نموذج رقم (1)

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Automated Complaint System Using Text Mining Techniques
(UNRWA Case Study)

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

DECLARATION

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

Student's name:	اسم الطالب: محمد رضوان عبد الحميد النجار
Signature:	التوقيع:
Date:	التاريخ: 2015/1/13

Islamic University of Gaza
Deanery of Higher Studies
Faculty of Information Technology
Information Technology

Automated Complaint System Using Text Mining Techniques

(UNRWA Case Study)

Submitted by

Mohammed R. A. ALNajjar (120110524)

Supervised by
Dr. Alaa EL-Halees

A Thesis Submitted in Partial Fulfillment of the Requirements
For the Degree of Master In Information Technology

2014

بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا هاتف داخلي: 1150

الرقم Ref ج س غ/35/

التاريخ Date 2014/12/28م

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **محمد رضوان عبدالحميد النجار** لنيل درجة الماجستير في **كلية *تكنولوجيا المعلومات*** **برنامج تكنولوجيا المعلومات** وموضوعها:

نظام شكاوي آلي باستخدام تقنية التنقيب عن النص (الأنروا كدراسة حالة)

Automated Complaint System using Text Mining techniques

وبعد المناقشة العلنية التي تمت اليوم الأحد 06 ربيع أول 1436هـ، الموافق 2014/12/28م الساعة العاشرة صباحاً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. علاء مصطفى الهليس	مشرفاً ورئيساً	.....
د. ربحي سليمان بركة	مناقشاً داخلياً	.....
أ.د. سامي سليم أبو ناصر	مناقشاً خارجيًا	.....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية *تكنولوجيا المعلومات*/ برنامج تكنولوجيا المعلومات.**

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

والله ولي التوفيق ،،،

مساعد نائب الرئيس للبحث العلمي والدراسات العليا

أ.د. فؤاد علي العاجز

P.O. Box 108, Rimal, Gaza, Palestine fax: +970 (8) 286 0800 فاكس Tel: +970 (8) 286 0700 ص.ب. 108 الرمال، غزة، فلسطين هاتف
public@iugaza.edu.ps www.iugaza.edu.ps

Acknowledgements

First of all, I thank Allah for all knowledge and education I gain which leads to achievement of this thesis.

Second, I would like to thank my advisor Dr. Alaa EL-Halees for giving me the opportunity to work on this project and his valuable guidance.

I would also like to thank my parents, and my wife for their constant support and encouragement.

Thanks for all members of response unit in UNRWA for giving me the required data and showing me the problems of the existed complaint system in UNRWA.

Thanks to anyone who participated in all the achievement of this thesis either directly or indirectly.

Abstract

Complaints System is the system that manages the process of how organizations handle, manage, respond and report to client's complaints. Manual organizing for large number of requests is extremely difficult, time consuming, error prone, expensive and often not feasible. Results also may differ according to the variety of expert's judgments. Not forgetting that there would be many questions that already been answered before. For example organization such as UNRWA, receive many complaints each day and make categorization for each request manually based on the contents of the message, forwarding the request to the responsible person according to its category to get the answer.

The problem of increasing the cost and efforts required to manage the complaints manually leads to the need to develop automated solutions to handle this problem by including text-mining techniques to substitute the human part. The solution will deal with Arabic content that is different from English which makes data analysis a complex task. Little researches have been conducted on Arabic corpuses mainly because it is highly rich and requires special treatments such as verbs order and morphological analysis.

In our work, we propose a new solution to overcome the manual system limitations that consists of three phases. First, we analyze the text message contents, categorize it by using text categorization algorithms and try to decide where to direct the question request automatically to the right person in order to get it answered. Then, we will use text similarity techniques to suggest the answers automatically. Finally, system will use summarization techniques to update the FAQ library with the most asked questions. As a result, the automated complaints system will improve the quality of answering questions by speeding the process and minimizing the required time and effort. We found that the process is efficient and effective. According to results analysis for the classification part, the developed classifier by SVMs achieved the highest average accuracy (74.69%). Also for the answers suggestion part, we obtained best F-Measure (72.45%) at similarity score (0.50). For Summarization part, we obtained the best results at compression rate =0.3, the best F-Measure was 71.56%.

Keywords: *Feedback Mechanism, Complaints mechanism, Text Mining, Text Categorization, Text Classification, Text summarization.*

عنوان البحث

نظام شكاوي آلي بإستخدام تقنية التنقيب عن النص (الأنروا كدراسة حالة)

ملخص

نظام الشكاوى هو النظام الذي يمكن المنظمات من التعامل مع الشكاوي والرد عليها .حيث ان المعالجة اليدوية لعدد كبير من الطلبات يعتبر صعب للغاية , ويستغرق وقتا طويلا , وايضا يكون عرضه للخطأ ومكلف , وغالبا ما يكون غير مجدي. و قد تختلف النتائج وفقا لأحكام الخبراء. ولا ننسى أن هناك العديد من الأسئلة تم الإجابة عليها من قبل . على سبيل المثال : الأونروا تتلقي العديد من الشكاوى كل يوم وتقوم بتصنيف كل طلب يدويا بناء على محتويات الرسالة، وبعدها تقوم بتوجيه الطلب إلى الشخص المسؤول حسب تصنيفه للحصول على الاجابة.

ولكن مشكلة زيادة التكلفة والجهد المطلوب لإدارة الشكاوى يدويا أدى إلى الحاجة لتطوير حلول آلية للتعامل مع هذه المشكلة عن طريق استخدام تقنيات التنقيب عن النص لاستبدال الجزء البشري. حيث سيقوم الحل المقترح بالتعامل مع المحتوى العربي الذي يختلف عن الإنجليزية مما يجعل تحليل البيانات مهمة معقدة. حيث أجريت أبحاث قليلة عن الجمل باللغة العربية التي تتطلب معالجة خاصة للنصوص مثل ترتيب الأفعال والتحليل الصرفي.

فإننا نقترح في عملنا وضع حلا جديدا للتغلب على قيود النظام اليدوي , الحل المقترح يتكون من ثلاث مراحل. أولا، نقوم بتحليل محتويات الرسالة النصية، ثم نقوم بتصنيفها بإستخدام خوارزميات تصنيف النصوص وبعدها يقوم النظام تلقائيا بتوجيه السؤال للشخص المناسب من أجل الحصول على الرد. وأيضا نستخدم تقنيات إيجاد النصوص المتشابهة لاقتراح الإجابات تلقائيا. وأخيرا، فإن النظام يستخدم تقنيات تلخيص النصوص لتحديث مكتبة الأسئلة المتكررة .

ونتيجة لذلك، فإن نظام الشكاوى الآلي سيقوم بتحسين نوعية الإجابة على الأسئلة من خلال تسريع العملية وتقليل الوقت والجهد المطلوب. حيث وجدنا أن عملية إدارة الشكاوى بشكل آلي تعمل بكفاءة وفعالية.

وفقا لتحليل النتائج لقد حقق النظام النتائج التالية : بالنسبة لجزئية **تصنيف الشكاوي** ، فقد وجدنا ان خوارزمية SVMs للتصنيف حققت أعلى دقة (74.69٪). أيضا بالنسبة لجزئية **اقتراح الإجابات** ، حصلنا على أفضل اداء (72.45٪) عند درجة التشابه (0.50). اما بالنسبة لجزئية **تلخيص الاسئلة** المتكررة ، حصلنا على أفضل أداء عند استخدام معدل ضغط 0.3، حيث كانت النتيجة 71.56٪.

الكلمات المفتاحية : آلية التغذية الراجعة , آلية الشكاوى , التنقيب عن النص , تصنيف النصوص , تلخيص النصوص.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

TM	Text Mining
TC	Text Classification / Text Categorization
VSM	Vector Space Model
IR	Information Retrieval
CA	Classical Arabic
NB	Naïve Bayes
CNB	Complement Naïve Bayes
NBM	Naïve Bayes Multinomial
DMNB	Discriminative Multinomial Naïve Bayes
SVMs	Support Vector Machines
TF-IDF	Term frequency–inverse document frequency
KNN	K Nearest Neighbors
DT	Decision Trees
ME	Maximum Entropy
NN	Neural Networks
DFE	Discriminative Frequency Estimate

Chapter 1
Introduction

A complaint system is a system that manages the process of how organizations handle, manage, respond and report to client's complaints. The manual categorization of the large number of complaints is extremely difficult, time consuming, error prone, expensive and is often not feasible, which results are dependent on variations experts' judgments. For example, according to large number of records that UNRWA has in its complaints system, it has become increasingly necessary for users to utilize automated system to find the desired information, and to track and analyze their usage patterns. These factors give rise to the necessity of using intelligent systems that can effectively mine for knowledge

To handle the manual system limitations, we propose to build a model that utilized by text processing and mining solutions that could uncover the trends, patterns and relationships inherent in the complaints. The automated complaints system will depend on Text mining techniques to understand the text message and try to categorize the complaints. We will use three text mining techniques: Text categorization, Similarity measures and Text summarization.

1.1 Complaints Mechanism

The complaints mechanism (feedback mechanism) is a mechanism that supports clients complaints, where the clients can send complaints / petitions about any problem that face them , and then the response unit receives the complaints and categorizes these complaints manually based on understanding of the received messages [1].

Therefore, the traditional Complaints systems as UNRWA complaints system depend on manual understanding for the received messages. So there is a human part responsible for all phases of the complaints process. While manually organizing for the large number of complaints is extremely difficult, time consuming, expensive and is often lead to un-satisfaction of the complainant.
The main problem for these systems is the required time for processing the complaint that affects the response time and quality of services for the wholly complaints mechanism, and also sometimes got wrong categorization or forwards the complaint to wrong person. While the time of answering the question affects directly on the user satisfaction.

So to improve the quality of service we need to minimize the processing time by replacing the human parties with automated parties as automatic categorization and answers suggestions.

1.2 UNRWA and its Complaints System

UNRWA is nonprofit organization, works in Gaza to serve refuges in many issues in education, health, employment and food distribution. According to the need to answer the refuges complaints, UNRWA developed a complaints system to enable the refuges to submit complaints in several issues. So the beneficiaries can use it to send complaint to UNRWA management directly through the UNRWA portal, and the system handle all received cases and process them with minting the quality of service (response time and response quality).
There is a unit in UNRWA responsible for managing these requests called Response Unit.
The Response Unit reviews complaints, categorize each complaint after reading the message and forward it to the department to get answer and then fills the reply and sends it back to the complainant.

According to large number of complaints received every day, there is an increase in the effort and time required to process the complaints that affected on the quality of response and

waiting time to get the answer. After analyzing a set of complaints, we found the period between request date and reply date affect on the feedback.
So there is need to improve the quality of services and decrease the cost and required effort.

Our goal of this project is improving the processes that need human efforts and time. The important processes are categorizing Arabic complaints, preparing answers and updating FAQ library using text mining techniques.

1.3 Text Mining

Text Mining is the automatic and semi-automatic extraction of implicit, previously unknown, and potentially useful information and patterns, from a large amount of unstructured textual data, such as natural-language texts [2, 3]. In text mining, each document is represented as a vector, whose dimension is approximately the number of distinct keywords in it, which can be very large. One of the main challenges in text mining is to classify textual data with such high dimensionality. In addition to high dimensionality, text-mining algorithms should also deal with word ambiguities such as pronouns, synonyms, noisy data, spelling mistakes, abbreviations, acronyms and improperly structured text. Text mining algorithms are two types:

- Supervised learning
- Unsupervised learning.

For Example: Support vector machines (SVMs) are a set of supervised learning methods used for classification and regression. Nonnegative matrix factorization is an unsupervised learning method.

We will use three techniques from text mining in our system which are Text categorization, Similarity measures and Text summarization.

1.3.1 Text Categorization (TC) is the task in which texts are categorized into predefined categories based on their contents [4]. For example, if texts are represented as a research paper, categories may represent “Computer Science”, “Mathematics”, “Medicine”, etc. The task of TC has various applications such as automatic email classification, web-page categorization and indexing [5].

1.3.2 Measure of similarity between two documents is therefore the Euclidean distance between their respective representatives points in space. The validity of this measure of “similarity” hypothesizes like documents share many of the same terms [6].

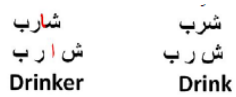
1.3.3 Text Summarization (TS) is the process of identifying the most salient information in a document or set of related documents and conveying it in less space (typically by a factor of five to ten) than the original text. In principle, TS is possible because of the naturally occurring redundancy in text and because important (salient) information is spread unevenly in textual documents.

1.4 Arabic Language

The complaints that we used in our system are written in Arabic language. Arabic language is a semantic language with a composite morphology. The words are categorized as particles, nouns, or verbs. There are 28 letters in Arabic, and the words are formed by linking letters of the alphabet. Letters of the alphabet differ in shape based on their position within the word (i.e. beginning, middle, or end). Unlike most Western languages, Arabic script is written from right to left. Furthermore, proper nouns do not start with capital letters, thus, extracting nouns and proper nouns is a challenging task for machines.

Also, in English, words are formed by attaching prefixes and suffixes to either or both sides of the root [7].In Arabic, additions to the root can be within the root (not only on the word sides) which is called a pattern. This causes a serious issue in stemming Arabic documents because it is hard to differentiate between root particles and affix letters.

For example, for the root "drink" in Arabic, adding the letter "ا" (infix) formed different words such as "drinker" can be formed by adding the letter "ا" (infixes).



Suffixes, prefixes and infixes are categorized based on their use. Similar to other Western languages, there are specific suffixes to convert the word from the singular form to the plural form and others to convert from masculine to feminine [7].

1.5 Research Problem

In any organization, manually organizing large number of complaints is extremely difficult, time consuming, error prone, expensive and is often not feasible, which results are dependent on variations expert's judgments. Also, there are some systems available in English, There is not any available in Arabic.
The response unit in UNRWA receives many requests each day and makes categorization for each request manually based on the contents of the message, followed by forwarding the request to the responsible persons according to its category to get the answers. Thus, it affects the response time and quality of services for the wholly complaints mechanism, and also sometimes got wrong categorization or forwards the complaint to wrong person. While the time of answering the question affects directly on the user satisfaction. So to improve the quality of service we need to minimize the processing time. Therefore, the existing Complaints mechanism systems as UNRWA complaints system depend on manual understanding for the received messages.

1.6 Research Objectives

1.6.1 Main objective

The main objective of this work is to develop an ***automated complaints system*** that uses text mining techniques to manage received complaints written in Arabic, where this system will minimize the human efforts and speed up answering of the complaints that leads to improve quality of the services. To measure the effectiveness of our approach, we will use UNRWA complaints system as a case study.

1.6.2 Specific objectives

The specific objectives of the thesis are:

- Study the current complaints systems
- Analyze the current manual complaints system in UNRWA and understand its limitations.
- Propose a solution by using text mining techniques [Text **categorization** for complaints classifications, **Similarity Measures** for answers suggestions and **Text Summarization** for updating FAQ library].
- Collect data and prepare it to be used to train the new system.
- Build an automated complaint model to manage client's complaints.
- Evaluate the effectiveness of the proposed systems and also evaluate the accuracy of the developed tool and compare it with other complaint systems.

1.7 Importance of the work:

Provide high quality system to manage the clients complaints in UNRWA. Overcame the Limitation of the existing manual complaints systems and quick processing, minimize the required staff for managing the system and minimize the required effort and time of processing the submitted complaints by using automated parties that analyze the Arabic contents .

1.8 Scope and limitations of the work

1.8.1 Scope

- The work focuses on the Arabic language.
- The system uses SQL Server methods for text collections.
- The new model deals with any type of complaints.
- The system will exchange data with other systems.

1.8.2 Limitations

- The new model doesn't answer all received questions automatically, but it answers the questions that are similar to answered questions in the system.
- The system needs human assistant to manage part of its process, so it's not fully automated system.
- We used already built tools in text mining to build the system.
- The system will not guarantee to suggest correct answers all the time because it depends on previous answers that may be wrong.

1.9 Methodology

In this work, as seen in Figure 1.1, we analyze the current complaints tools and study their limitations, and then design the model for the automated complaints system. For each part in the system, we use some text miming techniques and select the best method that achieved the best performance.

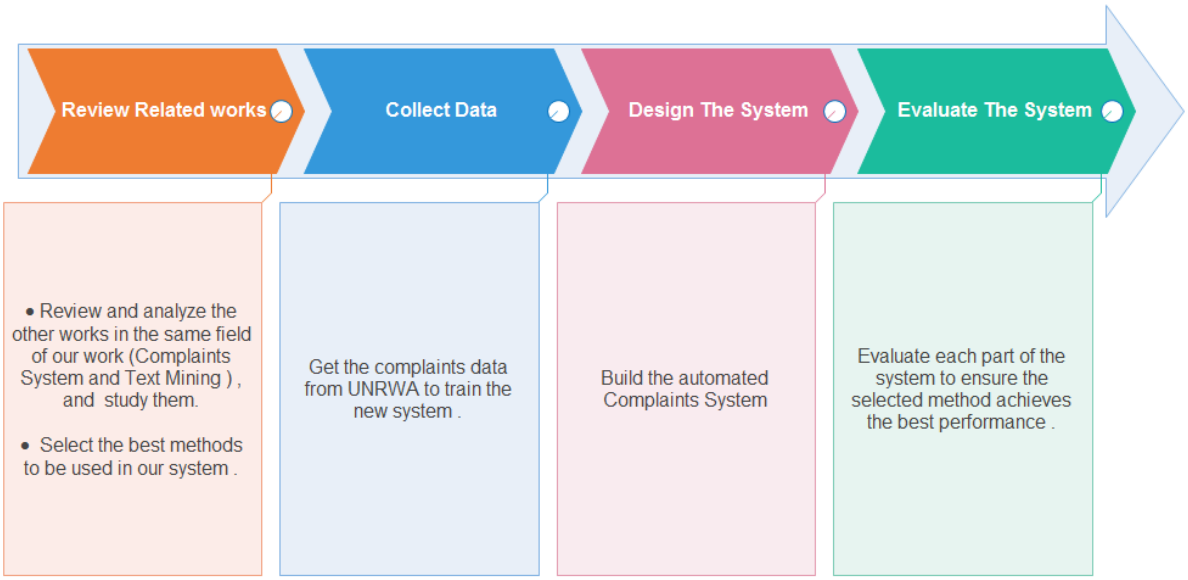


Figure 1.1: Research Methodology

The main steps are:

- **Review related works :**
  Study and analyze the current works about complaints systems and works about applying text mining methods in similar systems, and nominate the best methods to be used in our system.
- **Collect data :**
  Get the data for current complaints system in UNRWA and prepare the data to be used to train the new system (select the important fields, change the format and remove unnecessary data).
- **Design the system :**
  Design the system and implement it by using c# language, the main modules of the system are:
  - **Complaint Analyzer to** read received message details, and understands the meaning of it based on some rules.
  - **Complaint Classifier** to set the complaint category based on message text meaning by using text classification techniques.
  - **Answer Suggestion to** suggest the answer based on previous cases that similar to the current case by using text similarity algorithm.
  - **Complaints summarizer** to summarize set of complaints for selected topic and update the FAQ.
- **Evaluate System :**
  Evaluate each module and select the best method that achieved the highest performance.

