of Recall and precision and F-measures among proposed classifiers

Class name	Bi-gram			TF-IDF			TF-IDF + Bi-gram		
	F-measure	Recall	Precision	F-measure	Recall	Precision	F-measure	Recall	Precision
Business	88.50%	91.67%	85.56%	95.02%	94.51%	95.56%	92.39%	90.43%	94.44%
Entertainment	67.57%	57.36%	82.22%	80.78%	72.57%	91.11%	83.13%	90.79%	76.67%
Middle_east	80.42%	76.77%	84.44%	87.43%	86.02%	88.89%	84.46%	75.00%	96.67%
Scitech	85.20%	91.14%	80.00%	87.27%	96.00%	80.00%	91.86%	96.34%	87.78%
Sport	91.86%	96.34%	87.78%	96.55%	100.0%	93.33%	97.72%	100.0%	95.56%
World	75.15%	88.06%	65.56%	86.20%	89.29%	83.33%	88.63%	90.70%	86.67%
Average	81.46%	83.56%	80.93%	88.88%	89.73%	88.70%	89.70%	90.54%	89.63%

Depending on the prior figures (5-5) to (5-7) and according to table(5-2), the variation among the recall, precision and F-measure between the different suggested classifiers is clear , as an example,

in world class the F-measure when appliance bigram type is (75.15%), while when appliance the TF-IDF is (86.20%), also when applying the mixture of both bigram and TF-IDF is (88.63%),and if we need to have the average of F-measure for all classes when applying the bigram was (81.46%), while the average of F-measure for all classes when applying TF-IDF was (88.88%)and the average of F-measure for all classes when applying the combination of both bigram and TF-IDF was (89.70%), we can observe that the classifier which enhanced by utilizing both of TF-IDF and bigram accomplished the uppermost values, so we prove that the maximum effectiveness classifier between the three suggested classifier, it is the classifier that enhanced by utilizing the both of TF-IDF and bigram, after that the classifier that enhanced by utilizing only the TF-IDF, finally the classifier that enhanced by utilizing the bigram only.

The classifier that enhanced by utilizing both of the TF-IDF and the bigram accomplished the maximum values, so to prove that the maximum effectiveness classification tool between the three offered classifier it's the classifier that enhanced by utilizing both the TF-IDF and the bigram, then the classifier that enhanced by utilizing only TF-IDF and finally the classifier that enhanced by utilizing the bi-gram only. Figure (5-8) below shows the F-measures for all the classes when appliance all of the suggested classifiers , even that the average of the F-measure for the classifier who enhanced by utilizing only the TF-IDF less than the average of F-measure for the classifier who enhanced with the mixture of both the TF-IDF and the bigram, but in some of the classes the classifier that enhanced by only TF-IDF registers a higher F-measure value than the classifier that enhanced by the TF-IDF and the bi-gram , this phenomenon could be interrelated to the nature of the documents in these class, may be does not have expressive consecutive strings, nonetheless as an average the classifier that enhanced with the combination of both the TF-IDF and bigram annals a highest F-measure also a highest accuracy.

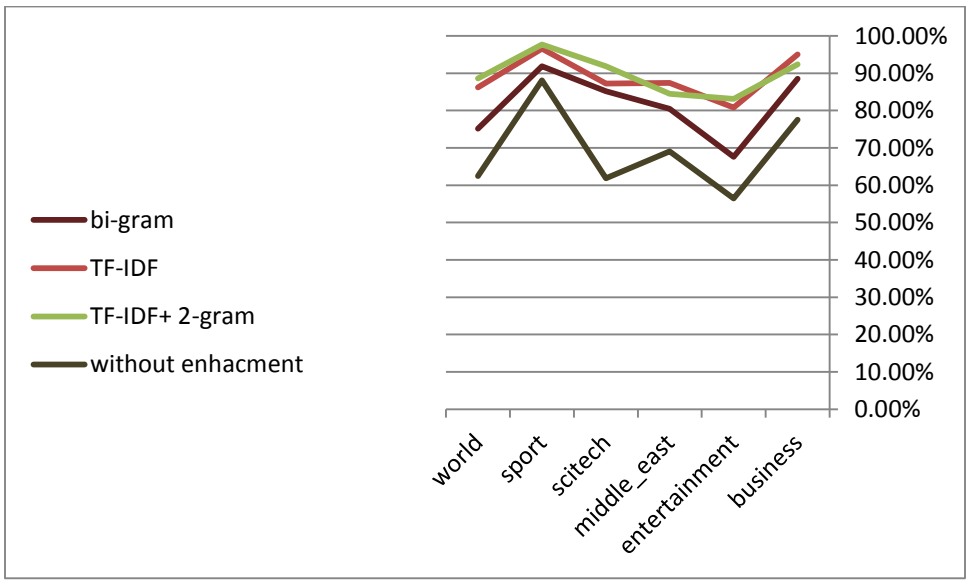


Figure 5-8 : F-measure for different proposed classifiers

Chapter Six

Conclusion and Future Work

6.1. Conclusion

This study was suggested three approaches to enhance the performance of the MNB classifier. The enhanced MNB classifiers have been evaluated to find the maximum enhancement technique. The MNB has been enhanced by implemented the n-gram and the TF-IDF, in addition to enhance it by using both of the n-gram and the TF-IDF. Different measures have been used to evaluate the performance of the three suggested classifiers such as the recall, precision, and F-measure. The average of F-measure for all classes when apply the bigram was (81.46%), while the average of F-measure for all classes when apply TF-IDF was (88.88%) and the average of F-measure for all classes when apply the combination of both bigram and TF-IDF was (89.70%). The variance F-measure between the different three suggested classifiers verified that the classifier which is enhanced by using both of the TF-IDF and bigram accomplished the highest values and it characterizes as the most effective classifier between the three suggested classifier. In the second stage of effectiveness, the classifier that enhanced by using only TF-IDF and finally the classifier which enhanced by using only the bigram.

6.2. Future work

As a future work:

- Another model of the Naïve Bays classifier will be investigated such as the multi-variate Bernoulli.
- Working with more complicated models of classifiers and discuss how to increase the performance of it.

- Discussing the effects of document size in the classification, looking for correlation between the Multinomial Naïve Bays and the document length.
- Experiment with the effect of varying the number of training document over the performance of "Multinomial Naïve Bayes" classifiers.

References

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S., & Al-Rajeh, A. (2008). Automatic Arabic text classification.

Alkafije, A., &Ajam, G. (2013). "Improving Document Processing and Indexing by Preprocessing and Tokenization. young, 103, 0-2543.

Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. Int. Arab J. e-Technol., 2(2), 124-128.

Al-Shalabi, R., &Obeidat, R. (2008, March). Improving KNN Arabic text classification with n-grams based document indexing. In Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt (pp. 108-112).

Anagnostopoulos, A., Broder, A. Z., &Punera, K. (2006, November). Effective and efficient classification on a search-engine model. In *Proceedings of the 15th ACM international conference on Information and knowledge management*(pp. 208-217). ACM.

Denoyer, L., Zaragoza, H., & Gallinari, P. (2001, March). HMM-based passage models for document classification and ranking. In Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research (pp. 126-135).

