



Imam Mohammad Bin Saud Islamic University College of  
Computer Science and Information Department of  
Information Systems

# Ra'ai Analyzer an Arabic Sentiment Analyzer for Twitter



**By:**

**Name**  
Mawaheb Altuwijri  
Tarfa Albuhairi  
Wejdan Alohaideb

**ID**  
431020432  
431042812  
431020515

**Supervisor:** Dr. Sarah Alhumoud

Project Submitted in Partial Fulfillment for the Degree of B.Sc. In "Computer Science"

Semester 1-2014

# Ra'ai Analyzer an Arabic Sentiment Analyzer for Twitter

**By:**

<b>Name</b>	<b>ID</b>
Mawaheb Altuwijri	431020432
Tarfa Albuhairi	431042812
Wejdan Alohaideb	431020515

We hereby certify that this project satisfies the project requirements



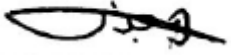
<b>Date of Approval</b>	<b>Approved by:</b>
	<b>Dr. Sarah Alhumoud</b>
<b>Signature</b>	

<b>Date of Approval</b>	<b>Approved by:</b>
	<b>Vice Dean of Educational Affairs</b>
<b>Signature</b>	

## Declaration

We Mawaheb Altuwijri, Tarfa Albuhairi, and Wejdan Alohaideb being members of the final year project of Ra'ai Analyzer an Arabic Sentiment Analyzer for Twitter, declare that this report contains only work accomplished by the members of our the group except for information obtained in a legitimate way from literature, company or university sources. All information from these other sources has been duly referenced and acknowledged in accordance with the University Policy on Plagiarism.

Furthermore, we declare that in completing the project, the individual group members had the following responsibilities and contributed in the following proportions to the final outcomes of the project:

<b>Student ID</b>	<b>Responsibility</b>	<b>% Contributed</b>	<b>Signature</b>
431020432	<ul style="list-style-type: none"><li>▪ Analysis.</li><li>▪ Implementation.</li><li>▪ Documentation.</li><li>▪ Testing.</li></ul>	33.33%	
431042812	<ul style="list-style-type: none"><li>▪ Analysis.</li><li>▪ Implementation.</li><li>▪ Documentation.</li><li>▪ Testing.</li></ul>	33.33%	
431020515	<ul style="list-style-type: none"><li>▪ Analysis.</li><li>▪ Implementation.</li><li>▪ Documentation.</li><li>▪ Testing.</li></ul>	33.33%	

## **Acknowledgment**

Foremost, we would like to offer our grateful to Almighty ALLAH for helping us to complete this project. Then we would to thank our families who supported and helped us all the time. Also, we would like to offer our sincere thanks and gratitude to our supervisor Dr. Sarah Alhumoud for her advice, supervision, guidance, encouragement and support. We also present our thanks to teacher Haifa Al Dayel, who explained some concepts that related to the project. And we would like to take this opportunity to thank all people who helps us during this project from our families and friends. At the end, a great thank to our university and especially our college for all their support and help.

## Abstract

Data has become the currency of this era as it is continuing its massive increase in size and generation rate. These big data can be of great value for organizations when analyzed properly. This research represents an implementation of sentiment analysis on Twitter's tweets which is one of the biggest public and freely available big data sources. It will analyze Arabic (Saudi dialect) tweets to extract sentiments of these tweets toward a specific topic. It used a small dataset consisting of 1000 tweets collected from Twitter. The collected tweets were analyzed using two approaches, supervised which based on machine learning and unsupervised which based on lexicon. The supervised approach used three algorithms which are *Support Vector Machine* (SVM), *Naive Bayes* (NB) and *K-Nearest Neighbor* (KNN). The obtained results by the cross validation option on the same dataset clearly confirm the superiority of the supervised approach namely SVM, NB and KNN with accuracies' degree of 98%, 95% and 94%, consecutively. While the unsupervised classifier has scored 78% of accuracy.