1.10 Tools, equipments and methods

- Visual Studio 2010
- SQL Server Database
- Rapid miner - Text Mining library
- Internet connection

1.11 Time Table

Task	Time
Literature survey	5 Weeks (18 Feb – 24 Mar)
Developing the Proposed Model	5 Weeks (25 Mar – 07 May)
Implementation	6 Weeks (08 May – 21 Jun)
Testing and Evaluation	5 Weeks (22 Jun – 26 Jul)
Results and Analysis	5 Weeks (01 Aug – 07 Sep)
Writing the report	9 Weeks (10 Sep – 13 Nov)
	Total 35 Weeks

1.12 Thesis organization

The thesis is composed of six chapters which present theoretical and practical aspects of the subject. Chapter 2 presents Literature Review for theoretical foundation of the research. Chapter 3 presents issues related to applying text mining methods in some real applications and present some researches about complaints management. Chapter 4 presents the experimental setup and describes automated complaints system model that has been chosen to be implemented in the work. Chapter 5 presents the experimental results and discusses the results obtained. Finally, Chapter 6 summarizes the work and outlines possible further extensions to the current work.

Chapter 2
Theoretical Foundation

This chapter introduces the theoretical foundation of the research. It includes the following topics: Types of text mining (Supervised learning and unsupervised learning), definition of unstructured datasets and documents collection, Term weight of text documents, Text categorization (TC), Similarity measure of documents and Text summarization (TS).

2.1 Supervised learning

Supervised learning is a technique in which the algorithm has a target attribute value and uses predictor to learn the predictor and target value relation. Techniques as Support Vector Machine (SVM) method is a supervised learning technique for creating a decision function with a training dataset. The training data consist of pairs of target values and predictor. Each predictor value is tagged with a target value. If the algorithm can predict a categorical value for a target attribute, it is called a classification function. Class is an example of a categorical variable. Positive and negative can be two values of the categorical variable class. Categorical values do not have partial ordering. If the algorithm can predict a numerical value then it is called regression. Numerical values have partial ordering [8].

2.2 Unsupervised learning

Unsupervised learning is a technique in which the algorithm uses only the predictor attributes values without having target attribute values, so the learning task tries to gain some understanding of relevant structure patterns in the data. Each row in a data set represents a point in n-dimensional space and unsupervised learning algorithms investigate the relationship between these various points in n-dimensional space. Examples of unsupervised learning are clustering, density estimation and feature extraction [8].

2.3 Unstructured data sources.

Data for analyzing the Text Mining algorithms can be obtained from various external and/or internal sources. The most important sources of external data are social services with thousands of posts, feedback, comments, etc. Minutes from conversations with customers, e-mails, business documents such as contracts and offers, publications, transcripts of call-centre, descriptions of insurance claims, police notes, open-ended questions in surveys, etc. are examples of internal sources of data [8].

2.4. Document collection

Categories must be predefined before applying the classification process. The categorization is related not only to text. It can be related to video, stock markets, health care etc. Example data collections can be downloaded from the Internet. Alternatively data collections can be created on our own systems or any source of data. This approach however may be a tough task if the data set is going to be big. In such case, the categories should be prepared first and then relevant documents put inside.

In majority of cases the document collections are divided into two sets:

- Training set
- Test set

The training set is used to construct a classifier. The test set is used to evaluate the classifier. Size of the data sets is an important issue that related to the process preparation. Authors in [9] strongly recommend splitting these two sets in proportion 2/3 for the training set and 1/3 for the test set. In some cases the classifier can be overloaded e.g. trained too much.

Obviously in such case such system will work but the trained function will not be able to recognize documents which are not very similar to the ones of the training set. For this reason it is necessary to have a function or functions which would be able to determine if the classifier was trained correctly.
It is impossible to predict when process should be finished during training the classifier. It may lead to complicated situations. If the classifier is undertrained or over-trained it may give wrong results. We can use another document collection working as a validation set to prevent such situation.

To make the learning algorithms brought satisfactory results, the training set should incorporate as many documents as possible. In such cases the learning process slows down but the learned hypotheses usually have better accuracy [10].

2.5 Term weight of text documents

In text mining the document is represented as a vector. The elements as words in the vector reflect the frequency of terms in documents. Table 2.1 represents a document word matrix with frequencies.

Table 2.1: Document word matrix with frequencies

	Word1	Word2	Word3	……	Wordm
Document1	3	1	3		
Document2	1	2	4		
Document3	2	3	0		
Document4	5	0	5		
……..					
Documentn					

In Table 2.1, the numbers in each row represent the term frequencies, tf, of the keywords in documents 1, 2, 3… n.

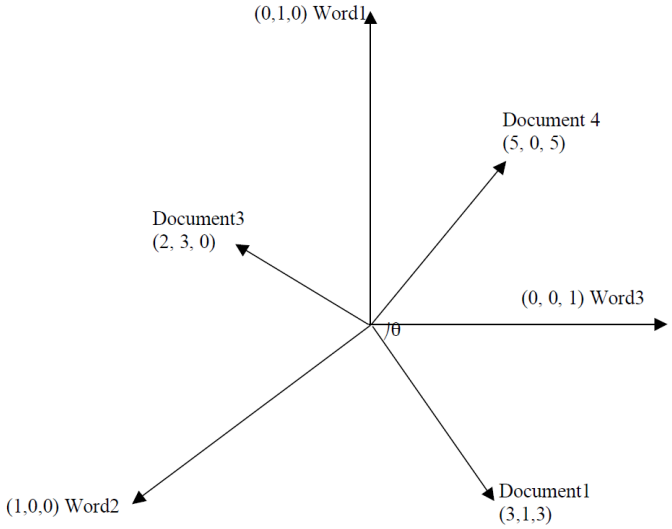


Figure 2.1: Three-dimensional term space

In text mining each word is represented as a dimension and documents are vectors as shown in Figure 2.1. Each word in a document has weights. These weights types can be: Local or global weights. If local weights are used, then term weights are normally expressed as term frequencies, tf. If global weights are used, Inverse Document Frequency, IDF values, gives the weight of a term.

tfi = Frequency of i-th term
dfi /D = Probability of selecting a document containing a queried term from a collection of documents.
dfi = Document frequency or number of documents containing term i
D = Number of documents in a database.
log(D/dfi) = inverse document frequency, IDFi, represents global information.

Table 2.2: Shows Frequency of word "PEN" in documents 1, 2, 3, 4 and 5.

Paragraph	Document 1	Document 2	Document 3	Document 4	Document 5
P1	PEN		PEN		
P2		PEN	PEN		
P3		PEN	PEN		
P4		PEN	PEN		
P5		PEN	PEN		
P6					

In Table 2.2, number of documents D = 5 and document frequency df = 3. Searching the system for 'PEN' word gives an IDF value of, log (D/dfi) = log (5/3) = 0.2218. It is possible to do improve term weighing by multiplying tf values with IDF values, using local and global information. Therefore total weight of a term = tf * IDF. It is referred to as, TF-IDF weighting.

2.6 Text categorization

Text categorization (TC) is the task in which texts documents are categorized into predefined categories based on their contents [4]. For example, if texts are represented as a research paper, categories may represent "Engineering", "Information Technology", "Medicine", etc. The task of TC is used in various applications such as automatic email classification, web-page categorization and indexing [5]. These applications are becoming increasingly important in today's information-oriented society especially with the rapid growth of online information, and therefore TC has become one of the key areas for handling and organizing textual data. As mentioned earlier, the goal of TC is the classification of documents into a fixed number of pre-defined categories in which each document can be in multiple, exactly one, or no category at all.

TC can provide conceptual views of document collections and has important applications in the real world. For example, organizing news stories by subject categories (topics), academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are often indexed from multiple aspects, sorting of files into folder hierarchies, topic identifications, dynamic task-based interests, automatic meta-data organization, text filtering and documents organization for databases and web pages [11,12,13]. Another common application of text categorization is spam filtering, where email messages are classified into the two categories spam and non-spam [14].

Automatic text categorization can significantly reduce the cost of manual categorization , For example , News sites which uses hundreds of expert people to manually categorize their web sites pages where it receives hundreds of pages daily [15, 16].

The main steps for TC task: Text pre-processing, text classification and classifier evaluation. Text pre-processing phase is to make the text documents suitable to train the classifier. Then, the classifier is constructed and tuned using a learning technique against the training data set.

Finally, evaluating the classifier by using some evaluation measures as recall, precision and F-Measure [5].

The **classifier** is usually built based on the content of the training data set, and utilized to predict the category for new document. This type of learning is called supervised where the input data set contains predefine classes / categories and the search for knowledge is restricted with target categories as in figure 2.2.

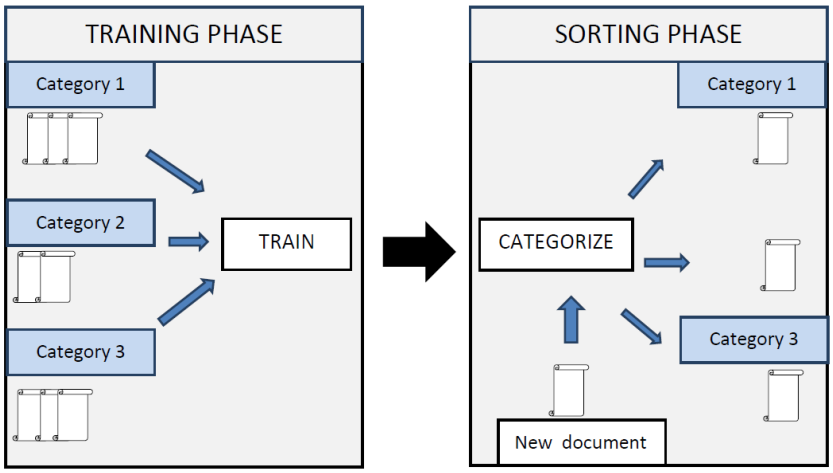


Figure 2.2: Training – Sorting Phase of categorization

The text classification problem is composed of several sub problems, which are the document indexing, the weighting assignment, dimensionality reduction, document clustering, threshold determination and the type of classifiers [13].

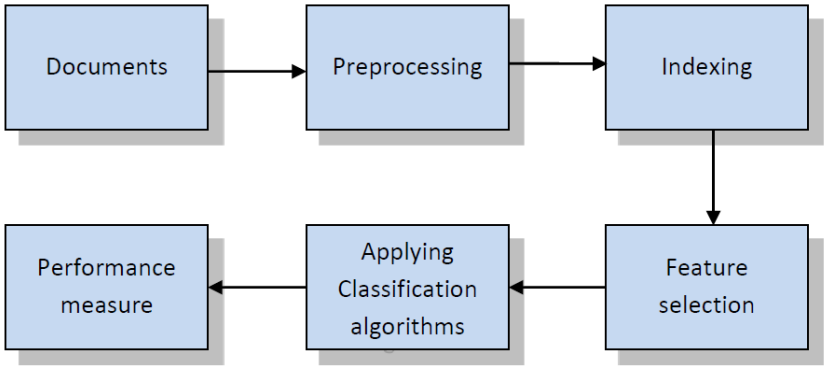


Figure 2.3: Steps of classifying documents

The common methods that used for text classification are : Support Vector Machines (*SVMs*), *K* Nearest Neighbor (*KNN*) , Naïve Bayes (*NB*) , Decision Trees (*DT*) , Maximum Entropy (*ME*) , *N-Grams* ,and Association Rules [17,18].

- **Decision Tree**
  A decision tree is a diagram like tree structure used to determine a course of action or show a statistical probability, where each branch of the decision tree represents a possible decision or occurrence, each internal node denotes a test on an attribute, and each leaf node (or terminal node) holds a class label [18]. The topmost node in a tree is the root node. During tree construction, attribute selection measures are used to select the attribute which best partitions the tuples into distinct classes. The popular attribute selection measures are Information Gain, Gain Ratio, and Gini Index. When decision tree is built, many of the branches may reflect noise or outliers in the training data. Tree pruning attempts to identify and remove such branches to improve classification accuracy on unseen data.

- **Naïve Bayes**
  Bayesian classifier is considered as statistical classifier. The Bayesian classifiers can predict class membership probabilities. Naïve Bayes classifiers are commonly studied in machine learning. The basic idea in NB approaches is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The assumption of word independence is the naïve part of NB methods, i.e. the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. This assumption leads to considering the computation of the NB classifiers far more efficient than the exponential complexity of non-naïve Bayes approaches because it does not use word combinations as predictors [18].

- **k-Nearest Neighbors**
  the k-Nearest Neighbors algorithm is a method used for classification and regression. Nearest neighbor classifiers are based on learning by analogy, that is by comparing a given test object with training objects which are similar to it. The training objects are described by *n* attributes. Each object represents a point in an *n*-dimensional space. In this way, all of the training objects are stored in an *n*-dimensional pattern space. When given an unknown object, a k-nearest neighbor (k-NN) classifier searches the pattern space for the *k* training objects which are closest to the unknown object. These *k* training object s are the *k*-nearest neighbors of the unknown object [17].

- **Support Vector Machine**
  In machine learning, Support vector machines are supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. So a support vector machine (SVM) is an algorithm that uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane where the hyperplane is a “decision boundary” separating the objects of one class from another. The SVM finds this hyperplane using support vectors (“essential” training objects) and margins (defined by the support vectors) [17].

- **Association rule**
  Association rule mining has been extensively studied in the data mining community. It finds interesting association or correlation relationships among a large set of data items. The discovery of interesting association relationships among huge amounts of transaction records can supoort many decision making processes .Since then, association rule mining has been studied and applied in many domains (e.g. network intrusion detection, credit card fraud, genetic data analysis). In every domain, Association rule mining is used to analyze data to identify patterns associating [18].

- **Maximum Entropy**
  The Maximum Entropy (MaxEnt) classifier is similar to a Naive Bayes classifier, except that, rather than allowing each feature to have its say independently, the model uses search-based optimization to find weights for the features that maximize the likelihood of the training data.

  The features you define for a Naive Bayes classifier are easily ported to a MaxEnt setting, but the MaxEnt model can also handle mixtures of boolean, integer, and real-valued features [19].

- **N-GRAMS**
  An N-gram [20] is an N-character slice of a string. The N-Gram method is language independent and works well in the case of noisy-text (text that contains typographical errors). It used for text classification. The trigrams of a string or token is a set of continuous 3-letter slices of the string. For example, the tri-grams for the word المودعين are: ال ,لم ,لمو ,مود , ودع , دعي ,عين . In general, a word of length w has w-2 tri-grams. According to Zipf's law [21] .

Text processing includes tokenizing string to words, normalizing tokenized words, remove predefined set of words (stopwords), morphological analysis, and finally term weighting [12, 14].

The main phases of building a text classification system which involve compiling and labeling text documents in corpus, selecting a set of features to represent text documents in a defined set classes or categories (structuring text data), and finally choosing a suitable classifier to be trained and tested using the compiled corpus . The constructed classifier system then can be used to classify new (unlabeled) text documents as shown in Figure 2.4.

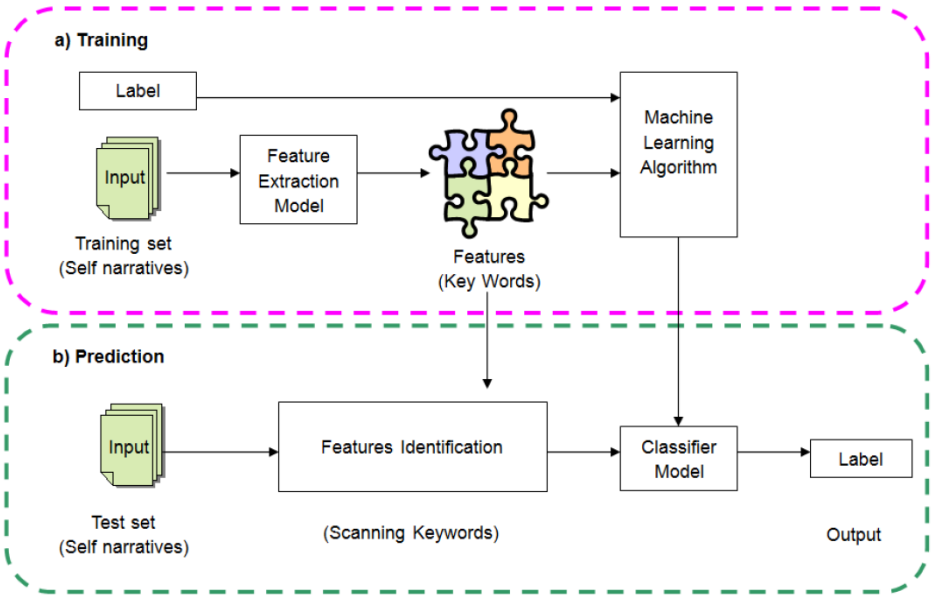


Figure 2.4: Classification Process

2.7 Similarity measure of documents

Measure of similarity between two documents is the Euclidean distance between their respective representative's points in space. The validity of this measure of "similarity" hypothesizes like documents share many of the same terms. If two documents describe similar topics, employing nearly the same keywords, these texts are similar and their similarity measure should be high. Usually dot product used to represent similarity of the documents.

The Euclidean distances of the two documents are used to normalize the dot product, we divide it by represented respectively by Doc1 and Doc2; i.e., <Doc1, Doc2> / (|Doc1||Doc2|).

Here |Doc1|, |Doc2| represent magnitudes of vectors Doc1 and Doc2 respectively and <Doc1, Doc2> is the dot product of the vectors Doc1 and Doc2. This ratio defines the cosine angle between the vectors, with values between 0 and 1 [16]. This is called cosine similarity.

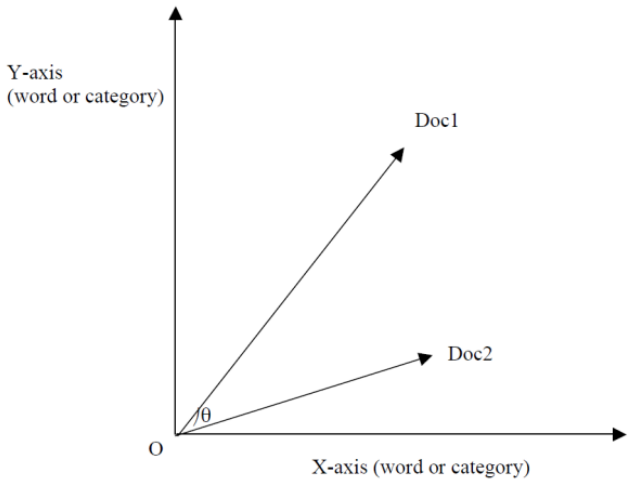


Figure 2.5: Straight lines in 2-Dimensional space represent Euclidean distances of document vectors Doc1 and Doc2, with origin O.

Cos θ = Similarity of the vectors Doc1 and Doc2 = <Doc1, Doc2> / |Doc1||Doc2| As the angle between the vectors, θ, decreases, the cosine angle approaches to 1, meaning that the two document vectors are getting closer, and the similarity of the vectors increases [22].

2.8 Text summarization

Text summarization (TS) is the process of identifying the most important information in a text document or set of related documents and conveying it in less space (typically by a factor of five to ten) than the original text. In principle, TS is possible because of the naturally occurring redundancy in text and because important (salient) information is spread unevenly in textual documents. Identifying the redundancy is a challenge that hasn't been fully resolved yet.

There is no single definition for salience and redundancy given that different users of summaries may have different backgrounds, tasks, and preferences. Salience also depends on the structure of the source documents. Since information that the user already knows should not be included in a summary and at the same time information that is salient for one user may not be for another, it is very difficult to achieve consistent judgments about summary quality from human judges. For this reason, it's difficult to evaluate (and hence, improve) automatic summarization [24].