Dharmadhikari, S. C., Ingle, M., & Kulkarni, P. (2012). Analysis of semi supervised learning methods towards multi label text classification. International Journal of Computer Applications, 42(16).

Domingos, P., &Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 29(2-3), 103-130.

Du, R., Safavi-Naini, R., &Susilo, W. (2003). Web filtering using text classification. *Faculty of Informatics-Papers*, 166.

Duwairi, R. M. (2006). Machine learning for Arabic text categorization. Journal of the American Society for Information Science and Technology, 57(8), 1005-1010.

Duwairi, R. M. (2007). Arabic Text Categorization. Int. Arab J. Inf. Technol.,4(2), 125-132.

El Kourdi, M., Bensaid, A., & Rachidi, T. E. (2004, August). Automatic Arabic document categorization based on the Naïve Bayes algorithm. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (pp. 51-58). Association for Computational Linguistics.

Frank, E., &Bouckaert, R. R. (2006). Naive bayes for text classification with unbalanced classes. In *Knowledge Discovery in Databases: PKDD 2006* (pp. 503-510). Springer Berlin Heidelberg.

Godbole, S., Harpale, A., Sarawagi, S., &Chakrabarti, S. (2004). Document classification through interactive supervision of document and term labels. InKnowledge Discovery in Databases: PKDD 2004 (pp. 185-196). Springer Berlin Heidelberg.

Goller, C., Löning, J., Will, T., & Wolff, W. (2000). Automatic Document Classification-A thorough Evaluation of various Methods. *ISI*, *2000*, 145-162.

Han, E. H. S., &Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results (pp. 424-431). Springer Berlin Heidelberg.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85.

Ikonomakis, M., Kotsiantis, S., &Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, *4*(8), 966-974.

Lam, W., Ruiz, M., & Srinivasan, P. (1999). Automatic text categorization and its application to text retrieval. *Knowledge and Data Engineering, IEEE Transactions on*, *11*(6), 865-879.

Li, Y. H., & Jain, A. K. (1998). Classification of text documents. *The Computer Journal*, *41*(8), 537-546.

Liu, X. (2008). Proposal of Document Classification with Word Sense Disambiguation.

Manning, C. D., Raghavan, P., &Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 496). Cambridge: Cambridge university press.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).
Mitchell, T. M. (1997). Machine learning. WCB.

Moulinier, I., & Jackson, P. (2002). Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization.

Mustafa, S. H. (2012). Word Stemming for Arabic Information Retrieval: The Case for Simple Light Stemming. Abhath Al-Yarmouk: Science & Engineering Series, 21(1), 2012.

Nanas, N., Domingue, J., Watt, S., & Motta, E. (2001). Literature Review: Information Filtering for Knowledge Management. Technical Report, KMI-TR-113, Knowledge Media Institute, The Open University.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, *39*(2-3), 103-134.

Padhye, A. (2006). Comparing Supervised and Unsupervised Classification of Messages in the Enron Email Corpus (Doctoral dissertation, UNIVERSITY OF MINNESOTA).

Patrick, T. (2009). An Introduction to Unsupervised Document Classification. University of Maryland Baltimore County.

Pop, I. (2006). An approach of the Naive Bayes classifier for the document classification. General Mathematics, 14(4), 135-138.

Power, R., Chen, J., Kuppusamy, T. K., & Subramanian, L. (2010, March). Document Classification for Focused Topics. In AAAI Spring Symposium: Artificial Intelligence for Development.

Rajaraman, A., & Ullman, J. D. (2012). Mining of massive datasets (Vol. 77). Cambridge: Cambridge University Press.

Ramdass, D., &Seshasai, S. (2009). Document classification for newspaper articles.

Rennie, J. D., Shih, L., Teevan, J., &Karger, D. R. (2003, August). Tackling the poor assumptions of naive bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. InIJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). IBM New York.

Saad, M. K., &Ashour, W. (2010, November). OSAC: Open Source Arabic Corpora. In *6th International Symposium on Electrical and Electronics Engineering and Computer Science, Cyprus* (pp. 118-123).

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. Expert Systems with Applications, 33(1), 1-5.

Shimodaira, H. (2014). Text Classification using Naive Bayes. Learning and Data Note 7. Informatics 2B.

Sivic, J., & Zisserman, A. (2009). Efficient visual search of videos cast as text retrieval. Pattern Analysis and Machine Intelligence, IEEE Transactions on,31(4), 591-606.

Steffen, J. (2004). N-Gram Language Modeling for Robust Multi-Lingual Document Classification. In LREC.

Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naive Bayes a good classifier for document classification?. *International Journal of Software Engineering and Its Applications*, *5*(3), 37-46.

Wijewickrema, C. M., &Gamage, R. (2013). An ontology based fully automatic document classification system using an existing semi-automatic system.

Zečević, A. (2011, September). N-gram based text classification according to authorship. In Student Research Workshop (pp. 145-149).

Appendices

A. The Implemented of Classes

To contrivance the system numerous classes have been utilized, Figure (A-1) shows the major classes that have been used to implement this system, and Figures from (A-2) to (A-9) show the architecture of each class, and Figures from (A-10) to (A-11) present the class diagrams for the two implemented class diagrams