## Arabic Abstract (ملخص عربي)

أصبحت البيانات هي عملة هذا العصر بسبب تزايدها المستمر، وهذه البيانات يطلق عليها البيانات الضخمة، والتي يمكن أن تكون مفيدة للمؤسسات، والحكومات، وكذلك الباحثين إذا تم تحليلها. هذا البحث يُطلق عليه "رأي"، والذي يقوم بتحليل الرأي الوارد في تغريدات موقع تويتر الذي يعتبر أحد أكبر مصادر البيانات الضخمة. يقوم "رأي" بتحليل التغريدات العربية ذات اللهجة السعودية العامية، لاستخراج رأي الأشخاص حول موضوع واحد. تم تحليل مجموعة من التغريدات يبلغ عددها 1000 تغريدة جُمعت من موقع تويتر. وقد أُستخدم نهجين في تحليلها، أولهما: النهج الإشرافي "Supervised" الذي يعتمد على تعليم الآلة، والنهج الثاني: غير الإشرافي "Unsupervised" الذي يعتمد على قاموس الكلمات. تم استخدام ثلاث خوارزميات في نهج الإشرافي بناء المصنفات، وهي: SVM، NB وKNN. أما في النهج غير الإشرافي فقد تم بناء خوارزمية خاصة بهذه الطريقة، واستخدمت أربعة قواميس تم بناؤها لتخدم اللهجة المستهدفة. أظهرت نتائج النهج الإشرافي باستخدام اختبار "Cross-Validation" أن أفضل نسبة دقة تصنيف حققها المصنف SVM حيث بلغت 98%، بينما حقق NB نسبة 95%، ويأتي KNN آخرًا بنسبة 94%. بلغت نسبة دقة المصنف غير الإشرافي 78%، مما يؤكد تفوق النهج الإشرافي على النهج غير الإشرافي.

## **Keywords**

Sentiment Analysis, Data Mining, Big Data, Machine Learning, Supervised Approach, Unsupervised Approach, Support Vector Machine, Naïve Bayes, K-Nearest Neighbor.

## List of Abbreviation

API: Application Program Interface.

MSA: Modern Standard Arabic.

NLP: Natural Language Processing.

POS: Part of Speech.

RA: Ra'ai Analyzer.

SA: Sentiment Analysis.

SVM: Support Vector Machine.

NB: Naïve Bayes.

KNN: K-Nearest Neighbor.

D-TREE: Decision Tree.

MaxEnt: Maximum Entropy.

TF-IDF: Term Frequency – Inverse Document Frequency.

URL: Uniform Resource Locator.

WWW: World Wide Web.

ZB: Zetta Bytes.



# Table of Contents

CHAPTER ONE: INTRODUCTION .....	1
1.1 Introduction.....	2
1.2 Problem Definition.....	3
1.3 Objectives.....	3
1.4 Targeted Users .....	3
1.5 Project Scope.....	3
1.6 Conclusion .....	4
CHAPTER TWO: LETERAURE REVIEW .....	5
2.1 Introduction.....	6
2.2 Overview of Data Mining .....	6
2.2.1 History.....	6
2.2.2 Functions.....	6
2.2.3 Techniques .....	6
2.2.4 Research Trends .....	6
2.2.5 Applications .....	7
2.3 Text Mining.....	7
2.4 Overview of Sentiment Analysis .....	8
2.5 General Process Flow of Sentiment Analysis in Twitter .....	8
2.5.1 Collecting.....	9
2.5.2 Preprocessing .....	9
2.5.3 Filtering.....	9
2.5.4 Classifying .....	10
2.6 Related Work .....	13
2.6.1 Related Work on Sentiment Analysis .....	13
2.6.2 Related Work on Twitter Sentiment Analysis.....	15
2.6.3 Related Applications .....	18
2.7 Conclusion .....	18
CHAPTER THREE: ANALYSIS AND METHODOLOGY .....	19
3.1 Introduction.....	20
3.2 User Requirements.....	20
3.2.1 Functional Requirements .....	20
3.2.2 Non Functional Requirements.....	20
3.3 System Requirements.....	20
3.3.1 Functional Requirements .....	20

3.3.2 Non Functional Requirements.....	21
3.3.3 System Characteristic.....	21
3.4 Methodology .....	22
3.5 Timeline .....	22
3.5.1 Description of Project Stages.....	23
3.5.1.1 Roles and Responsibilities .....	23
3.6 Functional modeling .....	24
3.6.1 Use Cases.....	24
3.6.2 Activity diagram .....	28
3.7 Conclusion .....	28
CHAPTER FOUR: DESIGN AND IMPLEMENTATION.....	29
4.1 Introduction.....	30
4.2 System Design.....	30
4.2.1 Modular Decomposition.....	30
4.2.2 Architectural Design .....	31
3.2.2.1 Collecting Tweets Component:.....	31
3.2.2.2 Preprocessing Component:.....	31
3.2.2.3 Filtering Component: .....	31
3.2.2.4 Classifying Component:.....	32
3.2.2.4.1 Supervised Approach .....	32
3.2.2.4.2 Unsupervised Approach .....	32
4.2.3 Logo Design.....	33
4.2.4 Interface Design .....	33
4.3 Implementation .....	34
4.3.1 Programming Languages .....	34
4.3.2 Tools.....	34
4.3.4 Package and Classes Description.....	35
4.3.5 Procedures Description .....	36
4.4 Conclusion .....	36
CHAPTER FIVE: TESTING.....	37
5.1 Introduction.....	38
5.2 Supervised Testing.....	38
5.2.1 NGram Comparison Test .....	38
5.2.2 Test Options Comparison.....	41
5.2.3 Detailed Accuracy by Class .....	41
5.2.4 Time Cost Comparison between Classifiers .....	43
5.2.5 Stop Words Effect over Classifiers Time.....	44