Most existing summarizers work in an extractive fashion, selecting sentences of the input documents that are believed to be more salient. Non-extractive summarization includes dynamic reformulation of the extracted content, involving a deeper understanding of the input text, and is therefore limited to small domains. Query-based summaries are produced in reference to a user query (e.g., summarize a document about an international summit focusing only on the issues related to the environment) while generic summaries attempt to identify salient information in text without the context of a query. The difference between single- and multi-document summarization (SDS and MDS) is quite obvious; however some of the types of problems that occur in MDS are qualitatively different from the ones observed in SDS: e.g., addressing redundancy across information sources and dealing with contradictory and complementary information. No true multilingual summarization systems exist yet, however, cross-lingual approaches have been applied successfully [23].

A number of evaluation techniques for summarization have been developed. They are typically classified into two categories. Intrinsic measures attempt to quantify the similarity of a summary with one or more model summaries produced by humans. Intrinsic measures include Precision, Recall, Sentence Overlap, Kappa, and Relative Utility. All of these metrics assume that summaries have been produced in an extractive fashion. Extrinsic measures include using the summaries for a task, e.g., document retrieval, question answering, or text classification [24].

Traditionally, summarization has been mostly applied to two genres of text: scientific papers and news stories. These genres are distinguished by a high level of stereotypical structure. In both these domains, simply choosing the first few sentences of a text or texts provides a baseline that few systems can better and none can better by much. Attempts to summarize other texts, e.g., fiction or email, have been somewhat less successful.

Recently, summarization researchers have also investigated methods of text simplification (or compression). Typically, these methods apply to a single sentence at a time. Simple methods include dropping unimportant words (determiners, adverbs). Complex methods involve reorganizing the syntactic parse tree of the sentence to remove sections or to rephrase units in shorter form. Language modeling approaches in TS have mostly focused on this method [23].

2.9 Summary:

This chapter has described popular text classification algorithms. We will use the most common classification methods which are SVM, KNN with Cosine similarity, Naïve Bayes and Decision Tree methods to classify the new complaints (For Complaints Categorization part) , and select the best of them to be in our system as the Complaints Classifier. And also it described representation of documents as vectors in text mining and how Measure of similarity between two documents. Also it described text summarization process for identifying the most salient information in a document or set of related documents.

Chapter 3
State of the Art and Related Works

In this chapter, we introduce some previous works that are relevant to our research, and point out their limitations. And also we introduce similar systems that applied to different domains such as disease recognition and emails filtering, and some previous works that handles categorization problems, text similarity and text summarization by using different text mining techniques.

3.1 State of the art

Complaints are an important way for the management of an organization to be accountable to the public, as well as providing valuable prompts to review agency performance and the conduct of people that work within and for it [25].

A complaint is an "expression of dissatisfaction made to an organization, related to its services, or the complaints handling process itself, where a response or resolution is explicitly or implicitly expected" (as defined by the *Australian Complaint Handling Standard ISO AS 10002-2006*) [25].

An effective complaint handling system provides three key benefits to agencies:

- It resolves issues raised by a dissatisfied person in a timely and cost-effective way.
- It provides information which can lead to improvements in service delivery.
- Where complaints are handled properly, a good system can improve the reputation of an organization and strengthen public confidence in an organization

There are needs for both organizations and the benefiters that trying to meet through the complaints systems [25].

The organization needs are:

- A user friendly system for accepting feedback.
- Clear delegations & procedures for staff to deal with complaints and provide remedies.
- A recording system to capture complaint data.
- To use complaint data to identify problems and trends.
- To improve service delivery in identified areas.

The beneficiaries needs are:

- A user friendly complaints system.
- To be heard and understood.
- To be respected.
- An explanation.
- An apology.
- Actions as soon as possible.

To assist the organization to develop ***effective internal complaints processes***, the Ombudsman group has developed a suite of guidelines as follows [25]:

- Effective handling of complaints made to organization.
- Having Complaint handling systems checklist, so organizations can use this checklist.
- In conjunction with the Ombudsman's guidelines for effective handling of complaints made to organization to assess their complaint handling system against the key features required for an effective system.
- Making complaint handling system accessible to make the complaint handling system accessible to all members of the community.
- Guidance for Complaint Handling Officers to offer assistance to Complaint Handling Officers in handling and investigating complaints made to organization.
- Good record keeping to explain who is responsible for answered complaints.

There are two types of the complaints systems which as:

- **Staff complaints system:**
  To manage the complaints of the organization staff members that face them in the work as complaints related to HR issues, Finance issues and others.
- **Public community complaints system :**
  To manage the complaints of public community that face them in many issues related to organization services.

3.2 Automatic complaint system

There are some researches in ***automatic complaint systems*** such as:

Chen in [26] tried to measure the efficiency of the complaints system in auction store by using text mining techniques, they have chosen questions that lie under category "finance" and applied text mining methods to analyze these questions. They used text mining to build the semantic network and topic to learn the reason for the problems in the complaints processing. He used TextAnalyst tool to analyze thousands of compliments from consumers in auction store. As a result he discovered the behavior types of the complainers and some problems in the categorization. But the limitation of his study is the accuracy of the results is not obvious, to use only classification and it is for English language.

Francis in [27] tried to discuss the National Health Service (NHS) complaints system and to list its limitations to make public services better and lead the way to make the complaints system better. They observe an increase in complaints where the NHS had failed to acknowledge mistakes or provide an appropriate solution when things go wrong. They found individuals were unhappy with the way their complaints were handled by the NHS. The results showed that 19% of them received Poor explanation and 7% unnecessary delays, 6% Factual errors in response to complaint and 3% Communication with complainant was unhelpful and ineffective, while the Lack of the available information and also no existing of the procedures to handle the complaints leaded to these problems. As a result they agreed that poor complaints system has a negative impact on the patients and others who seek to use it. Inadequate responses cause distress. The limitation of this work is failing to discover the reasons for wrong answers and delay factors, and didn't introduce solutions to overcome these problems.

Himmel and Reincke in [28] developed a scoring procedure to automatically classify lay requests to an internet medical forum about involuntary childlessness. The requests should be classified according to their subject matter (32 categories) and the sender's expectation (6 categories). Their text mining approach comprised the following steps: a large start list of relevant words and the calculation of the Cramer's V statistic for the association between relevant words and the 38 categories. To find the most nearest neighbors, they applied a formula, which gave high weight to singular value decompositions (SVDs). Also they considered the automatically classified subject matter of this 'new' request and to a lesser degree the sender's expectation.
As a result, the proposed approach precision and recall was above 80% in nearly for all categories.
One important limitation must be mentioned: although matches to a new request had to correspond with respect to the subject matter and the expectation and should be close to each other with regard to the SVDs, this does not protect them against mismatches due to false classifications. In this case, the experts' answers from former requests cannot meet the sender's information needs on principle. And visitors to an expert health forum will be disappointed if they do not receive a more adequate and individual answer in due time.

Urdziková and Jakábová in [29] tried to explore the nature of complaint satisfaction with particular emphasis on the qualities and behaviors that customers value during personal complaint handling service encounters. They found the reasons of dissatisfaction due to delay in answering the complaints and also wrong answers. As a result they suggested some procedures to improve the quality of service in complaints management, such as try to recruit individuals who have strong listening, questioning, and verbal skills as complaining customers take these skills for granted. Also, It improves the capability to analyze complaints messages and interpret their correct meanings. Also need to increase the knowledge base.

3.3Text mining:

There are various types of text mining techniques (text classification, text similarity and text summarization). In the following some works that depends on these techniques:

3.3.1 Classification

Anirban in [30] developed medical diagnosis tool for classifying patient records and reveal important vocabularies that characterize nursing and pathology records. They proposed a Minimum spanning tree algorithm to develop k-clusters of training data related to different liver diseases which are validated using Silhouette coefficient. A text classification algorithm is developed using cluster centers as training samples which uses a similarity measure to classify the categorical data. As a result the clusters were validated using silhouette coefficient. It is observed that an accuracy of 89% is reached in the proposed algorithm which is much superior to state of art k -NN algorithm for text categorization.

Sharef and Kasmiran in [31] used classification methods in classifying the incidents events ,They introduced fuzzy grammar as a technique for building text classifier and compare the performance of it with other machine learning methods such as support vector machine, statistic, nearest neighbor and boosting. The results have shown that fuzzy grammar has gotten promising results among the other benchmark machine learning methods. Where fuzzy grammar has obtained around 84% of F-score and has the highest precision (93.2%) in categorizing texts on bombing although lowest precision in categorizing texts on armed attack

Mesleh in [4] has implemented the SVM algorithm with the uses Chi square method as a feature selection method to classify Arabic documents. He has used an in-house collected corpus from online Arabic newspaper archives, including several news sites as Al-Jazeera, Al- Nahar, Alhayat and Al-Ahram, as well as a few other specialized websites. The collected corpus contains 1445 documents that vary in length. These documents fall into nine classification categories. The results showed that the SVM algorithm with the Chi-square method has outperformed Naïve Bayes and the KNN classifiers in term of F-measure, but rule based approaches have poor recall.

Hall in [9] proposed a system as an automated categorizer for email to try to eliminate the large amounts of manual email categorization that is currently done by many users. The categorization approach is derived from an instanced-based learning method that explores conditional probabilities of particular words. The results showed the Precision was 65% while recall was 17%. So rule based approaches have poor recall and require a time consuming job of building rules manually.

Harrag et al. in [32] used method to improve Arabic text classification by feature selection based on hybrid approach. he used decision tree algorithm and reported classification accuracy of 93% for scientific corpus, and 91% for literary corpus. Harrag collected 2 corpora; the first one is from the scientific encyclopedia "Do You Know" (هل تعرف). It contains 373 documents belonging to 1 of 8 categories (innovations, geography, sport, famous men, religious, history, human body, and cosmology), each category has 35 documents. The second corpus is collected from Hadith encyclopedia (موسوعة الحديث الشريف) from ―the seven pens (الاقلام السبعة). It contains 435 documents belonging to 14 categories.

Al-Shalabi et al. in [33] applied KNN on Arabic text; they used TF-IDF as a weighting scheme and got accuracy of 95%. They also applied stemming and feature selection. The authors reported in their paper the problem of lacking freely publically availability of Arabic 25 corpus. He collected a corpus from newspapers (Al-Jazeera, An-Nahar, Al-Hayat, Al-Ahram, and Ad-Dostor) and from Arabic Agriculture Organization website. The corpus consists of 621 documents belonging to 1of 6 categories (politics 111, economic 179, sport 96, health and medicine 114, health and cancer 27, agriculture 100). They preprocessed the corpus by applying stopwords removal and light stemming.

There are several studies compare classification algorithms on Arabic text, Hmeidi , Hawashin and El-Qawasmeh in [34] compared KNN and SVM for Arabic text classification. They used full word features and considered tf-idf as the weighting method for feature selection, and CHI statistics for ranking metrics. They showed that both SVM and KNN have superior performance, and SVM has better accuracy and time. Authors collected documents from online newspaper (Al-Ra'i and Ad-Dostor), They collected 2206 documents for training and 29 documents for testing. The collected documents belong to one of two categories (sport and economic).

In the work [35] about re-examination methods in the field of text categorization, Yiming Yang and Xin Liu examined five different classifiers: the k-Nearest Neighbours, the Neural Networks, the Last-squares linear fit the Naive Bayes classifier and the Support Vector Machine. Document collection chosen for test is Routers-21578 corpora. All unlabelled documents were eliminated from this corpus. Each category included at least one document in

the training set as well as in the test set used for examination. Though the selection of learning collection was carried out in full compliance with the supervised approach, the process resulted in 90 categories in the training set and test. 82% of the categories had less than 100 documents and 33% had less than 10 44 documents. Evaluation of the effectiveness of the system was provided by using recall precision and F1 measure.

Lam, Ruiz and Srinivasan in [36], investigated whether automatic categorization will have better retrieval performance than that achieved using manual categorization applied to medical documents (Lam, Ruiz, Srinivasan). They analyzed the retrieval performance on test queries to gain insights on the interaction of their categorizer and text retrieval.

The first part of their work dealt with automatic categorization including a category-extraction process. For their test documents they use a corpus of medical documents from the MEDLINE database that is referred to as the HERSH corpus. The authors ran a series of experiments on parameter selection to provide a metric and categorization results. Their results are broken down into category and document 4 perspectives. The category perspective results are related to sizes of categories ranging from 10 to 60 categories. Three different parameters were tested: C0, C35 and C50. C0 used all manually assigned categories that existed in the training set and test set. C35 and C50 limit the number of categories to those that have a document frequency greater than 35 or 50 per category. The document frequency is the number of documents that a specific category is assigned to. The F1 score is a weighted combination of recall and precision, with the scores being averaged to determine a mean. Their results for parameter selection can be seen in Table 3.1.

Table 3.1. Results for Parameter Selection in (Lam, Ruiz, Srinivasan).

Run	Parameter selection based on training set		N	M
	# of categories	F1 score		
C0	641	0.258	5	50
C35	58	0.468	5	40
C50	43	0.509	30	20

The results indicate that as the frequency threshold on the category set increases, the mean F1 score improves. N represents the number of documents while M was the number of categories.

3.3.2 Text Similarity

Hoi and Lyu in [37] compared four similarity measures on a collection of Yahoo! News pages. they extended the experiments by including the averaged KL divergence.

They found that the performance of the cosine similarity, Jaccard correlation and Pearson's coefficient are very close, and are significantly better than the Euclidean distance measure. This measure was more frequently used to assess the similarity between words, especially for such applications as word sense disambiguation. Information theoretic clustering algorithms such as the Information Bottleneck method rely on this measure and have shown considerable improvement in overall performance.

Wilson and Martinez in [38] performed a detailed study of heterogeneous distance functions (for categorical and continuous attributes) for instance based learning. The measures in their study are based upon a supervised approach where each data instance has class information in addition to a set of categorical/continuous attributes. There have been a number of new data mining techniques for categorical data that have been proposed recently. Some of them use notions of similarity which are neighborhood-based or incorporate the similarity computation into the learning algorithm .These measures are useful to compute the neighborhood of a point and neighborhood-based measures but not for calculating similarity between a pair of data instances.

Thabtah and Alzubaidi in [39] applied graph for representing the structure of the text as well as the relationship between sentences of the document. Sentences in documents are presented as nodes. The edges between nodes illustrate connections between sentences. These connections are introduced by similarity relation between contents. The similarity between two sentences is calculated and each sentence is scored. All the scores for one sentence are combined to form a final score for each sentence. When the graph is processed, the sentences are categorized by their scores and sentences in higher orders are chosen for final summary.

Inouye in [40] developed a hybrid TF-IDF algorithm. The idea of the algorithm is to assign each sentence within a document a weight that reflects the sentence's saliency within the document. The sentences are ordered by their weights from which the top sentences with the most weight are chosen as the summary.

In order to avoid redundancy, the algorithm selects the next top tweet and checks it to make sure that it does not have a similarity above a given threshold with any of the other previously selected tweets because the top most weighted tweets may be very similar. Another method in [41] collects a set of Twitter posts, clusters the tweets into a number of clusters based on a similarity measure and then summarizes each cluster by picking the most weighted post as determined by TF-IDF algorithm.

3.3.3 Summarization

Suzuki in [42] proposed a SumBasic algorithm for document summarization. In the system, words that occur more frequently across documents have higher probability of being selected for human created multi-document summaries than words that occur less frequently.

Ma, Yu and Liang in [23] developed multi-document summarization system for the web context. The system is useful in combining information from multiple sources. Information may have to be extracted from many different articles and pieced together to form a comprehensive and coherent summary. One major difference between single document summarization and multi-document summarization is the potential redundancy that comes from using many source texts. The solution presented is based on clustering the important sentences picked out from the various source texts and using only a representative sentence from each cluster.

Erkan and Radev [24] developed a LexRank algorithm for computing the relative importance of sentences or other textual units in a document or a set of documents. It creates an adjacency matrix among the textual units and then computes the stationary distribution considering it to be a Markov chain.

Hassel and Dalianis [43] developed automatic text summarizer called SweSum. It summarizes news text in HTML/text format on the WWW. During the summarization 5-10 key words - a mini summary is produced. Accuracy 84% at 40% summary of news with an average original length of 181 words. SweSum is available for Swedish, Danish, Norwegian, English, Spanish, French, Italian, Greek, Farsi (Persian) and German texts. SweSum is based on statistical, linguistic and heuristic methods. The system calculates the frequency of the key words in the text, in which sentences they appeared, and the location of these sentences in the text. It considers if the text is tagged with bold text tag, first paragraph tag or numerical values.

Douzidia in [30] developed the summarizer, Lakhas, by using extracting techniques to produce ten words summaries of a new article. Lakhas first summarizes the original Arabic document and then applies Machine Translation (MT), translating the summary into English. These systems support the single document summarization.

Sobh1, Darwish and Fayek in [32] integrated Bayesian and Genetic Programming (GP) classification methods in an optimized way to extract the summary sentences. The system is trainable and uses manually labeled corpus. Features for each sentence are extracted based on Arabic morphological analysis and part of speech tags in addition to simple position and counting methods. Initial set of features is examined and reduced to an optimized and discriminative subset of features. Given human generated summaries, the system is evaluated in terms of recall, precision and F-measure. It is a concept- based summarizer system that takes a bag-of-words representing a certain concept as the input to the system.

Leskovec et al. in [44] presented a method for summarizing document by creating a semantic graph of the original document and identifying the substructure of such a graph that can be used to extract sentences for a document summary. First, the method starts with deep syntactic analysis of the text and for each sentence; it extracts the logical form triples. After this step, it applies cross-sentence pronoun resolution, co-reference resolution, and semantic normalization to refine the set of triples and merge them into a semantic graph. This procedure is applied to both documents and corresponding summary extracts. In the evaluation phase, the method achieved an average recall of 75% and precision of 30% when compared with human summarization.

3.4 Summary

In this chapter we introduced some related works including works about complaints systems and works about applying text mining techniques in some fields including classification issue such as classifying emails and spam filtering by using some methods SVM, KNN and Naïve Bayes. From works about complaints systems, we found some limitations as: most of these systems depend on manual processing for complaints (review and classify complaints) that lead to some problems as delay in answering questions and wrong classification and inadequate answers.

Also this chapter included works about finding text similarity by using set of text mining methods as cosine similarity, Jaccard correlation and Levenshtein distance similarity algorithm .Also we discussed works about using text summarization techniques to build multi-document summarization systems by using LexRank, Lakhas and Centroid-based summarization algorithm.

Chapter 4
Proposed Complaints System

This chapter introduces our work phases. It includes the following sections: data acquisition, complaint workflow, text pre-processing steps, applying Rapid miner, complaints classification module, answers suggestion part, complaints summarization part and evaluation methods. In this work, we used some text mining techniques to construct an automated complaints system by using the UNRWA data. Figure 4.1 depicts the methodology steps. The first step: read received message and apply text processing steps to prepare data for manipulation and then go to next phases as in the workflow below.

Figure 4.1: Complaints System

- **The initial parts of our system are:**
  - **Requests Receiver:** for all requests, receive the request and forward it to the request analyzer.
  - **Requests Analyzer:** read each message details, and understand the meaning of it based on some rules and then set the request category based on message text meaning by using text mining techniques.
  - **Requests Dispatchers:** After categorizing the request under specific category, forward the request to the desired destination to process the request.
  - **Answer Suggestion:** In some cases, the system will suggest the answer based on previous cases that are similar to the current case by using text similarity algorithms as KNN algorithm.
  - Also the system includes additional feature to update the FAQ library with the most asked questions by using **summarization** techniques.