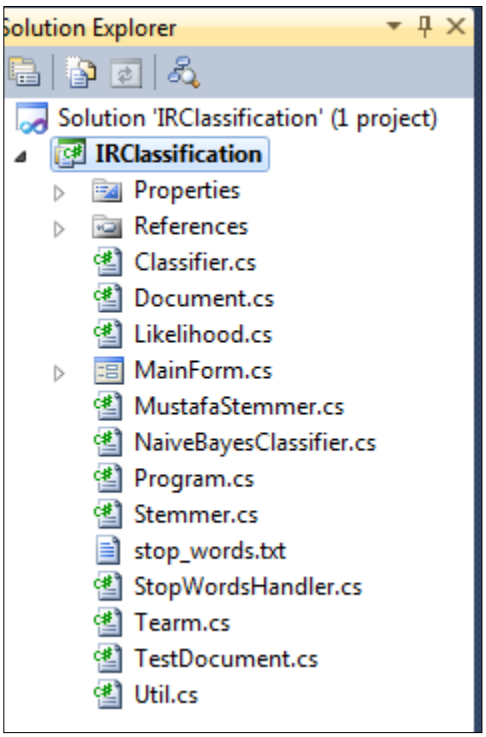


Figure A-1: the implemented classes

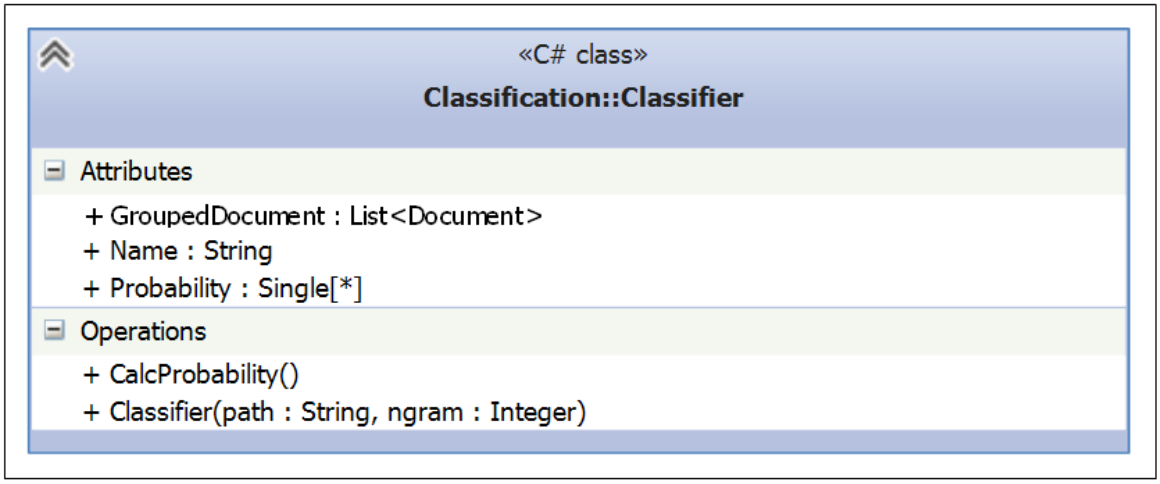


Figure A-2: the classifier classes

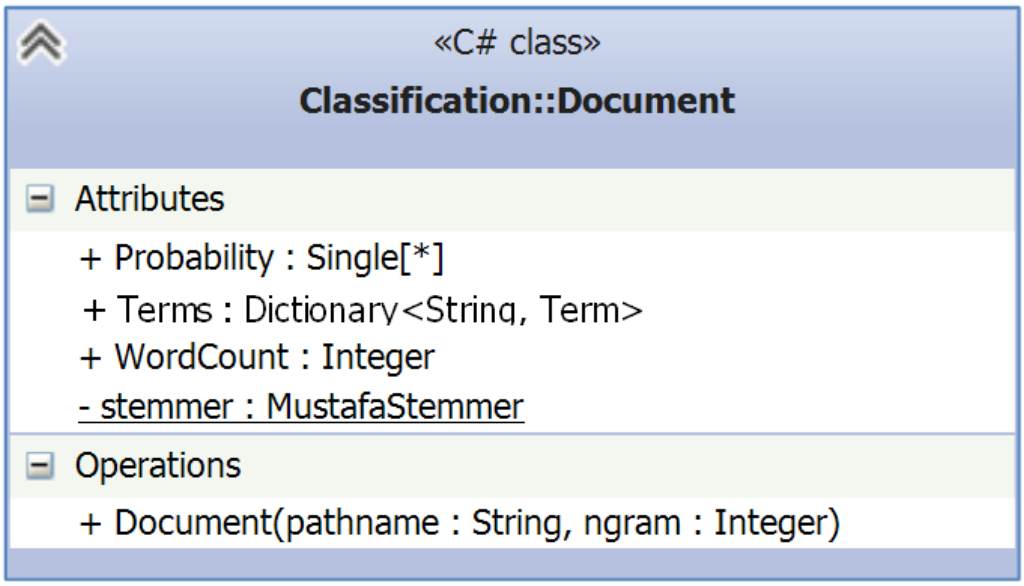


Figure A-3: Document class

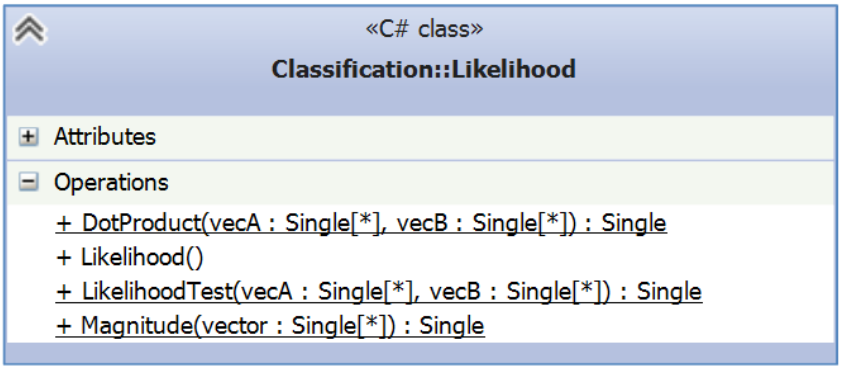


Figure A-4: likelihood class

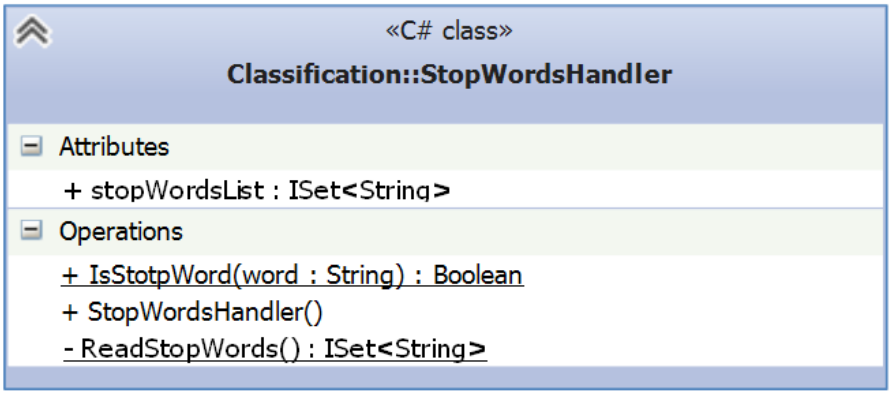


Figure A-5: stop word handler class

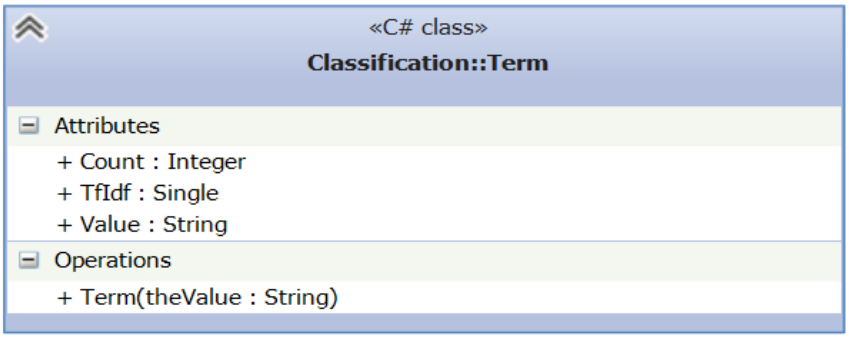


Figure A-6: Term class

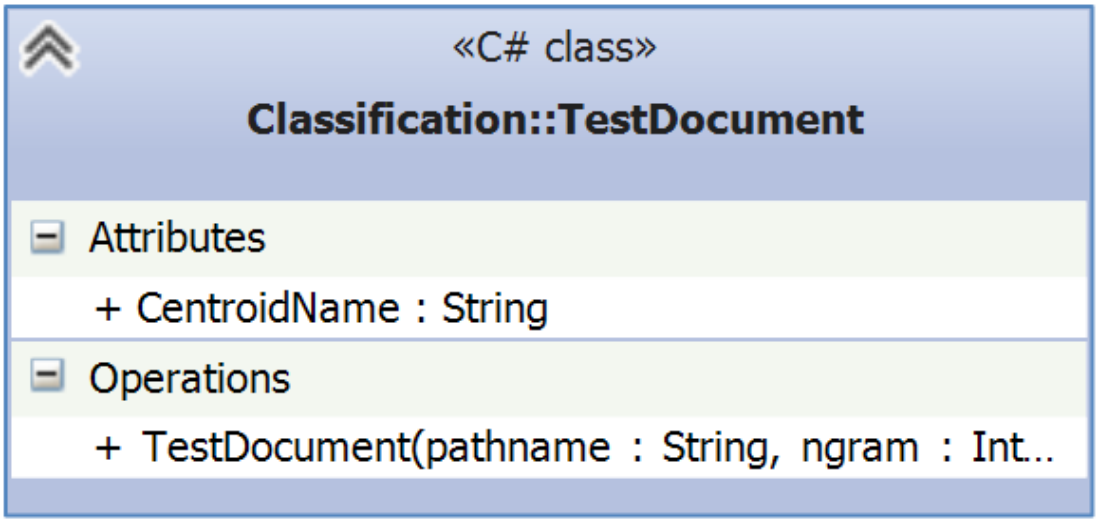


Figure A-7: Test document class

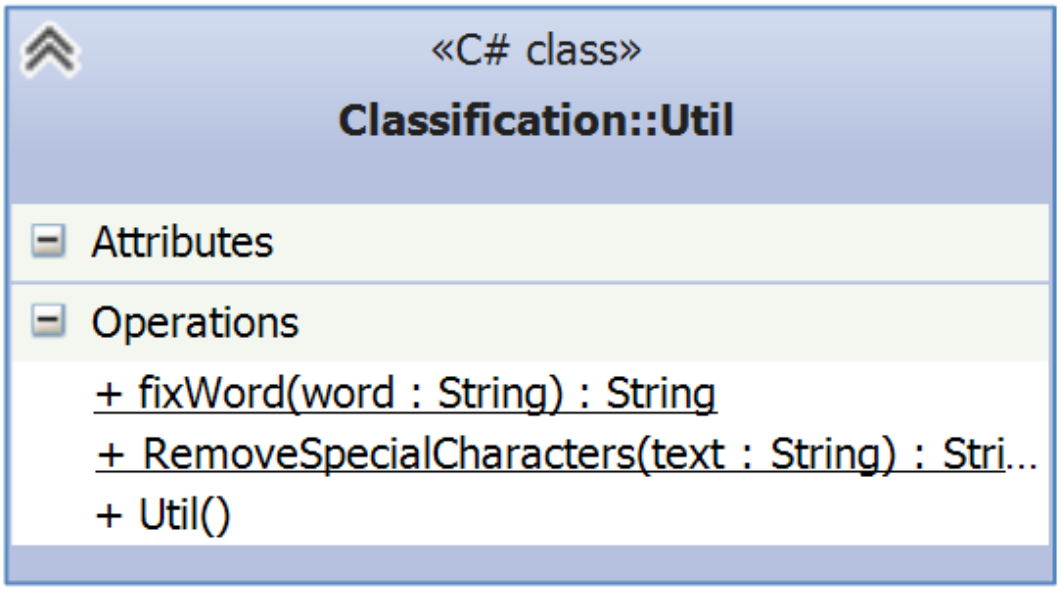


Figure A-8: Util class

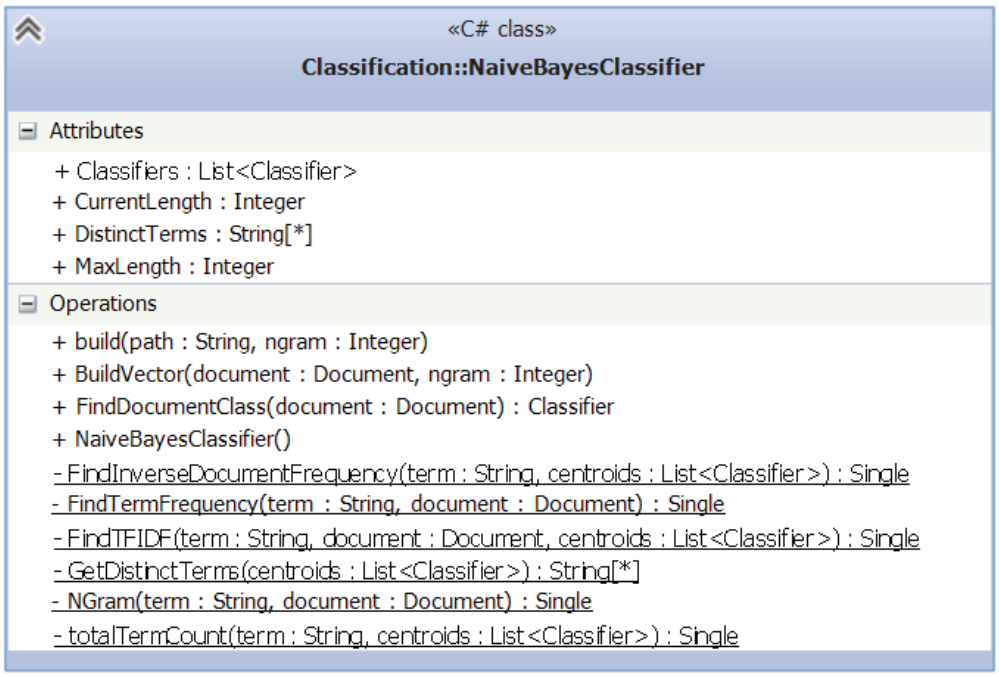