4.1 Data Acquisition

We used the UNRWA dataset for its complaints system that contains thousands of Arabic text messages of different lengths that belong to about 14 different categories .The data collected from Jan 2011 to Sep 2013. A total of 12,690 complaints were used to train and test our system. The data set contains 14 classes that describe the groups of complaints as ***finance class*** for complaints of financial problems, ***HR class*** for employments and hr complaints ***emergency class*** for food distribution problems, ***education class*** for educational problems**,** ***engineering class*** for housing problems and ***relief services class*** for refuges problems.

4.2 The Complaint workflow

The ***complaint workflow*** contains steps as shown in Figure 4.2: The first step is submitting complaint by the complainant after filling the complaint details. Then ***complaints unit*** receives the requests and checks if there are similar cases exists by using similarity techniques of text mining. If yes it will select the answer and send it back to the complainant. Otherwise, the system will categorize the complaint by using text mining techniques based on the message contents understanding, and then forwards the complaint to the right person to get the answer. The ***specialist in the department*** will receive a notification regarding new complaint is received, and then specialist will prepare the answer and send it to complaints unit.

The final step is reviewing the answer by the complaints unit and then sending it back to the complainant. The complainant will receive a message contains the answer and fills the feedback.

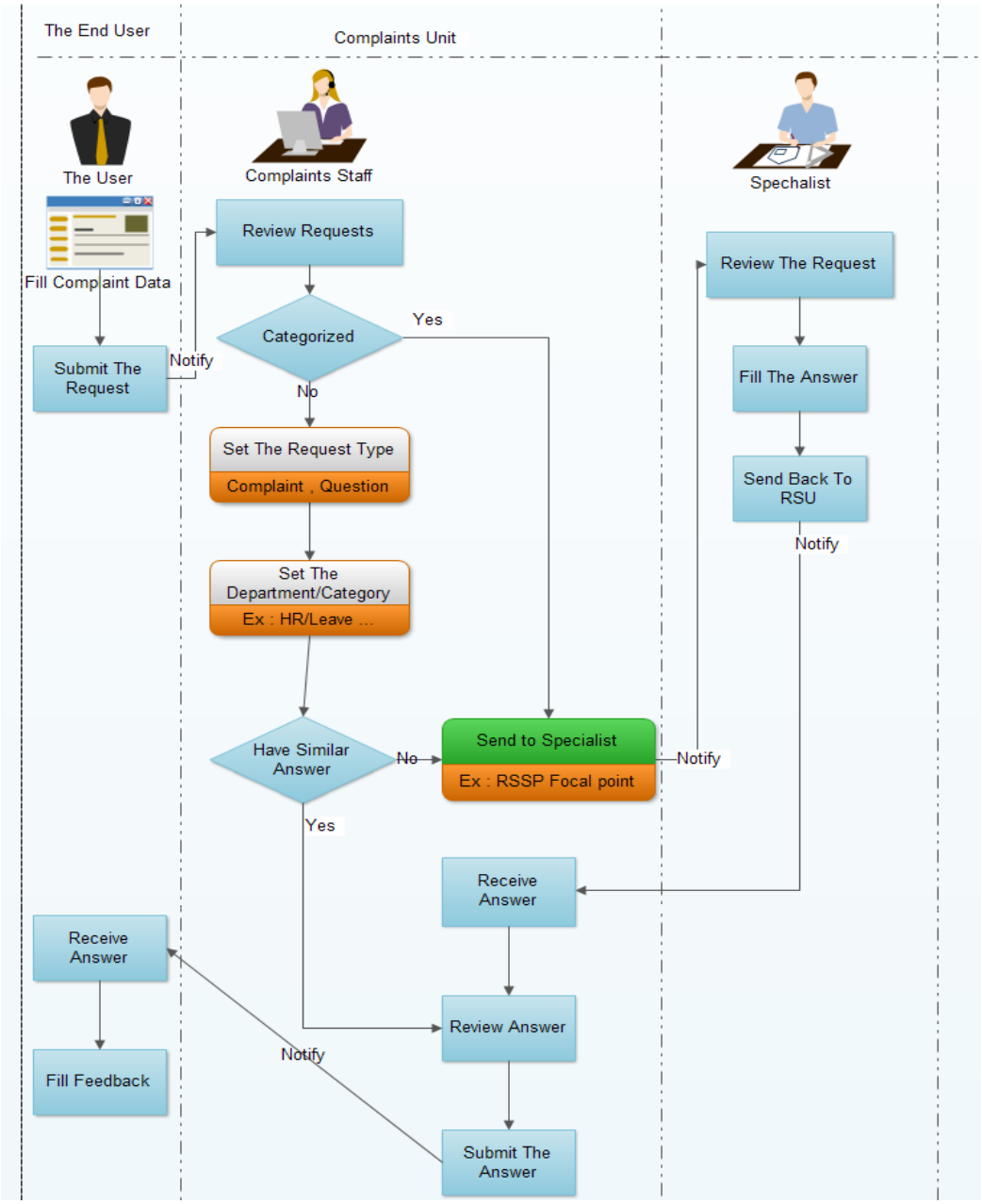


Figure 4.2: The Complaint work flow

4.3 Text pre-processing steps:

To use text mining we need to prepare our data to be ready for applying the mining methods. We aim to transform the Arabic text documents to a form that is suitable for the classification data mining algorithms. As shown in Figure 4.3 preprocessing includes the following steps:

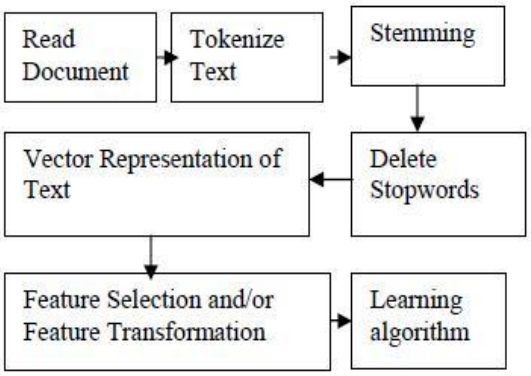


Figure 4.3: Preprocessing Process

4.3.1 Tokenization

The process of breaking a stream of text up into tokens that is words, phrases, symbols, or other meaningful elements is called Tokenization where the list of tokens is input to the next processing of text classification. Generally, tokenization occurs at the word level. Nevertheless, it is not easy to define the meaning of the "word". Where a tokenize process responds on simple heuristics, for instance: All contiguous strings of alphabetic characters are part of one token; similarly with numbers. Tokens are divided by whitespace characters, like a space or line break, or by punctuation characters. Punctuation and whitespace may or may not be added in the resulting list of tokens. In languages like Arabic still tokenization is not easy. Some ways to mention this problem are by improving more complex heuristics, querying a table of common collocations, or fitting the tokens to a language model that identifies collocations in a next processing [38].

4.3.2 Stemming

Stemming is the process of removing affixes (prefixes and suffixes) from features. This process is used to reduce the number of features in the feature space and improve the performance of the classifier when the different forms of features are stemmed into a single feature. Stemming usually used to convert words to root form; it dramatically reduces the complexity of Arabic language morphology by reducing the number of feature / keywords in corpora. For example: (درس, يدرس, دراسة), from the above example, the set of features is conflated into a single feature [7]. There are two types of stemming: root and light stemming. ***Stemming*** reduces words to their stems [45]. ***Light stemming***, in contrast, removes common affixes from words without reducing them to their stems.

We used light stemming in our system because most experiments in Arabic found that light stemming gives more accurate results than root stemming [45].

4.3.3 Stop word removal

This phase includes stop-word removal. **These stop words** can be classified into three types:

- **Frequent Words:** These words or characters are occurring more frequently in the text like common pronouns as (هي، هؤلاء، هم، هن) and some particles such as (ماذا , لماذا) .
- **Words with no particular meaning:** These words are not important words that appear in the context of text without indication to particular information about the text, these words like. (النظر، بغض بالذكر، بالرغم، الجدير، بالإضافة، بالنسبة،).
- **General Words and Numeral:** This type describes some general words likes (days, month, month name, day name, weeks .etc (and some numeral words such as (الاول، الثاني الاولى، الأول،) and many other words.

Typically in computing, stop words are filtered out prior to the processing of natural language data (text) which is managed by man but not a machine. A prepared list of stop words do not exist which can be used by every tool. Though any stop word list is used by any tool in order to support the phrase search the list is ignored.

Any group of words can be selected as the stop words for a particular cause. For a few search machines, these is a list of common words, short function words, like the, is, at, which and on that create problems in performing text mining phrases that consist them. Therefore it is needed to eliminate stop words contains lexical words from phrases to raise performance. Since the sequence of words is called a document. Thus every document is generally denoted by an array of words. The group of all the words of a training group is called vocabulary, or feature set [38].

4.3.4 Vector representation of the documents

Vector representation of the documents is an algebraic model for representing text documents and any objects as set of identifiers vectors, for example, index terms which will be utilized in information filtering, information retrieval, indexing and relevancy rankings where its primary use is in the SMART Information Retrieval System.

A sequence of words is called a text document [32]. Thus every document is generally denoted by an array of words. The group of all the words of a training group is called vocabulary, or feature set. Thus a document can be produced by a binary vector, assigning the value 1 if the document includes the feature-word or 0 if there is no word in the document. There are many types of representation , the most common is TF-IDF weight (term frequency–inverse document frequency) which is a weighting scheme that often used in the vector space model together with cosine similarity to determine the similarity between two documents ,

The TF-IDF is a weight often used in information retrieval and text mining. This weight is a statistical measure used to identify the importance of the word in document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus [7].

4.3.5 Feature Selection and Transformation

A feature selection method is used to decrease of the dimensionality of the dataset by eliminating features that are not related for the classification [43]. The transformation procedure is explained for presenting a number of benefits, involving tiny dataset size, tiny computational needs for the text categorization algorithms (especially those that do not scale well with the feature set size) and comfortable shrinking of the search space. The goal is to reduce the curse of dimensionality to yield developed classification perfection.

The other advantage of feature selection is its quality to decrease over fitting, i.e. the phenomenon by which a classifier is tuned also to the contingent characteristics of the training data rather than the constitutive characteristics of the categories, and therefore, to augment generalization. Feature Transformation differs considerably from Feature Selection approaches, but like them its aim is to decrease the feature set volume. The approach does not weight terms in order to neglect the lower weighted but compacts the vocabulary based on feature concurrencies [32].

4.4 10-Fold Cross Validations

When we have one dataset with the samples having predefined class for each data point, we can split this dataset into training and testing portion. The training portion is used to build a model of the dataset, and the testing version is used to test that model. We'll want to split the dataset multiple times at random places and then average the results.

Most common is **10-fold cross validation**. This means we choose 90% of the data to be the training set, and 10% to be the testing set. We evaluate the precision/recall/etc. with this split, then choose a different 90/10 split and do it again. Because there are 10 possible splits, we do it 10 times and average 10 results. Below figure 4.4 illustrates k-fold cross validation.

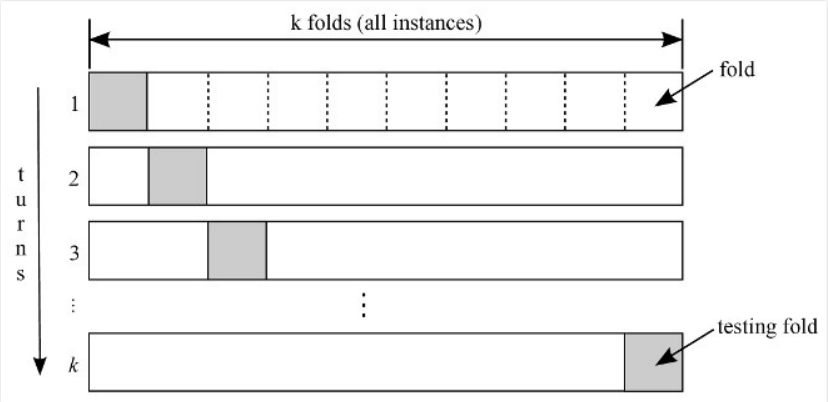


Figure 4.4: Fold Cross Validation

4.5 Using Rapid miner:

We used Rapid miner as shown in Figure 4.5 to perform **text pre-processing**:

a) Tokenization

We have applied **Tokenize** / Rapid miner on the input complaints dataset to break stream of text into list of tokens.

b) Stemming

We have applied **Stem (Arabic, Light)** / Rapid miner to reduce the number of feature / keywords by removing affixes (prefixes and suffixes) from input features.

c) Stop word removal

We have applied Filter **Stopwords (Arabic)** / Rapid miner to remove frequent words and words with no particular meaning.

d) Vector representation of the documents

We have used **TF-IDF** for vector creation / Rapid miner to represent text documents as set of identifiers vectors.

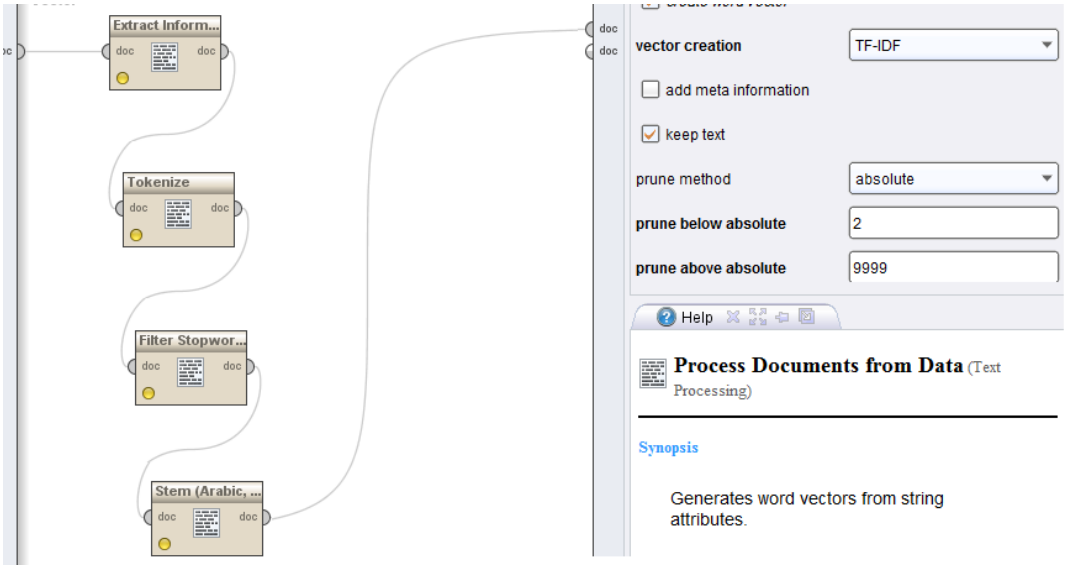


Figure 4.5: Text Pre-Processing

The result: Converting ***Complaints text messages*** to ***Word List*** as in figure 4.6.

WordList (Process Documents from Data) | ExampleSet (//Local Repository/data/data_ar)

Word	Attribute Name	Total Occurences	Document Occurences
ان	?	1727	1396
راقب	?	1376	1111
اجاز	?	1358	865
انا	?	1322	1148
موظف	?	1316	1108
شهر	?	1221	928
سلام	?	1129	1121
عليكم	?	1112	1104
ارجو	?	1108	1041
مدرس	?	1037	763
شكرا	?	955	945
عمل	?	930	757
له	?	890	779
معلم	?	878	678
لم	?	802	712
شكر	?	776	760
طلب	?	753	608
وكال	?	737	637
عام	?	727	590
خصم	?	717	496

Figure 4.6: Resulted Word List

The system includes set of text mining techniques to make the system automated, the system analyze, classify, summarize and find similar cases automatically based on the previous complaints cases, the main three parts of the system: (classify complaints, find similar cases to suggest answers and summarize common complaints to update the FAQ library) .

4.6 Complaints Classification module:

After reviewing various works such as of Hmeidi and Hawashin in [13] which compared KNN and SVM for Arabic text classification and showed that SVM has better accuracy and time, Mesleh [47] applied SVM to classify Arabic articles and showed that the SVM algorithm with the Chi-square method has outperformed Naïve Bayes and the KNN classifiers in term of F-measure. We decide to try the most common methods which are SVM, KNN with Cosine similarity, Naïve Bayes and Decision Tree methods to classify the new complaints (for complaints categorization part), and select the best of them to be in our system as the complaints classifier.

The classifier is built based on the content of the training data set of UNRWA that contained more that 12,000 complaints classified under 14 categories.

For the received complaints documents, text categorization steps are applied as shown in Figure 4.7.

- Apply text pre-processing to make the text documents suitable to train the classifier, it includes ***tokenization*** to convert input text to list of tokens, ***vector space model*** to represent them as a set of vectors, ***stop word removal*** to remove unnecessary words, ***stemming*** to remove suffixes of the resulted fractures and ***dimensionality reduction*** to select the important fields. Details of each step are described in section 4.3, page 30.
- Construct the classifier and tune it by using learning technique against the training data set
- Finally, evaluate the classifier by using some evaluation measures as [error rate, recall, precision and F-Measure].

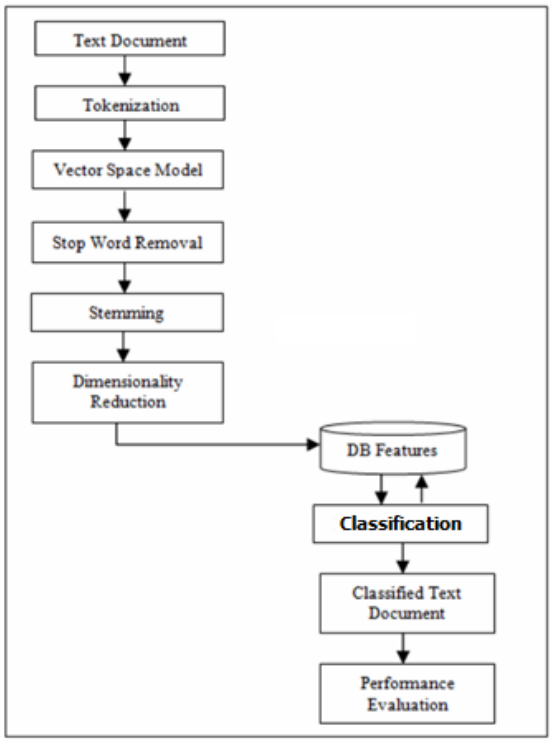


Figure 4.7: the proposed model for Categorization

So we have applied SVM, KNN, Naïve Bayes and Decision Tree methods on our data set (UNRWA data stet) to classify the new complaints , and then selected the best of them to be the ***Complaints Classifier*** in our system.

4.6.1 Support vector machines

A Support Vector Machines (*SVMs*) is a set of related supervised learning methods that analyze data and recognize patterns used for classification and regression. If we have a set of training objects, each one has a predefined category, *SVMs* training algorithm builds a model that predicts whether a new object falls into one category or the other. Intuitively, *SVMs* model is a representation of the objects as points in space, mapped so that the objects of the separate categories are divided by a clear gap that is as wide as possible. New objects are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [44, 48].

A support vector machine constructs a hyperplane or set of hyperplanes in a high dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [3].