Figure A-9: Naïve Bayes Classifier class

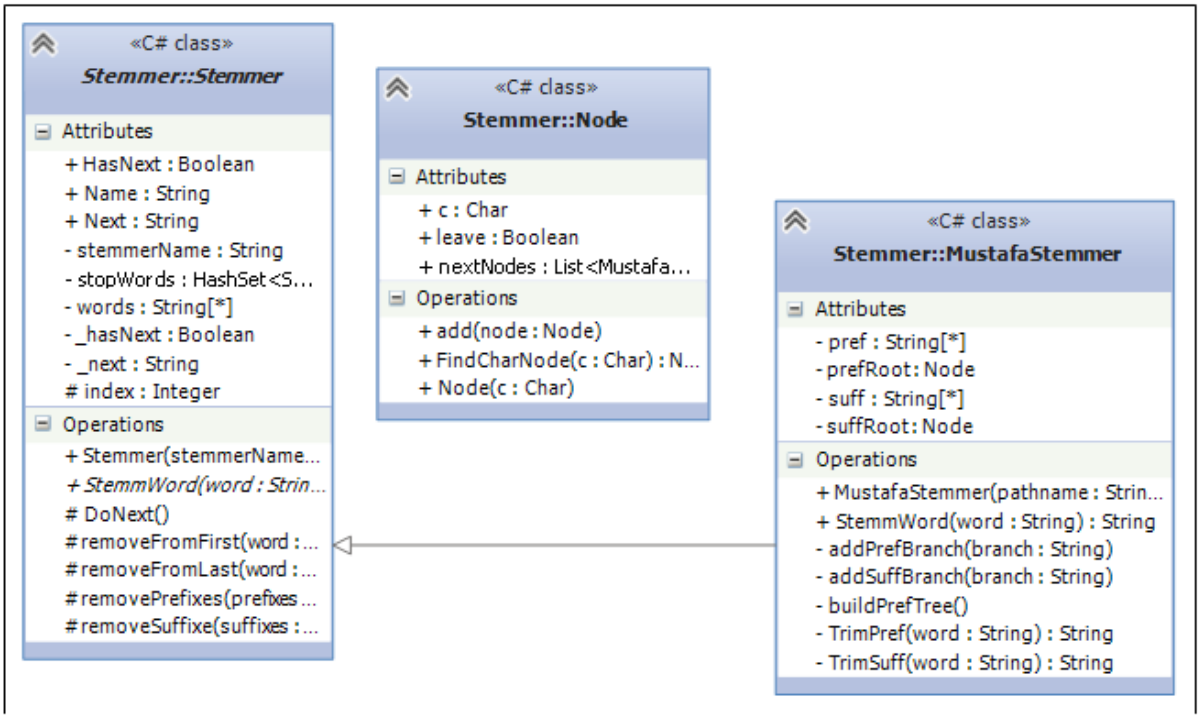


Figure A-10: the class diagram of the stemmer name space

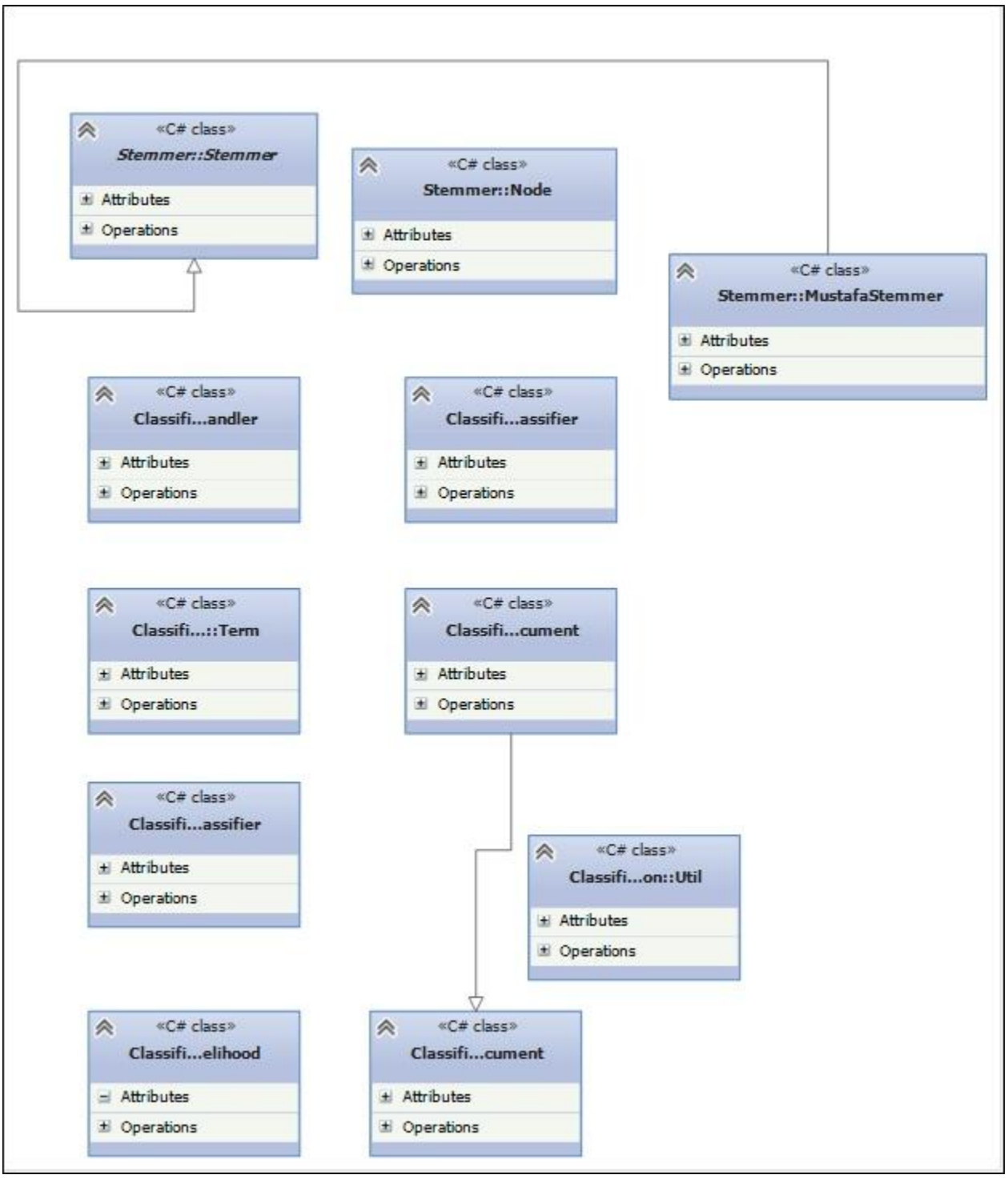


Figure A-11: class diagram for the classification name space

B. The Results

Multinomial Naïve Bayes without Enhancements

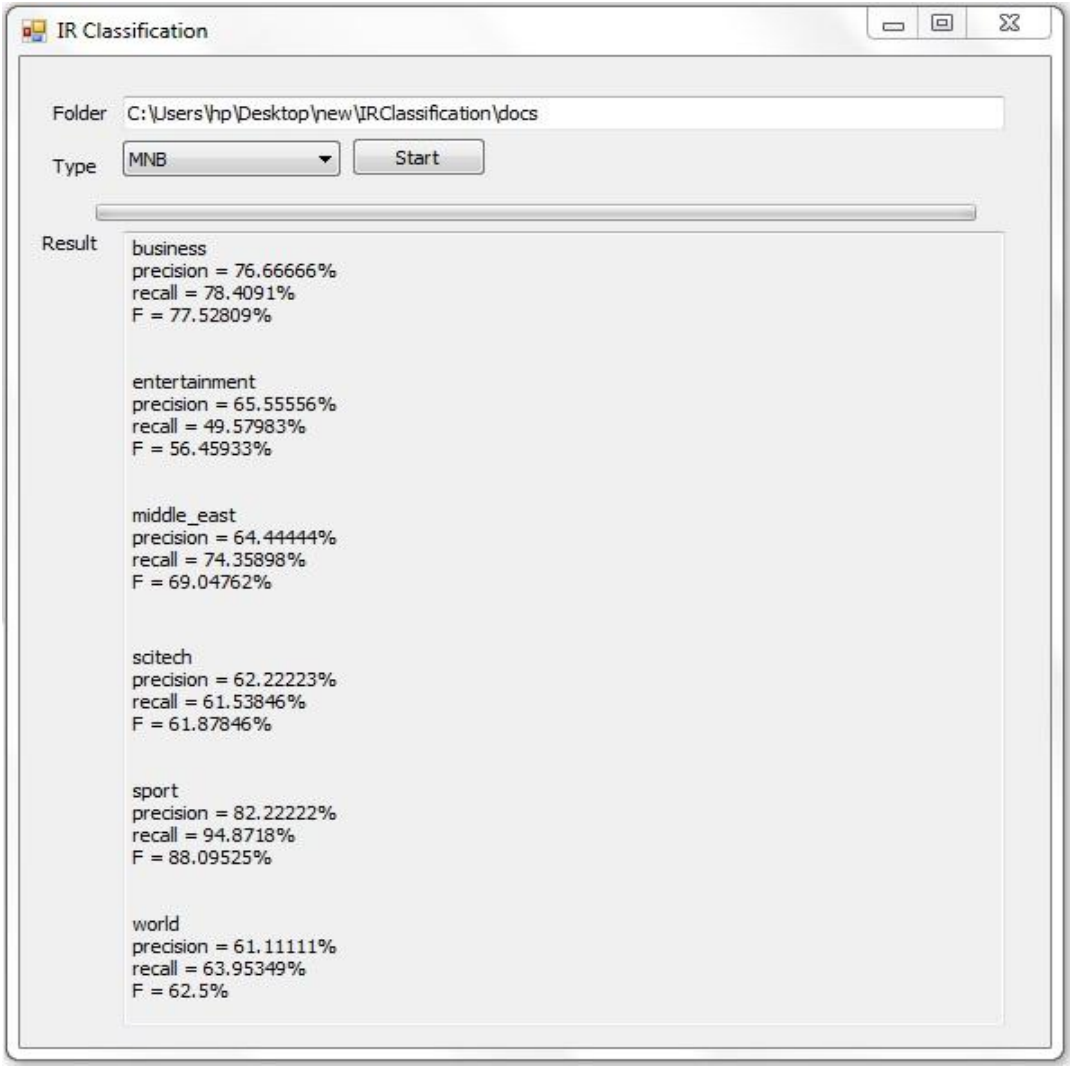


Figure B-1: the Recall and precision for the MNB classifier without enhancement

Enhanced Multinomial Naïve Bayes by Bi-gram Classifier

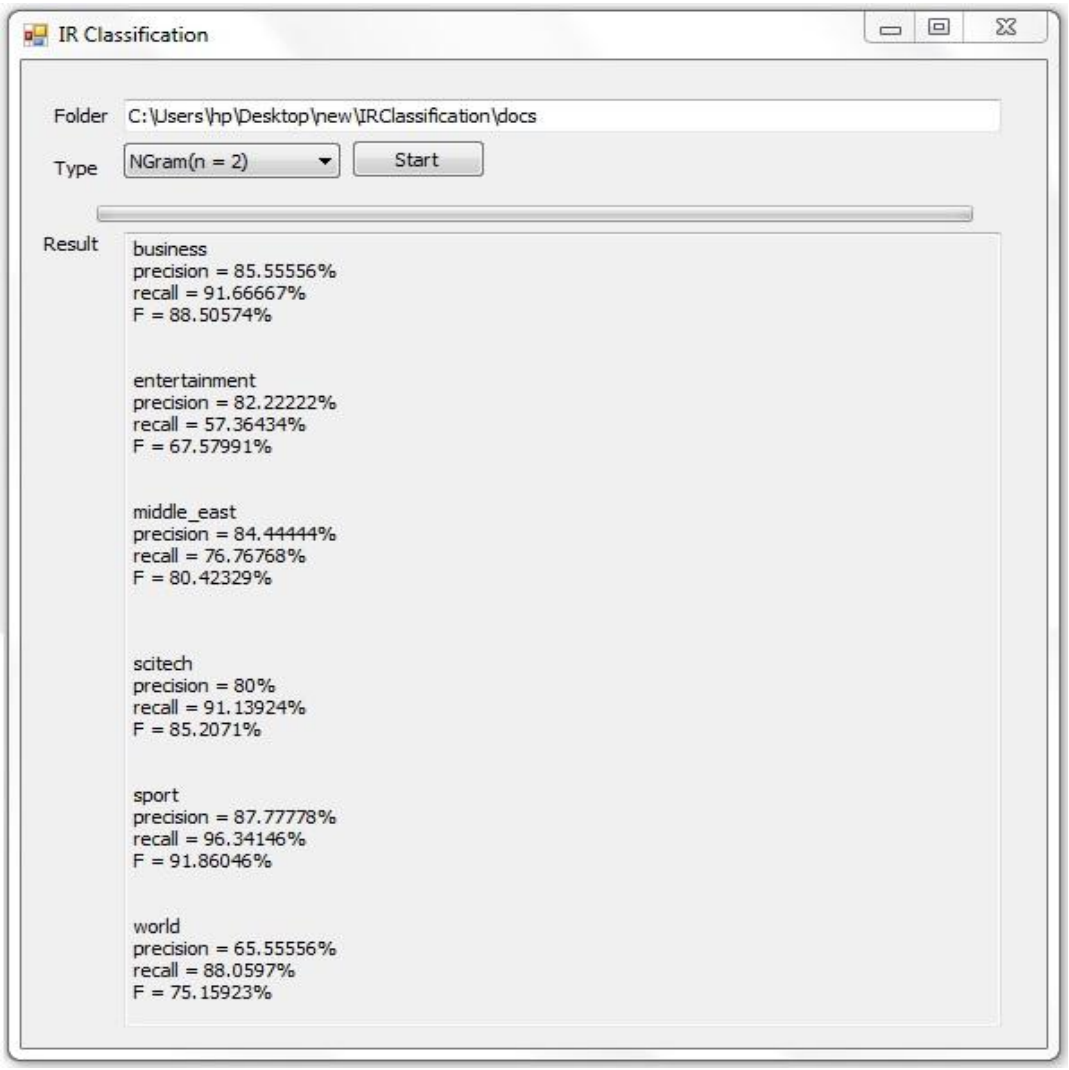


Figure B-2: The Recall and precision for the MNB classifier with bi-gram classifier