The main classification steps for SVM are:

- Use kernel function to Map the data to a predetermined very high-dimensional space.
- Find the hyper plane that maximizes the margin between the two classes.
- If data are not separable, find the hyperplane that maximizes the margin and minimizes the weighted average of the misclassifications.

We used rapid miner to apply SVM algorithm on our dataset as shown in Figure 4.8

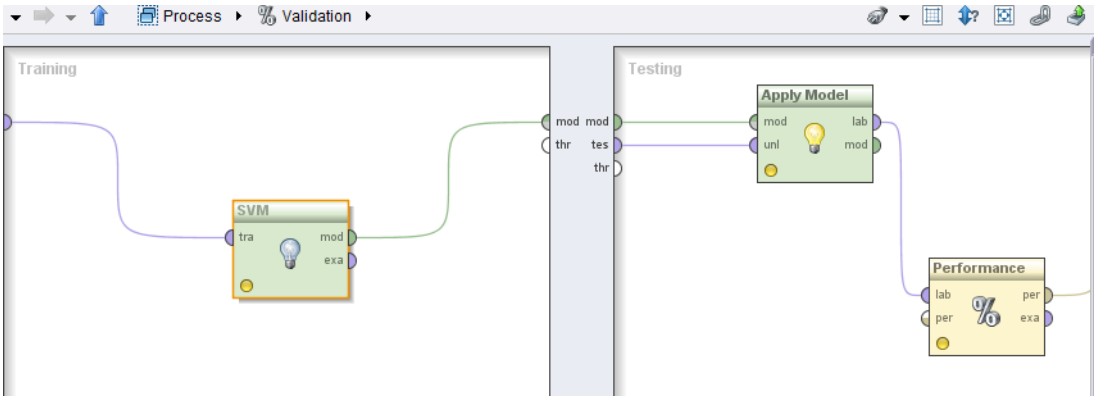


Figure 4.8: applying SVM method

4.6.2 Decision Tree Algorithm

Decision Tree is an algorithm used for classification by generating a tree where each branch of the decision tree represents a possible decision or occurrence. By using a set of training data, it builds the decision tree. At each node of the tree, *it* chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its 50 criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision.

Algorithm for decision tree induction constructs the tree in a top-down recursive divide-and-conquer manner. Below, the summary of the algorithm steps [3, 22]:

- First, all the training samples are at the root
- Samples are partitioned recursively based on selected attributes
- Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain)
- The algorithm stop partitioning in one of the following conditions:
  - All samples for a given node belong to the same class .
  - There are no remaining attributes for further partitioning – majority voting is employed for classifying the leaf.
  - There are no samples left.

We used Rapid miner to apply decision tree algorithm on our dataset as shown in Figure 4.9

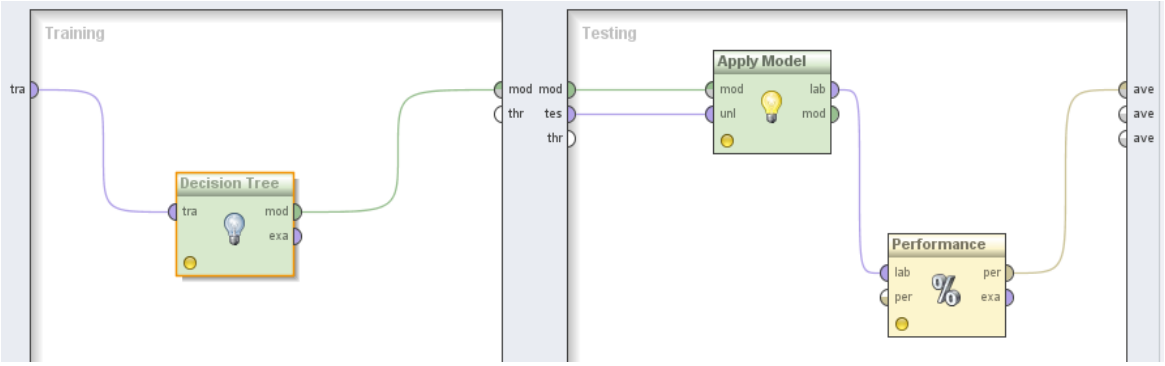


Figure 4.9: applying decision tree method

4.6.3 *K*-Nearest Neighbors (*KNN*)

K Nearest Neighbors algorithm (*KNN*) is a classification method for classifying objects based on nearest training samples in the feature space. *KNN* is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. *KNN* is considered the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of its nearest neighbor [3, 43].

KNN Directly estimates the a posteriori probabilities $P(C|X)$, i.e. bypass probability estimation and go directly to decision functions. *KNN* can center a cell about *x* and let it grows until it captures *kn* samples.

We used rapid miner to apply KNN algorithm on our dataset as seen in figure 4.10

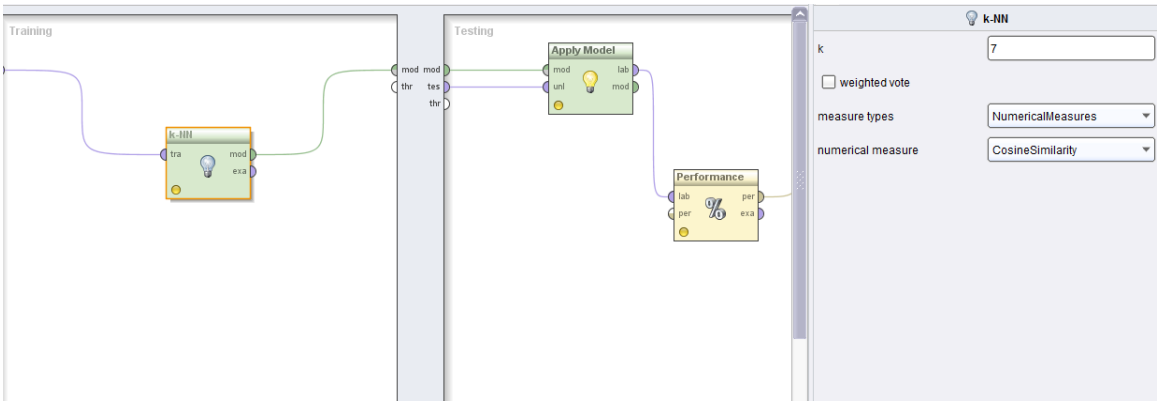


Figure 4.10: applying KNN method

4.6.4 Naïve Bayes

A Naïve Bayes classifier is a statistical classifier based on applying Bayes theorem with strong (naïve) independence assumptions.

Given the class variable, a naive Bayes classifier assumes that the value of a particular feature is unrelated to the presence or absence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and has other features. A naive Bayes classifier uses each of these features to compute the probability that this fruit is an apple, regardless of the presence or absence of the other features [46, 48].

The main feature of naive Bayes classifier is that the required data used to train the classifier is small amount of data, it's used to estimate the parameters means and variances of the variable necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix [37].

We used rapid miner to apply Naïve Bays algorithm on our dataset as seen in figure 4.11

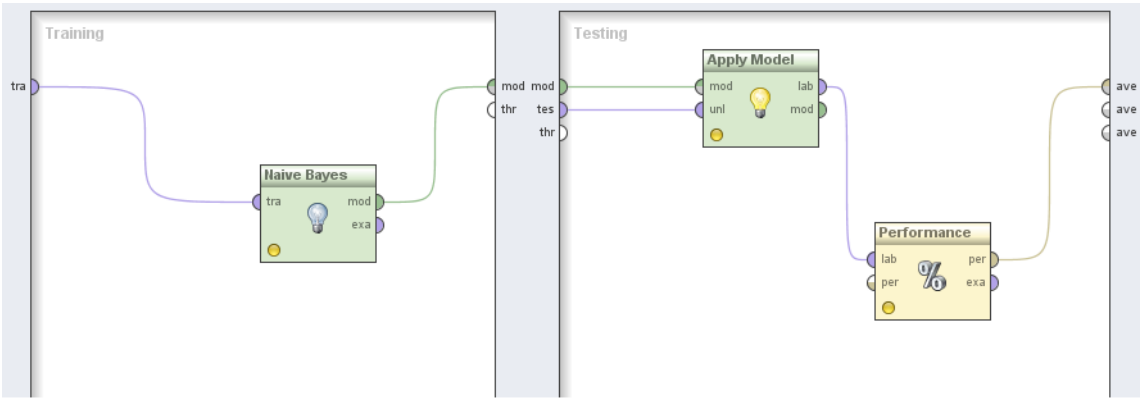


Figure 4.11: applying Naïve Bays method

4.7 Answers Suggestion part:

This part is responsible for suggesting answers based on the previous cases that are similar to the selected case by using text similarity techniques. It uses text similarity techniques to extract similar cases at determined similarity score and display suggested answers automatically as shown in Figure 4.12.

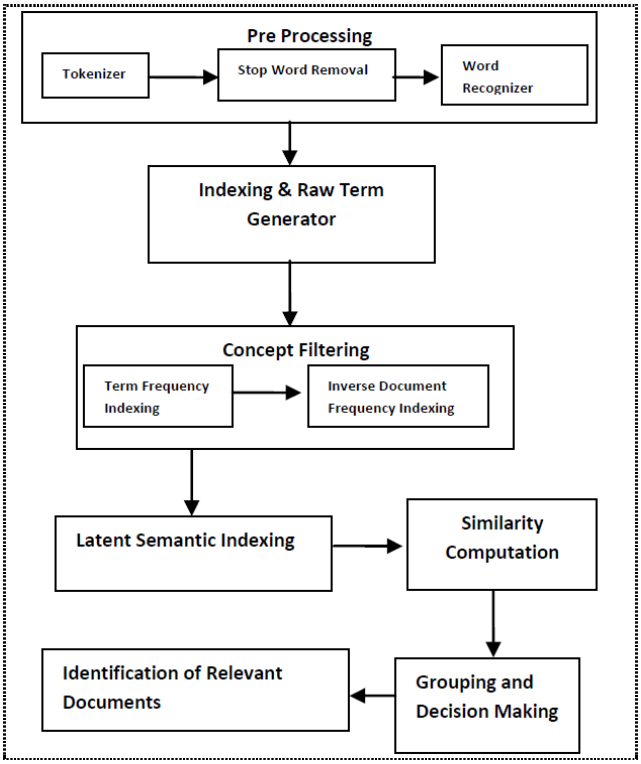


Figure 4.12: Finding similar cases by using text similarity

The main phases illustrated in the Figure 4.12 are:

A. Preprocessing of Learners Knowledge

Pre-processing is the process to prepare data set to be ready for applying the mining methods. The main objective is to optimize the list of terms that identify the collection. The pre-processing module is used to accept input text from the text corpora. The tokenizer is used to convert a text file into a set of tokens. Each of these tokens are passed to the stop word removal system where the stop words such as determiners and prepositions determiners and prepositions are removed from the source documents. Since these words appear in any contexts and they cannot provide useful information to describe a domain concept they can be removed. In our system, we construct a stop word file based on the standard stop word file. And then sort and store the obtained words after stop word removal in another text file.

B. Indexing and Raw Term Generation of Learners Knowledge

It's a process that gets the input from the pre-processing module as tokens. These tokens are collected from number of documents. Each document contains set of distinct words (ie five to ten terms) and then computes the occurrence of each term in every document. And then generates a matrix to show the terms in rows and columns for the document. Finally, arrange the terms in ascending order by using generated indexes.

C. Concept Filtering of Learners Knowledge

Concept filtering uses TF indexing to normalize the raw frequencies across a single document. For example, if a document had two words, one occurring twice and the other occurring thrice, the first word would be normalized to 2/5 (0.4) and the other to 3/5 (0.6).

The resulted term count in the given document is calculated from the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards a document to give a measure of the importance of the term *t* within the particular document *d*. Thus, we have the term frequency TF(*t*,*d*) in the document.

D. Latent Semantic Indexing

It's the process of extracting latent relationships among documents based on word co-occurrence. So if document A contains (w1, w2) and document B contains (w2,w3), we can conclude that there is something common between documents A and B. in this case we can say w2 is common between A and B [49].

E. Similarity Computation

Similarity computation is the process of computing the dependency between two entities based on mutual information. Different methods are used as Jaccard computation to compute the association weights among tokens [37].

We have used ***Levenshtein distance similarity*** algorithm to find similarities and achieved excellent results (F-Measure 72.45%).

4.7.1 Levenshtein distance similarity algorithm:

The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other. It is named after Vladimir Levenshtein, who considered this distance in [50] 1965 .

Levenshtein distance may also be referred to as edit distance, although that may also denote a larger family of distance metrics [41], It is closely related to pairwise string alignments. Mathematically, the Levenshtein distance between two strings a, b is given by equation 4.1

$$\mathrm{lev}_{a,b}(|a|,|b|) \qquad (4.1)$$

where

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise.

Note that the first element in the minimum corresponds to deletion (from a$_a$ to b), the second to insertion and the third to match or mismatch, depending on whether the respective symbols are the same.

So we can say the Levenshtein distance between two strings is the minimal number of insertions, deletions, and substitutions of one character for another that will transform one string into the other. So it's a global alignment of strings S_1 and S_2 is a way of lining up the

two strings (with spaces possibly inserted into one or both strings or at the ends) so that each letter or space in S_1 corresponds to a letter or space in S_2 and vice-versa. Note that a space indicates an insertion or deletion and needs to be distinguished from a blank if "blank" is a member of the alphabet [50].

D(i, j) = edit distance between S_1[1..i] and S_2[1..j]

Recurrence: D(i, 0) = i, D(0, j) = j, and D(i, j) = min[D(i-1, j)+1, D(i, j-1)+1, D(i-1, j-1) + (S_1(i) != S_2(j))], where (a!=b) has the value 1 if the characters a and b don't match and 0 if they match.

4.7.2 Implementation of answer suggestion part

The ***answer suggestion part*** was written in C#. We have implemented **(Levenshtein Edit Distance Algorithm) in the system,** we used two-dimensional arrays to store the distances of prefixes of the words compared, and return the amount of difference between the two strings based on the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character. The program starts by displaying received complaint documents.

<u>The main steps of answers suggestion part :</u>

- First select a complaint.
- And then compare it with the stored complaints in the database and return the similarity score.
- If the similarity score match the determined similarity score e.g. 0.5, add it to similar complaints list to display them in ***similar cases suggestion*** area.

Note: For each complaint document, apply preprocessing steps on it before passing it to similarity method.
See Figure 4.13 illustrates the ***Levenshtein distance similarity*** algorithm.

Algorithm 1 Control sequence comparison

```
1:  procedure LevenshteinDistance(A, B)
2:      define D[n+1][m+1]                          ▷ |A| = n, |B| = m
3:      set D[i][0..m] ← i                          ▷ 0 ≤ i ≤ n
4:      set D[0..n][j] ← j                          ▷ 0 ≤ j ≤ m
5:      for i ← 0, n do
6:          a ← getCharAt(A[i])
7:          for j ← 0, m do
8:              b ← getCharAt(B[j])
9:              if a = b then
10:                 cost ← 0
11:             else
12:                 cost ← 1
13:             end if
14:             D[i][j] = min{D[i-1][j]+1, D[i][j-1]+1,
                          D[i-1][j-1]+cost}
15:             w ← getWeighAt(i)
16:             D[i][j] = w×D[i][j]
17:         end for
18:     end for
19:     d ←D[n+1][m+1]                              ▷ distance
20:     return d                                    ▷ the similarity
21: end procedure
```

Figure 4.13: Levenshtein distance algorithm [50]

4.8 Complaints Summarization part:

Main Objective of this phase is to build a tool in the system that uses summarization techniques to update the FAQ library with the most asked questions. We proposed the model as shown in Figure 4.14.

Figure 4.14: Summarizer module

To summarize questions to update FAQ library, we use the following steps:

a. **Topics Selection :**
The first step in updating the FAQ library is nominating a set of topics based on the number of questions that lay under each topic and pass the selected topics for the second phase as shown in Figure 4.15.

Subject	count
التأمين الصحي	66
اقتراح	59
طلب قرض	54
التامين الصحي	51
شكر وتقدير	31
التعليم الصيفي	29
اضافة مولود	26
الزيادة السنوية	26
شكوي	25
شكر	25
العلاوة السنوية	25
علاوة الابناء	24
رواتب العاب الصيف	24
خصم من الراتب	23
طلب مساعدة	22
التدقيق العائلي	21
طلب نقل	21
اجازة الابوة	20
طلب	20
تقييم المعلمين	19
راتب التعليم الصيفي	18

Figure 4.15: Nominated Topics

b. **Select set of similar questions for selected topic :**
For the questions of each nominated topic, find the similarity of the latest questions with determined accuracy, and produce list of questions and count for each of them.

c. **Sort questions with descending order (highest first) :**
For each selected topic, select the top five questions from the sorted set and use them as input for our summarizer.

d. **Apply summarization and update the FAQ:**
After selecting the questions, supply our summarizer with these questions and update the FAQ library with the resulted summary.

4.8.1 Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. As the problem of information overload has grown, and as the quantity of data has increased, so has interest in automatic summarization. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. An example of the use of summarization technology is search engines such as Google.

4.8.2 A popular summarization methods that deal with Arabic text are:

Centroid-based summarization algorithm of multiple documents, LexRank algorithm And Continuos LexRank algorithm.

4.8.3 The selected algorithm for our summarizer:

After reviewing many papers about summarization techniques, we found the best techniques to be used in our system is ***Centroid-based summarization of multiple documents ,*** and then implemented it in our system .

4.8.4 Centroid-based summarization of multiple documents:

We used a ***multiple-document summarization*** method to summarize the most asked question and update FAQ library with the latest complaints. So our system extract a summary from ***multiple questions*** based on the document cluster centroids. This summarization technique is a cluster- based, extractive summarization method, where passages are first clustered based on similarity, prior to the selection of passages that form the extractive summary of the documents.

The sentences are then issued a timestamp based on the order of their occurrence in the original document, thereby ensuring the chronological order of sentences. Passage clustering forms a main component in this system that aims to extract the most relevant sentences of the documents at the same time keeping the summary non-redundant.

Centroid-based works as follows: First: the sentence scorer gives a value to each sentence based on a linear combination of their features. Sentences are then ordered according to their scores. The sentence re-ranker then adds sentences to the summary beginning with the highest scoring sentence. The re-ranker calculates the similarity of the sentence about to be added with all of the sentences already in the summary. If the similarity is above a given threshold, the sentence is not added to the summary and the re-ranker moves on to the next sentence. Sentences are added to the summary until the amount of sentences in the summary corresponds to the compression rate. **So** the passages are first clustered based on similarity, prior to the selection of passages that form the extractive summary of the documents

4.8.5 Implementation of summarization part

The **complaints summarization *part*** was written in C#. We have implemented **(*Centroid-based summarization of multiple documents* Algorithm)** in the system and integrated it with other parts.

This part works as follow:

- Select specific topic and extract the complaints for selected topic.
- Find the top five similar questions from the list.
- And then send the result to summarization method.
- The summarizer read the received multiple complaints documents and do summarization with selecting the compression rate e.g. 0.3.
- Display the result in the ***summary Area***.

See below in Figure 1.16 how the implemented algorithms do summarization.

Algorithm Steps:

- For all sentences in the cluster
  Begin
  1. Sort the sentences in descending order based on the obtained score values after the reduction of the redundancy penalty.