Enhanced Multinomial Naïve Bayes by Tri-gram Classifier

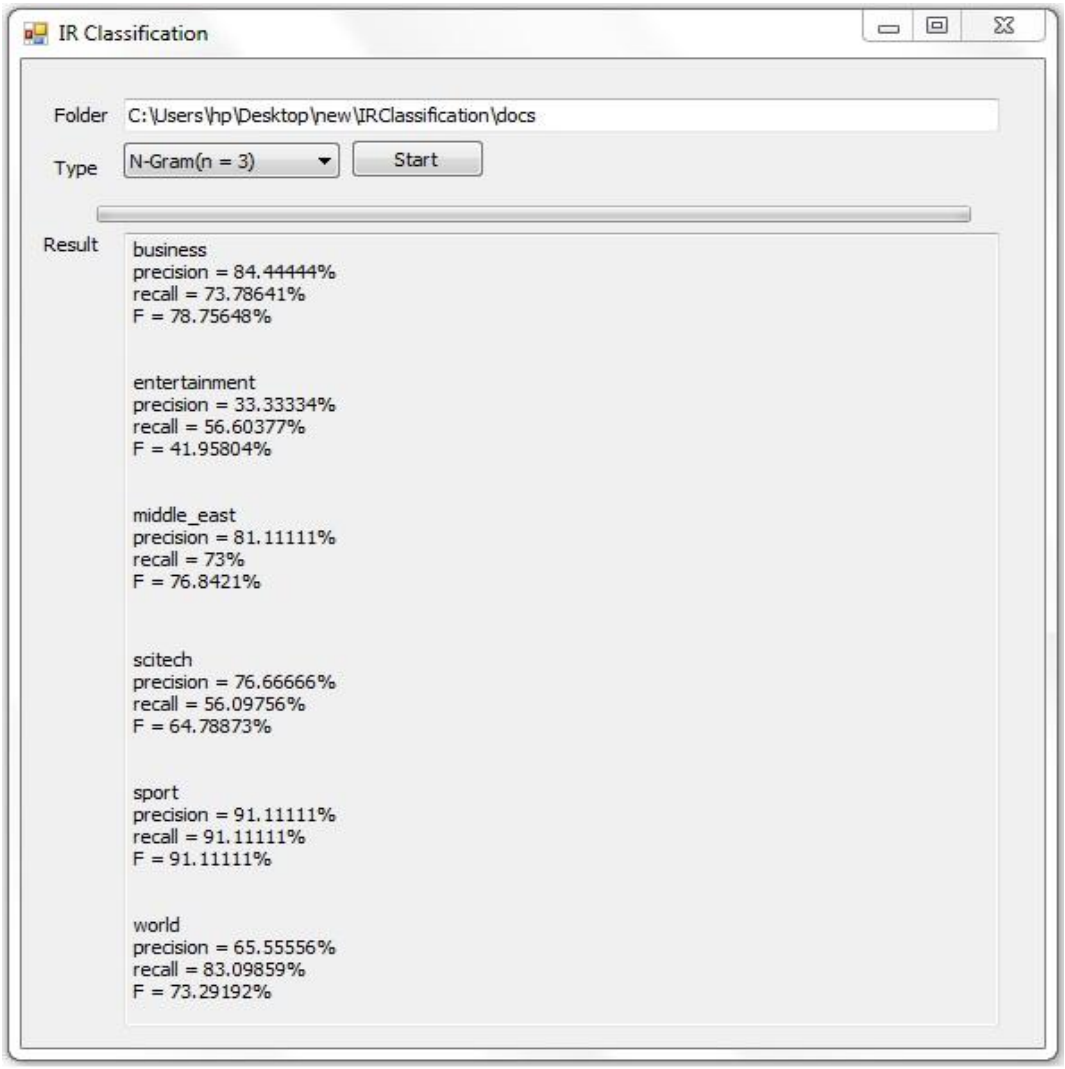


Figure B-3: the Recall and precision for the MNB classifier with tri-gram classifier

Enhanced Multinomial Naïve Bayes by 4-gram Classifier

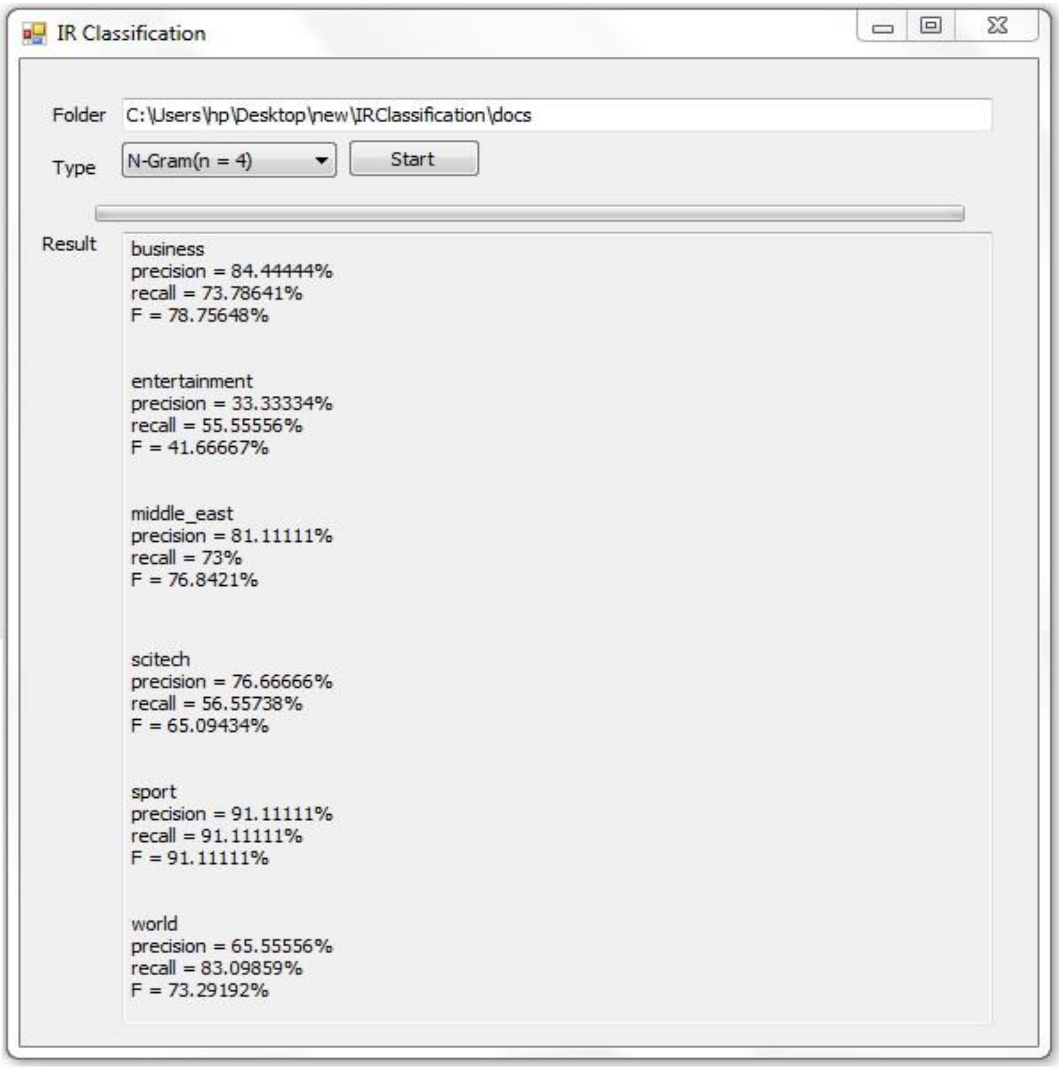


Figure B-4: the Recall and precision for the MNB classifier with 4-gram classifier