  End

  Begin
  1. Get the compression rate from the user
  2. Select the required number of sentences based on the compression rate.
  3. Sort the sentences in the ascending order depending on the timestamps
  4. If the Timestamps are the same

     Begin
     - Compare the score values
     - Sentence with the higher score value will appear first

     End

  End

Figure 4.16: Centroid-based summarization algorithm

4.9 Evaluation Methods

We used the following evaluation methods to evaluate the implemented parts:

4.9.1 Evaluating Text Similarity and Classifier modules:

We calculated recall, precision and F-measure to evaluate our modules, and determined what is the best F-Measure based on similarity score.

- **Precision :** is the number of correct results divided by the number of all returned results [equation 4.2]:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (4.2)$$

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P_n [34].

For example: for a text search on a set of documents precision is the number of correct results divided by the number of all returned results.

Precision is also used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.

- **Recall**: is the number of correct results divided by the number of results that should have been returned, Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.
For example: for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned

In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

- **F-measure:** is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score:

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

The traditional F-measure or balanced F-Score (F1 score) is the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

4.9.2 Evaluating Summarization Module

In this work, the measures that used in the evaluation are the precision, recall and F-Measure.

To illustrate how these two measures are used to evaluate text summarization; consider an example document for summarization and let X be the set of sentences in its summary (generated manually by an expert in the field), and Y be the set of sentences that are extracted by the system from the text, and Z be the set of sentences in the intersection of the sets X and Y as illustrated in Figure 4.17.

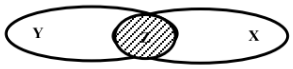


Figure 4.17: Sentences intersection.

The recall and precision can be computed as:

Recall R is the percentage of the target sentences that the system extracted.

$$R = \frac{|Z|}{|X|}$$

Precision P is the percentage of the extracted sentences that the system got right

$$P = \frac{|Z|}{|Y|}$$

F-measure F is to combine precision and recall into a single measure of overall performance.

$$F = \frac{2PR}{P+R}$$

4.10 Summary:

In this chapter we described our application, and presented algorithms for categorizing, answers suggestions and summarization of the complaints using a text mining techniques. We used UNRWA dataset to train the system and build the automated complaint system. The proposed system analyses the text message contents, categorizes it by using text categorization algorithms and tries to decide where to direct the question request automatically to the right person in order to get it answered.
Also it uses text similarity techniques to suggest the answers automatically and system use summarization techniques to update the FAQ library with the most asked questions. The next chapter will be about the results of our experiments.

Chapter 5
Experimental Results and Analysis

This chapter describes the results and the analysis of evaluating the main parts of our system. It also describes the comparisons between used methods according to the results to achieve the best performance. Each method was evaluated using precision, recall, F-measure. Each experiment was performed with the same dataset so that the results could be compared.

We have used C# language to implement the system, and implemented the following parts: Text Processing, Complaints Classifier, Answers Suggestion part and Complaints Summarizer, and evaluated the performance for each part in the system. This chapter contains sections describing the experiments for each system part: The preprocessing part which is common for all other parts, classifying new complaints part, answers suggestion part and summarizing complaints part.

The experimental environment used for all experiments was: CPU / Intel Pentium i5 processor, Memory is 4 GB DDR2 RAM, Windows 7. Also, we used the following software: visual studio 2010, Excel 2007 and Rapid miner.

4.1 Text Preprocessing

The first step is preparing the data to be ready for applying text mining methods, to transform the Arabic text messages to a form that is suitable for used algorithms. In our experiments, we used tokenizer, light stemmer, Stop word removal and vector representation for preparing data as described in chapter 4. The result was converting complaints text messages to Word List that contains the occurrence of each word in the category as shown in Table 5.1.

Word	Total Occur...	Document ...	الطوارئ	الإدارة	المالية	التعليم	الإعاقة	مدير العمليات	تمويل المشاريع ...
درأ	12	12	1	8	0	3	0	0	0
درب	8	5	0	8	0	0	0	0	0
درج	67	37	5	48	0	6	0	5	1
درس	209	110	8	78	6	94	0	11	0
دعا	6	5	0	2	0	0	0	3	1
دعم	2	2	0	1	0	0	0	1	0
دفتر	2	2	0	1	0	1	0	0	0
دفع	21	15	4	10	3	0	0	1	1
دفئ	11	10	0	8	0	0	0	2	0
دكتور	11	5	0	9	0	1	0	0	0
دكتوراة	2	2	0	2	0	0	0	0	0
دلل	10	5	1	0	0	1	0	6	1
دمي	11	11	2	6	0	0	0	3	0
دنا	9	9	3	4	0	1	0	0	0
دوأ	16	9	6	4	0	1	0	0	0
دور	172	91	15	94	4	29	0	16	5

Table 5.1: Word List

The resulted number of classes is 14 classes, Table 5.2 contains the resulted vectors number for each class:

Table 5.2: Resulted support vectors for classes

Class (Eng)	Class (Arb)	Number of support vectors for class
Administration	الإدارة	3750
Emergency Programme	الطوارئ	744
Education	التعليم	1334
Finance	المالية	974
Relief & Social Services	الإغاثة والخدمات الاجتماعية	77
Office of DUO-G	مدير العمليات	506
Microfinance & Microenterprise	تمويل المشاريع الصغيرة	51
Mental Health Programme	الصحة النفسية	21
Staff Response Unit	وحدة الإستجابة للموظفين	109
Environmental Health	صحة البيئة	14
UNRWA Adminsitration (HQ)	رئاسة وكالة الغوث (HQ)	35
Procurement	التوريدات	177
Health	الصحة	28
Engineering	الهندسة	108

5.2 Complaints Classification

We carried four types of classifiers for classifying the new complaints, and compared them to select the best classifier in the system.

We used Rapid miner to apply the classification methods on our dataset. We used 12.699 complaints in these experiments. In order to ensure the reliability of the results, 10-folds cross validation test was followed. The data set is divided into 10 equal subsets. Each of them is used once as testing data where the other 9 subsets are the training data. So we have applied SVM, KNN, Naïve Bays and Decision Tree methods on our data set and compared them to select the method that achieved the ***highest accuracy*** to construct the classifier.

We used Rapid miner to evaluate the selected classification methods to construct our classifier in the system, see Figure 5.1 as the classification process.

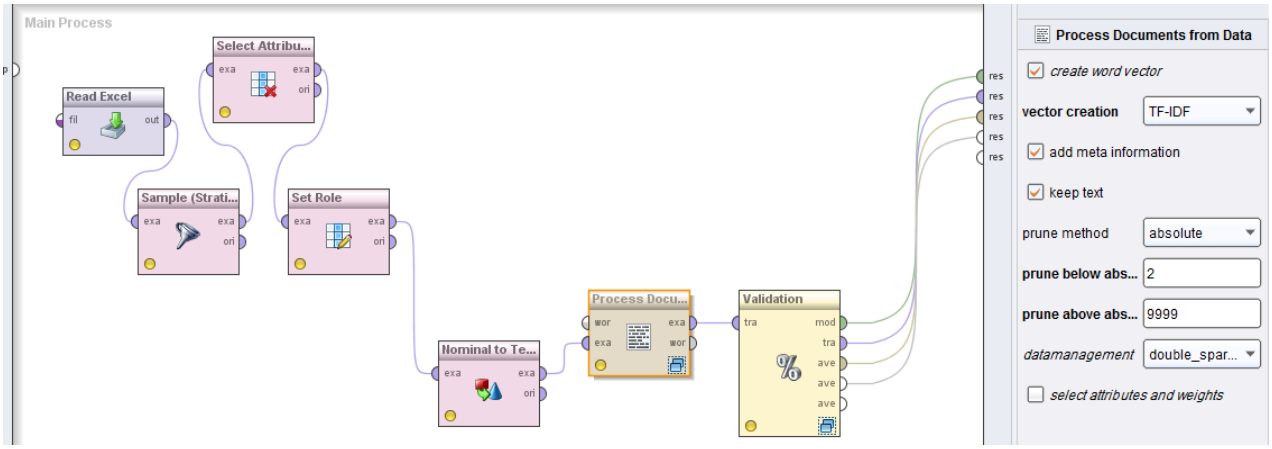


Figure 5.1: Classification Process

5.2.1 SVM Algorithm

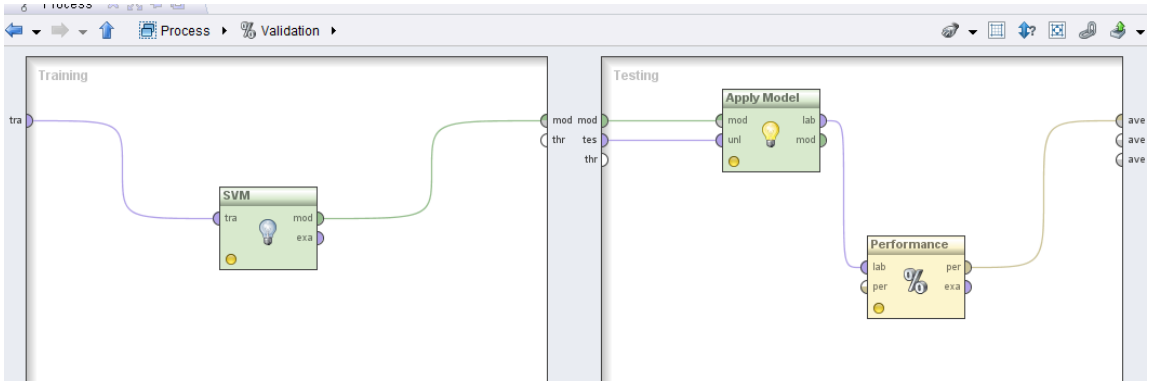


Figure 5.2: Applying SVM method

For SVM, we have tested it with changing the sample size and see the results. We noticed that the accuracy increase with increasing the sample size, below the summary of these experiments, see Figure 5.3.

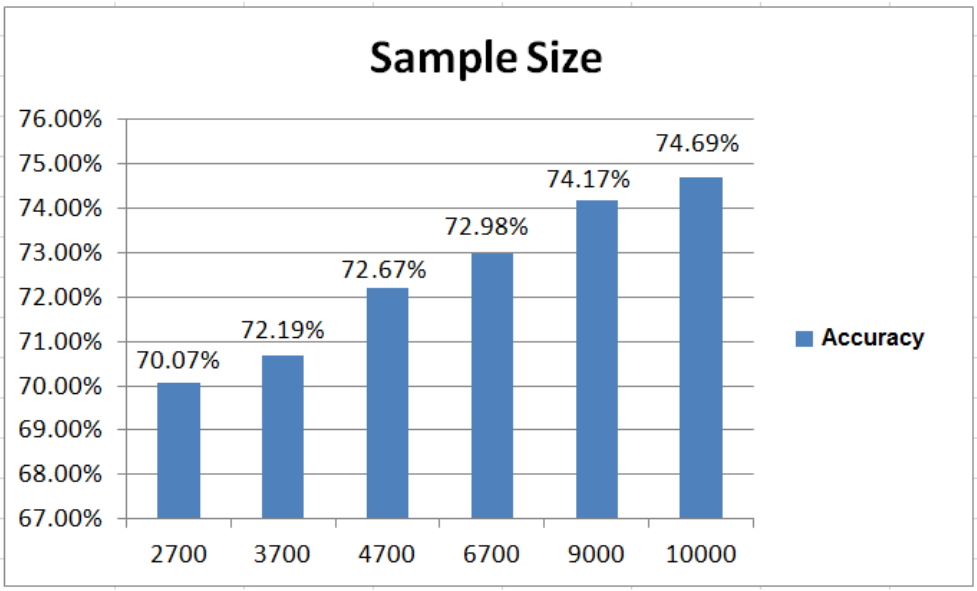


Figure 5.3: SVM Accuracy for different samples

So the best accuracy for SVM was 74.69%, and also we calculated Precision, recall and f-measure to compare it with the other methods, the results as shown in Table 5.3:

Table 5.3 : SVM results

Precision	74.96%
Recall	74.69%
F-Measure	74.82%
Accuracy	74.69%

Table 5.4 shows a confusion matrix of 14 categories and the 10,000 test complaints. This shows which complaint was "confused" with one another and which categories were clearly identified.

Table 5.4: SVM method classification results

	true الطوارئ	true الإدارة	true المالية	true التعليم	true الإغاثة	true مدير العمليات	تمويل المشاريع	الصحة النفسية	true الصحة	true صحة البيئة	الإستجابة للموظفين	رئاسة وكالة	true التوريدات	true الهندسة	Total	class precision
pred. الطوارئ	789	46	10	9	34	25	0	0	3	2	9	7	0	4	938	0.84
pred. الإدارة	108	4883	389	600	26	339	17	19	119	9	87	18	31	17	6662	0.73
pred. المالية	19	122	687	29	7	8	29	0	0	0	2	1	1	0	905	0.76
pred. التعليم	13	204	17	858	1	48	1	1	6	0	1	2	2	2	1156	0.74
pred. الإغاثة	0	1	1	0	7	2	0	0	0	0	1	0	0	0	12	0.58
pred. مدير العمليات	5	13	2	9	1	83	0	0	1	0	3	3	1	0	121	0.69
pred. تمويل المشاريع الصغيرة	0	0	0	0	0	0	4	0	0	0	0	0	0	0	4	1.00
pred. الصحة النفسية	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1.00
pred. الصحة	0	4	0	1	0	0	0	0	49	0	0	0	0	0	54	0.91
pred. صحة البيئة	0	0	0	0	0	0	0	0	0	3	0	0	0	0	3	1.00
pred. وحدة الإستجابة للموظفين	0	0	0	0	1	2	0	0	0	0	5	0	0	0	8	0.63
pred. رئاسة وكالة الغوث (HQ)	0	0	0	0	0	0	0	0	0	0	0	4	0	0	4	1.00
pred. التوريدات	0	22	10	3	0	0	0	0	0	0	1	0	91	0	127	0.72
pred. الهندسة	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	1.00
Total	934	5295	1116	1509	77	507	51	21	178	14	109	35	126	28		
class recall	0.84	0.92	0.62	0.57	0.09	0.16	0.08	0.05	0.28	0.21	0.05	0.11	0.72	0.18		

5.2.2 Decision Tree:

We applied Decision Tree algorithm on our data set and analyzed the results.

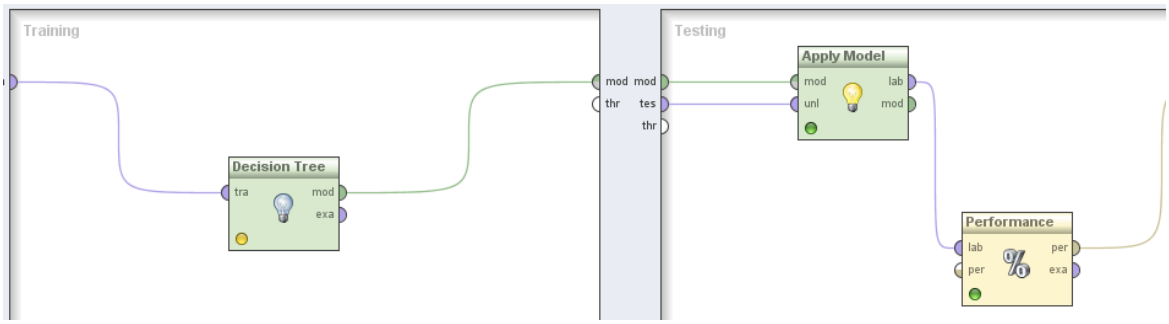


Figure 5.4: Applying Decision Tree method

The accuracy for **Decision Tree** was 52.95%, and also we calculated Precision, recall and f-measure to compare it with the other methods, the results as shown in Table 5.5:

Table 5.5: Decision Tree results

Precision	28.04%
Recall	52.95%
F-Measure	36.66%
Accuracy	52.95%.

Table 5.6 shows a confusion matrix of 14 categories and the 10,000 test complaints. This shows which complaint was "confused" with one another and which categories were clearly identified.

Table 5.6: Decision Tree classification results

	true الطوارئ	true الإدارة	true المالية	true التعليم	true الإغاثة	true مدير العمليات	true تمويل	true الصحة	true الصحة	true صحة البيئة	true وحدة الإستجابة	true رئاسة	true التوريدات	true الهندسة	TotalRetrived	class precision
pred. الطوارئ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. الإدارة	934	5295	1116	1509	77	507	51	21	178	14	109	35	126	28	10000	52.95%
pred. المالية	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. التعليم	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. الإغاثة	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. مدير العمليات	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. تمويل المشاريع الصغيرة	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. الصحة النفسية	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. الصحة	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. صحة البيئة	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. وحدة الإستجابة للموظفين	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. رئاسة وكالة الغوث (HQ)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. التوريدات	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. الهندسة	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
Total	934	5295	1116	1509	77	507	51	21	178	14	109	35	126	28		
class recall	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%		

5.2.3 KNN Algorithm

We used rapid miner to apply K-NN with Cosine similarity on our dataset, with changing the K value to get the best performance.

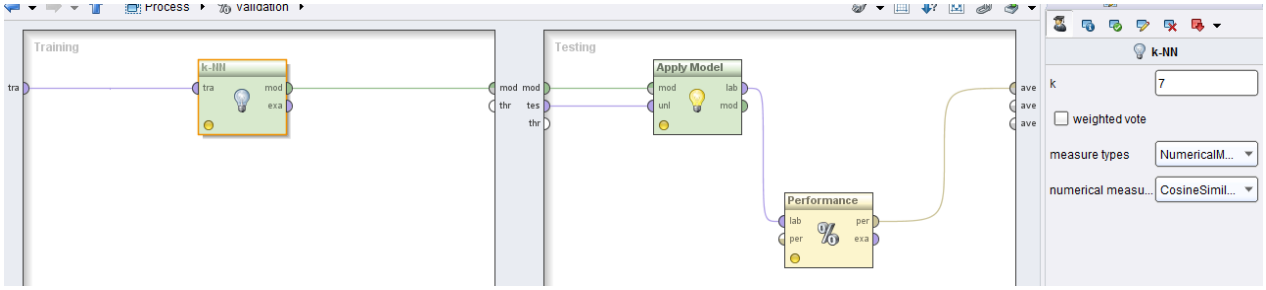


Figure 5.5: Applying KNN method

For K-NN, we have tested it with changing the k value, and we got the best accuracy at k=7 , the accuracy was 68.32%, and also we calculated Precision, recall and f-measure to compare it with the other methods, the results as shown in Table 5.7:

Table 5.7: KNN results

Precision	65.28%
Recall	68.32%
F-Measure	66.76%
Accuracy	68.32%

Table 5.8 shows the resulted confusion matrix for KNN method:

Table 5.8: KNN classification results

	true الطوارئ	true الإدارة	true المالية	true التعليم	true الإغاثة	true مدير العمليات	تمويل المشاريع	الصحة النفسية	true الصحة	true صحة البيئة	الإستجابة للموظفين	رئاسة وكالة	true التوريدات	true الهندسة	TotalRetrived	class precision
pred. الطوارئ	445	10	60	18	1	29	30	4	2	5	4	1	3	3	615	72.36%
pred. الإدارة	25	397	162	74	26	4	16	2	0	3	1	2	1	0	713	55.68%
pred. المالية	93	227	2794	417	6	14	191	79	9	55	15	25	3	10	3938	70.95%
pred. التعليم	7	11	137	385	1	3	32	5	0	2	0	6	0	2	591	65.14%
pred. الإغاثة	0	1	0	0	3	0	0	0	0	0	0	0	0	0	4	75.00%
pred. مدير العمليات	0	1	4	0	0	0	0	1	0	0	0	0	0	0	6	0.00%
pred. تمويل المشاريع الصغيرة	1	2	15	4	0	0	17	3	0	1	2	0	0	0	45	37.78%
pred. الصحة النفسية	1	0	3	1	0	0	0	17	0	0	0	0	0	0	22	77.27%
pred. الصحة	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. صحة البيئة	1	0	2	0	0	0	0	0	0	0	0	0	0	0	3	0.00%
pred. وحدة الإستجابة للموظفين	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
pred. رئاسة وكالة الغوث (HQ)	0	7	9	3	0	1	0	1	0	0	0	38	0	0	59	64.41%
pred. التوريدات	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	100.00%
pred. الهندسة	0	0	1	0	0	0	0	0	0	0	0	0	0	2	3	66.67%
Total	573	656	3187	902	37	51	286	112	11	66	22	72	8	17		
class recall	77.66%	60.52%	87.67%	42.68%	8.11%	0.00%	5.94%	15.18%	0.00%	0.00%	0.00%	52.78%	12.50%	11.76%		

5.2.4 Naïve Bays Algorithm

We applied Naïve Bays algorithm on our data set and noticed the results.

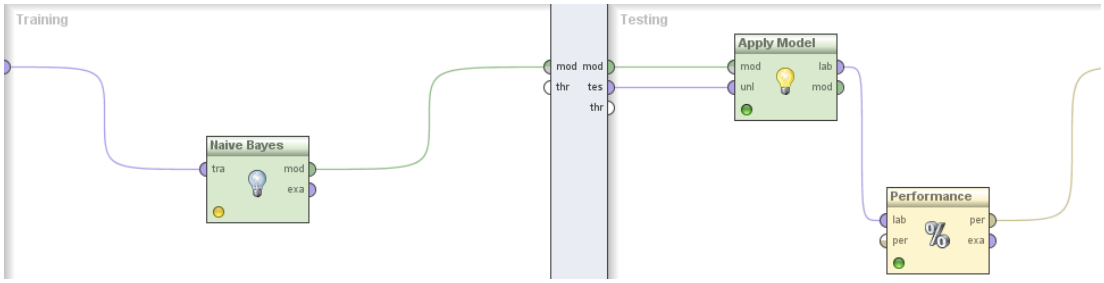


Figure 5.6: Applying Naïve Bays method

After applying the algorithm on 12,000 complaints, the accuracy was 56.42%. And the other results see Table 5.9.

Table 5.9: Naïve Bays results

Precision	59.07%
Recall	56.42%
F-Measure	57.72%
Accuracy	56.42%

Table 5.10 shows the resulted confusion matrix for **Naïve Bays** method:

Table 5.10: Naïve Bays classification results

	true الطوارئ	true الإدارة	true المالية	true التعليم	true الإغتة	true مدير العمليات	true تمويل	true الصحة	true الصحة	true صحة البيئة	true وحدة الإستجابة	true رئاسة	true التوريدات	true الهندسة	TotalRetrived	class precision	precision Summation
pred. الطوارئ	585	223	59	54	14	30	1	0	7	0	7	3	7	4	994	58.85%	549.659
pred. الإدارة	234	3103	233	444	12	185	7	1	65	2	36	2	80	8	4412	70.33%	3723.9735
pred. المالية	15	527	719	63	1	10	3	0	2	2	2	1	9	0	1354	53.10%	592.596
pred. التعليم	47	865	36	791	6	83	6	0	26	0	13	3	12	5	1893	41.79%	630.6111
pred. الإغتة	5	1	8	1	40	3	0	0	0	0	1	0	0	0	59	67.80%	52.206
pred. مدير العمليات	39	344	16	99	4	185	1	0	3	0	10	0	3	1	705	26.24%	133.0368
pred. تمويل المشاريع الصغيرة	0	3	7	0	0	0	32	0	0	0	0	0	0	0	42	76.19%	38.8569
pred. الصحة النفسية	0	1	0	0	0	0	0	20	0	0	0	0	0	0	21	95.24%	20.0004
pred. الصحة	1	107	4	18	0	3	1	0	70	0	2	0	0	0	206	33.98%	60.4844
pred. صحة البيئة	0	0	0	0	0	0	0	0	0	10	0	0	0	0	10	100.00%	14
pred. وحدة الإستجابة للموظفين	5	72	28	19	0	8	0	0	3	0	36	0	0	0	171	21.05%	22.9445
pred. رئاسة وكالة الغوث (HQ)	1	2	0	0	0	0	0	0	0	0	0	26	0	0	29	89.66%	31.381
pred. التوريدات	1	42	6	20	0	0	0	0	0	0	2	0	15	0	86	17.44%	21.9744
pred. الهندسة	1	5	0	0	0	0	0	0	2	0	0	0	0	10	18	55.56%	15.5568
	934	5295	1116	1509	77	507	51	21	178	14	109	35	126	28			
class recall	62.63%	58.60%	64.43%	52.42%	51.95%	36.49%	62.75%	95.24%	39.33%	71.43%	33.03%	74.29%	11.90%	35.71%			
Recall Summation	584.9642	3102.87	719.0388	791.0178	40.0015	185.0043	32.0025	20.0004	70.0074	10.0002	36.0027	26.0015	14.994	9.9988			

5.2.5 Results Analysis for our classifier

Among four classifiers applied on the dataset, *SVMs* achieved the highest average accuracy (74.69%), then *KNN* with average accuracy of 68.32. Decision Tree was the worst with average accuracy of 52.95%. So we selected SVM method to be our classifier in the system.

Generally, *SVMs* and *KNN* achieved the best average classification accuracy. *SVMs* achieved the best accuracy because it is a robust classifier, it maps data points into new dimension space, this makes different term weighting schemes have no impact on *SVMs* performance. In addition, *SVMs* is effective on high dimensional data because the complexity of trained classifier is characterized by the number of support vectors rather than the dimensionality of the data, see Table 5.11.

Table 5.11: Classification methods performance

Classification Method	Precision	Recall	F-Measure
SVM	74.96%	74.69%	74.82%
Decision Tree	28.04%	52.95%	36.66%
KNN	65.28%	68.32%	66.76%
Naïve Bays	59.07%	56.42%	57.72%

5.3 Answers Suggestion Part

We have evaluated **answers suggestion** part that implemented by using ***Levenshtein distance similarity*** algorithm in our system with changing similarity score, and after several experiments we got the best F-Measure at similarity score 0.50 , the number of experiments is 60 .

The system has access to the dataset, when a new complaint / question is received, the system compares it with all stored complaints, and return the similar cases including the similarity score for each of them . And the results sorted by ***similarity score***, the highest first. See Figure 5.7 as example of new complaints about getting compensation of strike days.

Figure 5.7: Answers Suggestion in our system.

We have tested several **cases**, for most of them the algorithm return [3-6] similar cases from the stored complaints (12,000 cases).

Figure 5.8 and Figure 5.9 show samples of the tested cases and the results. Each sample represents received complaint, and the system display similar cases for selected complaint in descending order according to similarity score for each suggested complain.

Case 1	هل سيتم تعويض المعلمين الخصم في الراتب بسبب الإضرابات
Similar Cases	هل سيتم تعويض راتب أيام الإضراب في راتب هذا الشهر؟؟ Score = 0.6086956 متى سيتم تعويض المعلمين عن خصم ايام الإضراب Score = 0.6004348 السلام عليكم هل سيتم استرداد الخصم على المعلمين في راتب يناير Score = 0.5882353 بعد التحية بدي اعرف هل ستعوض ايام الاضراب للمعلمين في راتب شهر 1 Score = 0.5576923 هل يمكن تعويض ايام الاضراب بالخصم من الاجازات الطارئة Score = 0.5434783 ارجو افادتي باسباب الخصم من الراتب لشهرين متتاليين Score = 0.5217391 هل سيتم اعادة الخصومات في راتب يناير للمدرسين Score = 0.5217391 الأخوة الكرام هل سيتم تعويض الخصومات بالكامل هذا الشهر ديسمبر Score = 0.5192308

Figure 5.8: Similar cases samples

Case 2	احاول ان اتقدم بطلب اجازة ابويه ولكن لم اجد الا مسميات الاجازات المعهودة . كيف استطيع التقدم باجازة ابويه من خلال موقعكم وشكرا
Similar Cases	أريد ان أتقدم لأجازة أبوة" الاسبوع الثاني "خلال شهر 8 تبدأ من صباح 20/8 وتنتهي مساء 25/8 وذلك من خلال بوابة الموظفين ولكن لا أستطيع .كيف اقدم هذه الاجازة Score = 0.504 يرجى من سيادتكم افادتي متى يستطيع الموظف الحصول على اجازة الابوة وما هي الاجراءات المتبعة ما هي الاوارق المطلوبة لاعتماد الاجازة Score = 0.5085185 انا مدرس هل استطيع اخذ اجازة الابوة في فترة الاختبارات الموحدة سواء كانت في نهاية الفصل الاول او الثاني ؟ ولكم الشكر Score = 0.50

Figure 5.9: Similar cases samples

We have tested our text similarity part (answers suggestion), by using three ***similarity score*** value [0.45, 0.50 and 0.55] and see the results. also we tested it for less than 0.5 and more than 0.6, the result was:

- When the similarity score less than 0.5, the result includes many irrelevant complaints.
- When the similarity score more than 0.5 (0.6 and more), the result includes little complaints and similar complaints didn't appear in the result.
- So, we have tested it by using the best similarity scores [0.45, 0.50 and 0.55] and compared the results.

First: similarity score =0.55

We have used group of complaints for testing, and apply our answer suggestion part on these samples , for each tested case: we recorded the ***true suggested answers*** , ***total suggested results*** and all similar cases in the dataset, after that we calculated the recall , precision and F-measure . Figure 5.10 contains the details of the experiment results at similarity score =0.55

	Compalint Case	Similarity Score	Total Similar Result	Total Result	Total Similar Dataset	Precision	Recall
2	هل يحق للموظف أن يعمل كرت مواصلات باتجاه واحد فقط ؟ ذهاباً فقط ؟ والرجاء تزويدي برقم هاتف للاستفسار عن ذلك أكثر ؟ شكرا لكم	0.55	1	1	3	1	0.33333333
3	بأعطاء علاوه بدل مخاطره لجميع موظفين دائرة الصحه، حيث تم استثناء كتبة العياداء ظلما وجورا مما أدى ذلك الى الإستياء الشديد من قبل كتبة العيادات	0.55	2	2	3	1	0.66666667
4	الاستفسار عن كيفية التسجيل في الملعب الصيفي لذا كمعلمين ومتى يتم التسجيل	0.55	2	2	4	1	0.5
5	أريد أن أتقدم بطلب قرض للبناء فكم المبلغ المعطى علما بانني اعمل في المهنة ست سنوات ارجو الافادة	0.55	1	2	4	0.5	0.25
6	كيف يمكن التسجيل بوظيفة شاغرة من خلال البوابة	0.55	3	3	4	1	0.75
7	متى سيعقد امتحان لوظيفة نائب مدير المدرسة التي تقدمنا لها	0.55	4	11	5	0.363636364	0.8
8	هل يحق لي الحصول على علاوة لابني الجامعي مع انني لم اتقاضى عليه علاة ابناء	0.55	3	3	6	1	0.5
9	ارد الاستفسار هل يمكنني تقديم طلب علاوة زوج وابناء حيث ان زوجي يعمل في السلطة ولايتقاضى اي علاوة لي أو لولدنا وماهي الخطوات	0.55	3	3	6	1	0.5
10	زوجتي انهت دراستها الجامعية هل يمكن لها أن تعمل على برنامج النطالة في حال تقدمت بطلب لها	0.55	2	2	7	1	0.28571429
11	يوجد أدني تأمين صحي وغير مضاف في قسيمة الراتب (لا يظهر قيمة التأمين في قسيمة الراتب) حيث يتم الخصم منه كل شهر ونود إلغاء التأمين ما هي الاجراءات المتبعة	0.55	3	4	8	0.75	0.375
12	فقدت كلمة المرور	0.55	4	4	8	1	0.5
13	كيفية طباعة قسيمة الراتب	0.55	1	2	8	0.5	0.125
14	متى سيتم صرف بدل الايجار للغير موظفين من أصحاب البيوت المهدمة	0.55	2	2	9	1	0.22222222
15	اريد الاستفسار عن تقييم الأداء	0.55	4	4	10	1	0.4
16	أرجو منكم التكرم بإفادتي عن موعد بداية العمل بالمشروع السعودي ب و الجدول الزمني المتوقع لانجاز المشروع و موعد التسليم	0.55	4	5	10	0.8	0.4
17	صفحة التدقيق الحالتي على موقع بوابة الموظفين لا تظهر في صفحتي الشخصية	0.55	4	4	10	1	0.4
18	متى موعد صرف نقود التصحيح لأوراق الامتحانات الفصل الدراسي الاول	0.55	7	8	14	0.875	0.5
19	أرسلت لسيادكم رسائل سابقة بخصوص إستعادة الكابونة وتم زيارتي من قبل الباحثين مرتين بعد قطعها وإنني توقعت سأستلم ضمن هذة الدورة ولكن لم يتم تنزيل إسمي أرجوا من سيادتكم مساعدة	0.55	3	3	16	1	0.1875
20	بدي اسال عندي في راتب هذا الشهر ماريس اخر خصم من الراتب بخصوص ثلاثة اضعاف الراتب السؤال هل يجوز لي التقدم لطلب اخر (3 اضعاف الراتب) هذا الشهر ام انتظر حتى ا	0.55	2	2	23	1	0.08695652
21	افيد سيادتكم انني قدمت بطلب بخصوص افقر الفقراء ثلاث مرات على مدار 3 شهور وفي كل مرة يتم وعدي باخراج كابونة علما بان انت الى منزلي باحتة ورفعت الي الادارة الى الادارة	0.55	5	5	34	1	0.14705882
22	متى ستصرف رواتب التعليم الصيفي	0.55	43	44	58	0.977272727	0.74137931
23	هل سوتم تعويض المعلمين الخصم في الراتب بسبب الإضرابات	0.55	4	4	14	1	0.28571429

Figure 5.10: Experiment results at similarity score (0.55)

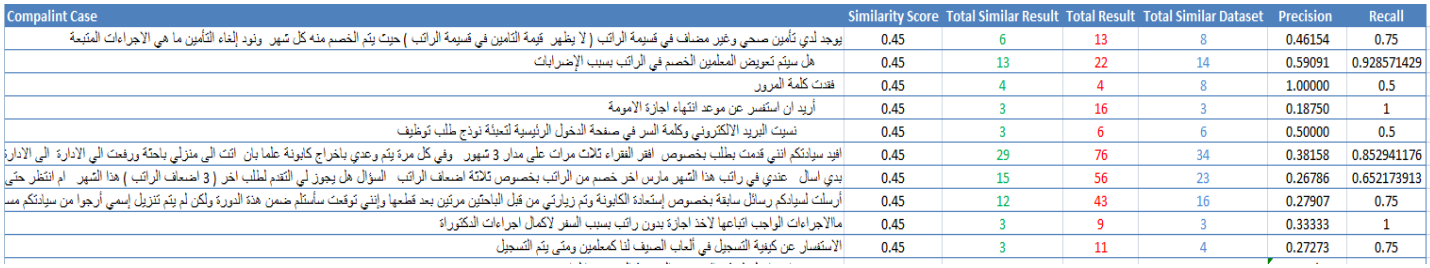
See Table 5.12 shows the total experiment results at similarity score (0.55):

Table 5.12: Total Results at score 0.55

Precision	87.55%
Recall	42.63%
F-Measure	57.34%

Second : similarity score =0.45

Figure 5.11 contains the results of the experiment for text similarity part at similarity score = 0.45.



Compalint Case	Similarity Score	Total Similar Result	Total Result	Total Similar Dataset	Precision	Recall
يوجد لدي تأمين صحي وغير مضاف في قسيمة الراتب (لا يظهر قيمة التأمين في قسيمة الراتب) حيث يتم الخصم منه كل شهر ونود إلغاء التأمين ما هي الاجراءات المتبعة	0.45	6	13	8	0.46154	0.75
هل سيتم تعويض المعلمين الخصم في الراتب بسبب الإضرابات	0.45	13	22	14	0.59091	0.928571429
فقدت كلمة المرور	0.45	4	4	8	1.00000	0.5
أريد ان استفسر عن موعد انتهاء اجازة الامومة	0.45	3	16	3	0.18750	1
نسيت البريد الالكتروني وكلمة السر في صفحة الدخول الرئيسية لتعبئة نوذج طلب توظيف	0.45	3	6	6	0.50000	0.5
افيد سيادتكم انني قدمت بطلب بخصوص افقر الفقراء ثلاث مرات على مدار 3 شهور وفي كل مرة يتم وعدي باخراج كابونة علما بان انت الى منزلي باحثة ورفعت الي الادارة الى الادارة	0.45	29	76	34	0.38158	0.852941176
بدي اسأل عندي في راتب هذا الشهر مارس اخر خصم من الراتب بخصوص ثلاثة اضعاف الراتب السؤال هل يجوز لي التقدم لطلب اخر (3 اضعاف الراتب) هذا الشهر ام انتظر حتى	0.45	15	56	23	0.26786	0.652173913
أرسلت لسيادكم رسائل سابقة بخصوص إستعادة الكابونة وتم زيارتي من قبل الباحثين مرتين بعد قطعها وإنني توقعت سأستلم ضمن هذة الدورة ولكن لم يتم تنزيل إسمي أرجوا من سيادتكم مس	0.45	12	43	16	0.27907	0.75
ماالاجراءات الواجب اتباعها لاخذ اجازة بدون راتب بسبب السفر لاكمال اجراءات الدكتوراة	0.45	3	9	3	0.33333	1
الاستفسار عن كيفية التسجيل في ألعاب الصيف لنا كمعلمين ومتى يتم التسجيل	0.45	3	11	4	0.27273	0.75

Figure 5.11: Experiment results at similarity score (0.45)

Table 5.13 shows the total experiment results at similarity score (0.45).