Enhanced Multinomial Naïve Bayes by TF-IDF Classifier

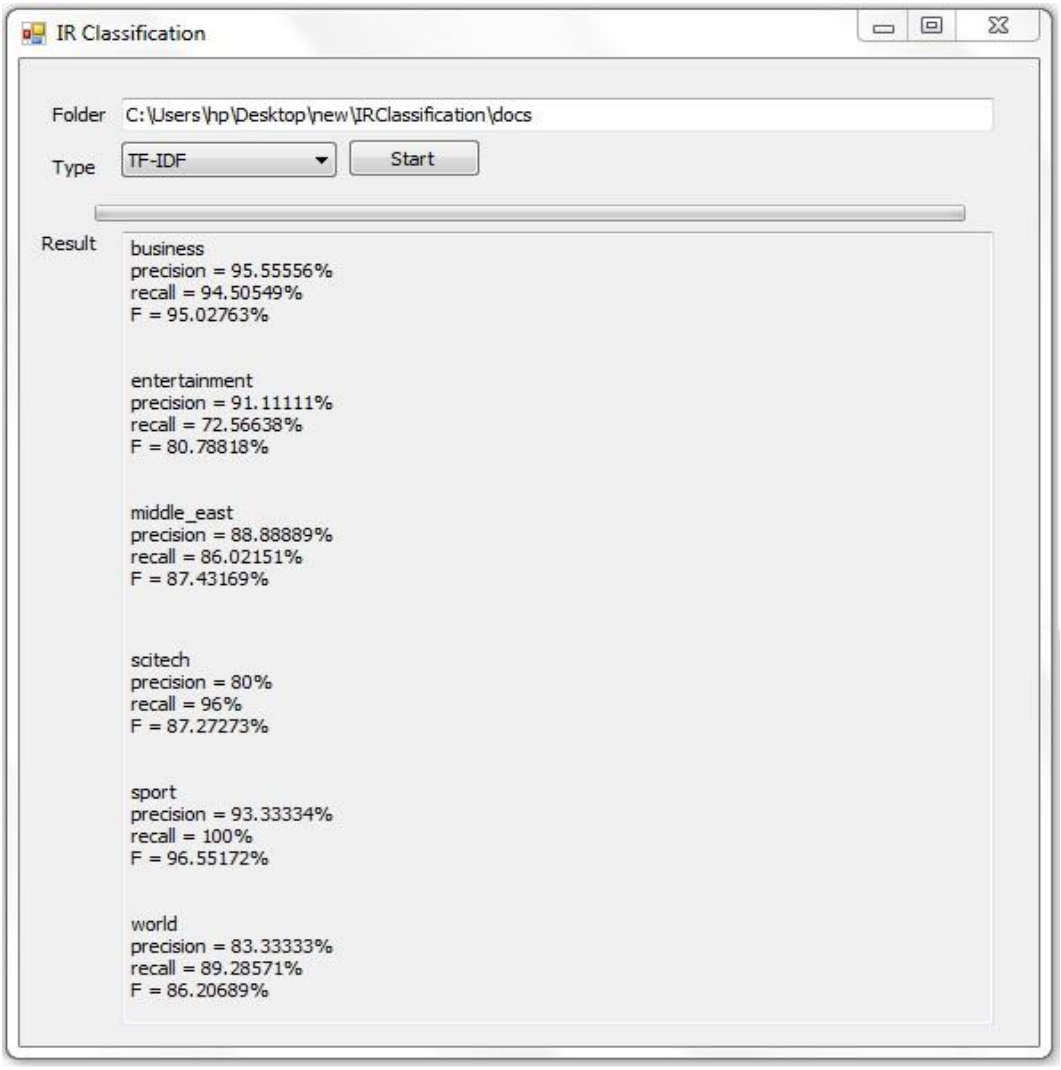


Figure B-5: the Recall and precision for the MNB classifier with TF-IDF classifier

Enhanced Multinomial Naïve Bayes by Bi-gram and TF-IDF Classifier

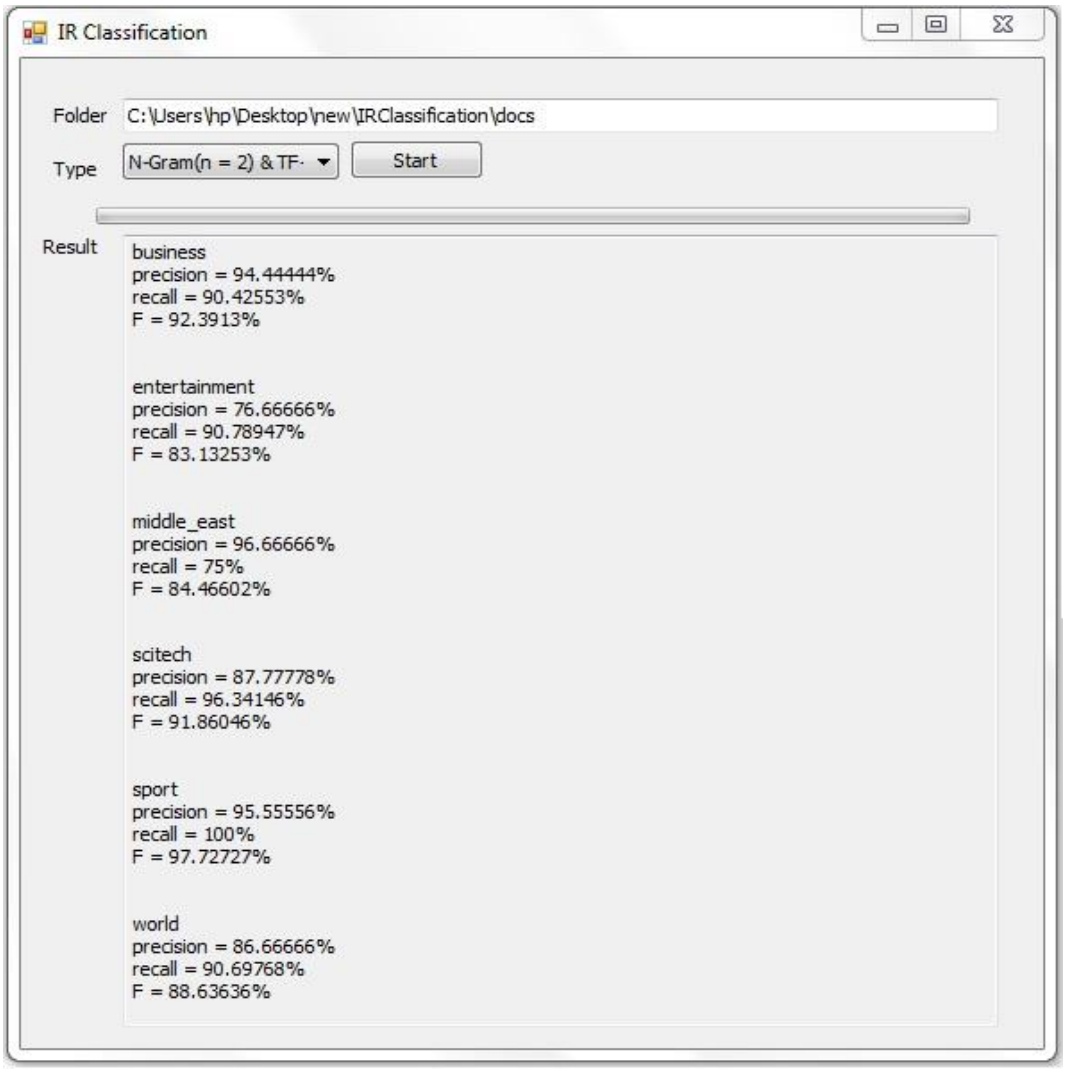


Figure B-6: the Recall and precision for the MNB classifier with TF-IDF and bi-gram classifier