Table 5.13: Total Results at score 0.45

Precision	42.64%
Recall	78.95%
F-Measure	55.37%

Third: similarity score =0.50

Figure 5.12 contains the results of the experiment for text similarity part at similarity score =0.50

Compalint Case	Similarity Score	Total Similar Result	Total Result	Total Similar Dataset	Precision	Recall
يوجد لدي تأمين صحي وغير مضاف في قسيمة الراتب (لا يظهر قيمة التأمين في قسيمة الراتب) حيث يتم الخصم منه كل شهر ونود إلغاء التأمين ما هي الاجراءات المتبعة	0.5	2	3	8	0.66667	0.25
هل سيتم تعويض المعلمين الخصم في الراتب بسبب الإضرابات	0.5	11	13	14	0.84615	0.785714286
فقدت كلمة المرور	0.5	4	4	8	1.00000	0.5
أريد ان استفسر عن موعد انتهاء اجازة الامومة	0.5	1	2	3	0.50000	0.333333333
نسيت البريد الالكتروني وكلمة السر في صفحة الدخول الرئيسية لتعبئة نوذج طلب توظيف	0.5	2	2	6	1.00000	0.333333333
افيد سيادتكم انني قدمت بطلب بخصوص افقر الفقراء ثلاث مرات على مدار 3 شهور وفي كل مرة يتم وعدي باخراج كابونة علما بان انت الى منزلي باحثة ورفعت الي الادارة الى الادارة	0.5	22	30	34	0.73333	0.647058824
بدي اسأل عندي في راتب هذا الشهر مارس اخر خصم من الراتب بخصوص ثلاثة اضعاف الراتب السؤال هل يجوز لي التقدم لطلب اخر (3 اضعاف الراتب) هذا الشهر ام انتظر حتى	0.5	9	21	23	0.42857	0.391304348
أرسلت لسيادكم رسائل سابقة بخصوص إستعادة الكابونة وتم زيارتي من قبل الباحثين مرتين بعد قطعها وإنني توقعت سأستلم ضمن هذة الدورة ولكن لم يتم تنزيل إسمي أرجوا من سيادتكم مس	0.5	13	18	16	0.72222	0.8125
ماالاجراءات الواجب اتباعها لاخذ اجازة بدون راتب بسبب السفر لاكمال اجراءات الدكتوراة	0.5	2	2	3	1.00000	0.666666667
الاستفسار عن كيفية التسجيل في ألعاب الصيف لنا كمعلمين ومتى يتم التسجيل	0.5	3	7	4	0.42857	0.75
متى سيعقد امتحان لوظيفة نائب مدير المدرسة التي تقدمنا لها	0.5	5	15	5	0.33333	1
متى موعد صرف نقود التصحيح لاوراق الامتحانات الفصل الدراسي الاول	0.5	13	15	14	0.86667	0.928571429
هل يحق لي الحصول على علاوة لابني الجامعي مع انني لم اتقاضى عليه علاة ابناء	0.5	4	6	6	0.66667	0.666666667
هل يحق للموظف أن يعمل كرت مواصلات بإتجاه واحد فقط ؟ ذهاباً فقط ؟ والرجاء تزويدي برقم هاتف للاستفسار عن ذلك أكثر ؟ شكرا لكم	0.5	2	2	3	1.00000	0.666666667
زوجتي انهيت دراستها الجامعية هل يمكن لها أن تعمل على برنامج البطالة في حال تقدمت بطلب لها	0.5	5	6	7	0.83333	0.714285714
كيفية طباعة قسيمة الراتب	0.5	3	4	8	0.75000	0.375
اريد الاستفسار عن تقييم الأداء	0.5	7	8	10	0.87500	0.7
أعطاء علاوه بدل مخاطره لجميع موظفين دائرة الصحه، حيث تم استثناء كتبة العياداة ظلما وجورا مما أدى ذلك الى الإستياء الشديد من قبل كتبة العيادات	0.5	3	3	3	1.00000	1

Figure 5.12: Experiment results at similarity score (0.50)

Table 5.14 shows the total experiment results at similarity score (0.50).

Table 5.14: Total Results at score 0.50

Precision	73.59%
Recall	71.33%
F-Measure	72.45%

Results Analysis :

According to our experiments results, we noticed when the similarity score was 0.55, the precision increased and recall decreased, but when similarity score was 0.50 or 0.45, the precision decreased and recall increased.

So we got best F-Measure (72.45%) at similarity score (0.50) due to expressing the complaints messages in ***indirect way*** , so you find many statements in the message , but small part of the message describe the complaint clearly and others just additional .

5.4 Complaints Summarization

This part was implemented by using ***Centroid-based summarization of multiple documents algorithm*** as described in summarizer module in our system and then tested by using set of real cases from UNRWA data set. Figure 5.13 describes a page that contains group of complaints for selected topic and the resulted summary.

Figure 5.13: Complaints summarizer in our system.

For Our Experiments: We used samples of our data set (UNRWA data) and apply our summarizer on these samples, and calculate recall, precision and F-measure to evaluate our summarizer. We used different compression rates 0.2, 0.3 and 0.4.

The following tables show the details of our experiments and contain samples of the used complaints titles to evaluate our summarizer module, and the table included x-expert column: is the set of sentences in its summary (generated manually by an expert in the field), and y-system column: is the set of sentences that are extracted by the system from the text, and Z column be the set of sentences in the intersection of the sets X and Y.

<u>Compression Rate : 0.2</u>

Figure 5.14 shows the details of our experiment at compression Rate 0.2 and the results.

Compalint Title	Comparation Rate	X-Expert	Y-System	Z	Precision	Recall
كيف استطيع التقدم بإجازة ابويه من خلال الموقع	0.2	1	1	0	0.00000	0
الاستفسار عن كيفية التسجيل في ألعاب الصيف لنا كمعلمين ومتى يتم التسجيل	0.2	5	2	2	1.00000	0.4
الإجراءات المتبعة لإلغاء التأمين الصحي	0.2	1	1	1	1.00000	1
الية الحصول على كابونة لعائلة مصنفة افقر الفقراء ثلاث مرات على مدار 3 شهور	0.2	2	3	1	0.33333	0.5
بخصوص تعويض راتب ايام الاضراب	0.2	1	1	1	1.00000	1
بخصوص مدة اجازة الأمومة المسموحة	0.2	1	1	1	1.00000	1
ماهي آلية استحقاق الدرجة 11 ؟	0.2	3	1	1	1.00000	0.333333333
متى يحق للموظف سحب مبلغ من الادخار وما المبلغ المسموح به ؟	0.2	4	2	1	0.50000	0.25
نسيت البريد الالكتروني وكلمة السر في صفحة الدخول الرئيسية لتعبئة نوذج طلب توظيف	0.2	2	1	1	1.00000	0.5
هل يحق للموظف أن يعمل كرت مواصلات باتجاه واحد فقط ؟	0.2	2	2	1	0.50000	0.5

Figure 5.14: Summarization experiment Samples

After applying our summarizer on the set of complaints and calculated the Precision and recall, we got the results at compression rate **0.2** as in the Table 5.15.

Table 5.15: Total Results at compression rate 0.2

Precision	73.33%
Recall	54.83%
F-Measure	62.75%

Compression Rate : 0.3

Figure 5.15 shows the details of our experiment at compression Rate 0.3 and the results.

Compalint Title	Comparation Rate	X-Expert	Y-System	Z	Precision	Recall
كيف استطيع التقدم بإجازة ابويه من خلال الموقع	0.3	1	2	1	0.50000	1
الاستفسار عن كيفية التسجيل في ألعاب الصيف لنا كمعلمين ومتى يتم التسجيل	0.3	5	7	4	0.57143	0.8
الإجراءات المتبعة لإلغاء التأمين الصحي	0.3	1	1	1	1.00000	1
الية الحصول على كابونة لعائلة مصنفة افقر الفقراء ثلاث مرات على مدار 3 شهور	0.3	2	4	2	0.50000	1
بخصوص تعويض راتب ايام الاضراب	0.3	2	1	1	1.00000	0.5
بخصوص مدة اجازة الأمومة المسموحة	0.3	1	1	1	1.00000	1
ماهي آلية استحقاق الدرجة 11 ؟	0.3	3	3	2	0.66667	0.666666667
متى يحق للموظف سحب مبلغ من الادخار وما المبلغ المسموح به ؟	0.3	4	5	3	0.60000	0.75
نسيت البريد الالكتروني وكلمة السر في صفحة الدخول الرئيسية لتعبئة نوذج طلب توظيف	0.3	2	2	1	0.50000	0.5
هل يحق للموظف أن يعمل كرت مواصلات باتجاه واحد فقط ؟	0.3	2	3	1	0.33333	0.5

Figure 5.15: Summarization experiment Samples

After applying our summarizer on the set of complaints and calculated the Precision and recall, we got good results at compression rate **0.3** as in the Table 5.16.

Table 5.16: Total Results at compression rate 0.3

Precision	66.71%
Recall	77.17%
F-Measure	71.56%

Compression Rate : 0.4

Figure 5.16 shows the details of our experiment at compression Rate 0.4 and the results.

Compalint Title	Comparation Rate	X-Expert	Y-System	Z	Precision	Recall
كيف استطيع التقدم بإجازة ابويه من خلال الموقع	0.4	1	2	1	0.50000	1
الاستفسار عن كيفية التسجيل في ألعاب الصيف لنا كمعلمين ومتى يتم التسجيل	0.4	5	12	4	0.33333	0.8
الإجراءات المتبعة لإلغاء التأمين الصحي	0.4	1	2	1	0.50000	1
الية الحصول على كابونة لعائلة مصنفة افقر الفقراء ثلاث مرات على مدار 3 شهور	0.4	2	7	2	0.28571	1
بخصوص تعويض راتب ايام الاضراب	0.4	2	2	1	0.50000	0.5
بخصوص مدة اجازة الأمومة المسموحة	0.4	1	1	1	1.00000	1
ماهي آلية استحقاق الدرجة 11 ؟	0.4	3	4	2	0.50000	0.666666667
متى يحق للموظف سحب مبلغ من الادخار وما المبلغ المسموح به ؟	0.4	4	6	3	0.50000	0.75
نسيت البريد الالكتروني وكلمة السر في صفحة الدخول الرئيسية لتعبئة نوذج طلب توظيف	0.4	2	1	1	1.00000	0.5
هل يحق للموظف أن يعمل كرت مواصلات باتجاه واحد فقط ؟	0.4	2	6	2	0.33333	1

Figure 5.16: Summarization experiment Samples

After applying our summarizer on the set of complaints and calculated the Precision and recall, we got the results at compression rate **0.4** as in the Table 5.17.

Table 5.17: Total Results at compression rate 0.4

Precision	54.52%
Recall	82.17%
F-Measure	65.55%

Results Analysis :

After doing many experiments by changing the compression rates, we noticed the following**:**

- We got the best results at **compression rate** =0.3 , the best **F-Measure** was **71.56%**
- When applying compression rate **less than** 0.3, the resulted summary didn't contain many words of the expert summary.
- When applying compression rate **more than** 0.3, the resulted summary contained a lot of unnecessary words.
- Also we noticed when ***decreasing the compression rate*** >> the **recall** decreased and **Precision** increased
- And we noticed when ***increasing the compression rate*** >> the **recall** increased and **Precision** decreased.

5.5 Summary

This chapter describes experiments results and analysis of the main parts of our system (Complaints classifier, Answers Suggestion part and Complaints summarizer), and also describes the comparisons between used methods according to the results to achieve the best performance.

According to results analysis for the classifiers, we can say among four classifiers applied on the dataset, *SVMs* achieved the highest average accuracy (74.69%). Also according to results analysis for the answers suggestion part, we got best F-Measure (72.45%) at similarity score (0.50). For Summarization part, we performed many experiments by changing the compression rates, we noticed the best results at compression rate =0.3, the best F-Measure was 71.56%

Chapter 6

Conclusion and Future works

6.1 Conclusion:

In this thesis, we designed and implemented an automated complaints system that integrates some text mining techniques. UNRWA dataset were used in this work. All of them came from the previous complaints submitted in the period from 2011 to 2013. The data set included 12 thousands complaint that belongs to14 categories used for learning.

This thesis examined automatic text categorization of complaints documents by using set of complaints methods (SVM, KNN, Naïve bays and decision tree) and according to the results we noticed that SVMs achieved the best average classification accuracy and then KNN. Final recall and precision results were 74.69% and 74.96% respectively.

Also we conducted several experiments to test answers suggestion part by changing similarity score. According to our experiments results, we noticed when the similarity score was 0.55, the precision increased and recall decreased, but when similarity score was 0.50 or 0.45, the precision decreased and recall increased.

Thus, we conducted several experiments to test summarization module by changing compression rate, we noticed that when decreasing the compression rate, the recall decreased and Precision increased. And also we noticed when increasing the compression rate, the recall increased and Precision decreased.

In addition, experimental results showed that Light stemming greatly reduced features to average of 30% and 50% of the original feature space. Also we conclude that light stemming and term pruning is the best feature reduction technique because light stemming is more proper than stemming from linguistics and semantic view point, and it has the least preprocessing time, it also has superior average classification accuracy.

6.2 Future work:

The work presented here can be developed further to improve quality of answers by using data mining tools to discover new knowledge from the existing data that can help us to know the factors that affect the quality and also discover new rules that help in prediction for users needs and requests. For example, try to know the effects of delay of reply on the feedback. Also know the limitation of the current feedback mechanism by discovering the errors and inconsistent data by using some data mining methods as outlier analysis.

Also try to develop new service called ***automatic answering*** to answer received complaints directly based on existing of similar cases. The work can be developed further to handle English content.

References

[1] D. Jatin Das, S. Arun Kumar, B. Ramakantha Reddy, S. Shiva Prakash ,"Web Data Refining Using Feedback Mechanism and k-mean Clustering" , 2011.

[2] Burges, C., "A Tutorial on Support Vector Machines for Pattern Recognition, data mining and Pattern Recognition" , 2006.

[3] Cherkassky, V., Ma, Y., "SVM-based Learning for Multiple Model Estimation", 2006.

[4] Mesleh A.,"Chi Square Feature Extraction Based SVM Arabic Language Text Categorization System". Journal of Computer Science 3(6): 430-435 , 2007.

[5] Min Song, Yi-Fang Brook, Handbook of research on text and web mining technologies, information science reference, IGI global, 2009.

[6] Douglas Fisher , "Iterative Optimization and Simplification of Hierarchical Clusterings" , Journal of Artificial Intelligence , 2007.

[7] Shadi Saleh ,Mosab Shaheen ,Zain Saqer , "Arabic Document Classifier" ,2013.

[8] Neelma Guduru, "Text mining with support vector machines and non negative matrix algorithms", 2006.

[9] Scott R. Hall, "Automatic text categorization applied to email", 2002.

[10] Zhu, M., Zhu, J., & Chen, W., "Effect analysis of dimension reduction on support vector machines", In the Proc. of the Natural Language Processing and Knowledge Engineering IEEE NLP-KE, Wuhan, China, pp. 592–596, 2005.

[11] Han J., and Kamber M., "Data Mining: Concepts and Techniques", (2nd Ed), the Morgan Kaufmann Series in Data Management Systems, 2006.

[12] Hill T., Lewicki P. , "STATISTICS Methods and Applications", 2007.

[13] An NSF Workshop: Language Engineering for Students and Professionals Integrating Research and Education, (2010, August), [Online]. Available: www.clsp.jhu.edu/ws99/projects/mt.

[14] Feldman R., Sanger J., "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", Cambridge University Press, 2007.

[15] Al-Shalabi R., Kannan G., Gharaibeh H., "Arabic text categorization using KNN algorithm", In the Proc. of Int. multi conf. on computer science and information technology CSIT06, 2006.

[16] Callan J., "human language technologies, text categorization", [Online]. Available: http://www.cs.waikato.ac.nz. 2004.

[17] Al-Harbi S., Almuhareb A., Al-Thubaity A., Khorsheed M., Al-Rajeh A., "Automatic Arabic Text Classification", In JADT'18, France, 2008, pp. 77-83.

[18] Al-Zoghby A., Eldin AS., Ismail NA., Hamza T., "Mining Arabic Text Using Soft Matching association rules", In the Int. Conf. on Computer Engineering & Systems, ICCES'17, 2007.

[19] Wai Lam, Miguel Ruiz , Padmini Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval" ,2000.

[20] Xindong Wu, Chengqi Zhang, and Shichao Zhang, "Efficient Mining of Both Positive and Negative Association Rules", ACM Transactions on Information Systems , 2004.

[21] Witten, Frank, Data mining: Practical machine learning tools and techniques with java implementations, 1999.

[22] Richard Alterman. Summarization in the Small. In N. Sharkey, editor, Advances in Cognitive Science, Chichester, England, 1986.

[23] Xiao-Chen Ma, Gui-Bin Yu, Liang Ma. "Multi-document summarization using clustering algorithm"; 2009.

[24] Erkan G, Radev D. ,"LexRank: graph-based centrality as salience in text summarization". J Artif Intell Res 2004;22:457–80.

[25] Wombudsman, “Effective handling of complaints” , [Online]. Available: http:// www.ombudsman.wa.gov.au/Agencies/Complaints_processes.htm, 2013 .
[26] Kuan C. Chen , Text Mining e-Complaints Data From e-Auction Store With Implications , 2009.
[27] Robert Francis,“The NHS hospital complaints system. A case for urgent treatment”, 2013.
[28] Wolfgang Himmel, Ulrich Reincke and Hans Wilhelm Michelmann , “Using Text Mining to Classify Lay requests to a Medical Expert Forum and to Prepare Semiautomatic Answers”, 2008.
[29] Jana Urdziková and Martina Jakábová, “Handling Customer Complaints Effectively”, 2011.
[30] Chakrabarty, Anirban, “A Framework for Medical Text Mining using a Novel Categorical Clustering Algorithm”, 2013.
[31] Nurfadhlina Mohd Sharef and Khairul Azhar Kasmiran , “Examining Text Categorization Methods for Incidents Analysis” , 2012.
[32] Ibrahim Sobh3, Nevine Darwish, Magda Fayek , “Evaluation Approaches for an Arabic Extractive Generic Text Summarization System “, 2008 .
[33] Douzidia FS, Lakhas Lapalme G. , “An Arabic summarizing system”. In: Proceedings of the document understanding conferences (DUC) workshop, DUC, p. 128–35, 2004.
[34] Hmeidi I., Hawashin B., El-Qawasmeh E., “Performance of KNN and SVM classifiers on full word Arabic articles”, Journal of Advanced Engineering Informatics 22, pp. 316–111, 2008.
[35] Yiming Yang and Xin Liu, “Re-examination methods of text categorization methods”, 2007.
[36] Wai Lam, Miguel Ruiz , Padmini Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval" ,2000.
[37] Chu-Hong Hoi , Michael R. Lyu , "A Novel Distance Similarity Measure on Learning Techniques & Comparison with Image Processing" 2012 .
[38] D. R. Wilson and T. R. Martinez. “Improved heterogeneous distance functions”. J. Artif. Intell. Res. (JAIR), 6:1(34, 1997).
[39] Jezek Karel, Steinberger Josef. Automatic text summarization: the state of the art 2007 and new challenges. Znalosti 2008:1–12.
[40] David Inouye, Kalita Jugal K. ,”Comparing Twitter summarization algorithms for multiple post summaries”. In: IEEE international conference on privacy, security, risk, and trust, and IEEE international conference on social, computing; 2011.
[41] Bettina Berendt,”Data Mining for Information Literacy” , 2011 .
[42] Vanderwende L, Suzuki H, Brockett C, Nenkova A. Beyond SumBasic: task-focused summarization with sentence simplification and lexical expansion. Inf Process Manage 2007;43(6):1606–18.
[43] Martin Hassel , Hercules Dalianis , “Automatic Text Summarizer” ,2003 .
[44] Leskovec J., Grobelnik M. and Frayling N. ,”Learning Semantic Sub-graphs for Document Summarization”,2002.
[45] Khoja S., Garside R., “Stemming Arabic text”, Computer Science Department, Lancaster University, Lancaster, UK, 1999.
[46] Burges, C., “A Tutorial on Support Vector Machines for Pattern Recognition, data mining and Pattern Recognition” , 2006.
[47] Mesleh A., “Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System”, Journal of Computer Science, 1(6), pp. 411-435, 2007.
[48] Richard Alterman. “Text Summarization”. In S. C. Shapiro, editor, Encyclopedia of Arti_cial Intelligence, volume 2, pages 1579{1587. John Wiley & Sons, Inc., 1992.

[49] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C. (2002). “Text Classification with String Kernels” ,Journal of Machine Learning Research, 2 (2) 419-444, 2008.
[50] Black, Paul E., "Levenshtein distance", Dictionary of Algorithms and Data Structures [online], U.S. National Institute of Standards and Technology, 2008.