5.3 Unsupervised Testing..... 45

    5.3.1 Unsupervised Functions Testing..... 45

5.4 Comparing between Supervised and Unsupervised..... 47

5.5 Conclusion ..... 47

CHAPTER SIX: CONCLUSION AND FUTURE WORK..... 48

6.1 Introduction..... 48

6.2 Challenges..... 48

6.3 Acquired Skills..... 49

6.4 Future Work..... 49

6.5 Conclusion ..... 49

## Table of Figures

Figure 2. 1: Relation between SA, data mining and text mining .....	7
Figure 2. 2: General process flow of SA in Twitter .....	8
Figure 2. 3: SA approaches .....	10
Figure 2. 4: Classification of data (supervised approach) .....	11
Figure 2. 5: Clustering of data (unsupervised approach) .....	11
Figure 2. 6: Defining the boundaries between data (semi supervised approach).....	11
Figure 2. 7: Most used data sources for Arabic SA .....	14
Figure 2. 8: Most used Arabic dialect in SA .....	15
Figure 3.1: Modified water fall model. ....	22
Figure 3. 2: Project timeline.....	22
Figure 3. 3: Use cases diagram.....	24
Figure 3. 4: Activity diagram .....	28
Figure 4. 1: System architecture.....	30
Figure 4. 2: RA logo.....	33
Figure 5. 1: The effect of use UniGram and BiGram on SVM.....	39
Figure 5. 2: The effect of use UniGram and BiGram on NB .....	39
Figure 5. 3: The effect of use UniGram and BiGram on IBK.....	40
Figure 5. 4: The effect of UniGram on classifiers accuracy .....	40
Figure 5. 5: The effect of BiGram on classifiers accuracy.....	40
Figure 5. 6: Accuracy of classifiers based on the used test options. ....	41
Figure 5. 7: Detailed accuracy by class.....	42
Figure 5. 8: Code reults .....	42
Figure 5. 9: WEKA GUI results.....	43
Figure 5. 10: Dataset size effect on time cost in seconds.....	44
Figure 5. 11: Stop words effect on classification time in seconds. ....	44
Figure 5. 12: Overall sentiment bar plot. ....	46
Figure 5. 13: Words cloud function. ....	46
Figure 5. 14: View tweets function .....	46
Figure 5. 15: Comparing the results and process flow of supervised and unsupervised classifiers.....	47

## Table of Tables

Table 2. 1: Comparison between supervised and unsupervised approach .....	13
Table 2. 2: Data sources for published papers on Arabic SA .....	14
Table 2. 3: Comparison between techniques that were used for Arabic SA in terms of SA stages .....	17
Table 3. 1: Team’s roles and responsibilities. ....	23
Table 3. 2: Enter keyword use case. ....	25
Table 3. 3: Validate keyword use case. ....	25
Table 3. 4: Retrieve tweets use case. ....	26
Table 3. 5: Extract text field use case. ....	26
Table 3. 6: Preprocess tweets use case. ....	26
Table 3. 7: Filter tweets use case. ....	26
Table 3. 8: Classify tweets use case. ....	27
Table 3. 9: View overall sentiment use case. ....	27
Table 3. 10: View the words cloud use case. ....	27
Table 3. 11: View tweets info use case. ....	28
Table 3. 12: View user manual use case. ....	28
Table 3. 13: View policy use case. ....	28
Table 4. 1: List of the used packages. ....	35
Table 4. 2: List of the used classes. ....	35
Table 4. 3: List of the unsupervised classifier procedures. ....	36
Table 5. 1: Classifier accuracy. ....	39
Table 5. 2: Test options comparison. ....	41
Table 5. 3: Detailed accuracy by class. ....	42
Table 5. 4: Time cost over dataset size .....	43
Table 5. 5: Stop words effect on classification time in seconds. ....	44
Table 5. 6: Example of classifying tweets.....	45
Table 5. 7: Requirements test results .....	45

# **CHAPTER ONE: INTRODUCTION**

# 1.1 Introduction

The data available online is doubling in size every two years [1]. While the amount of online data that was generated in 2013 was 4.4 Zettabytes (ZB). Moreover, in 2020 it will reach 44 ZB [1]. Individual users are the main source, contributing 75% to the overall produced data [2].

Big data could be described by the 3'Vs model. Those are variety, velocity and volume. Variety means the variation of data available online that include both structured and the unstructured data such as emails, videos, audios, images, click streams, logs, posts, search queries. Velocity refers to how the processing and storing of these huge and complex data need to be fast to accommodate the increasing and continuous requests. Volume indicates the massive size of generated data, as stated previously [3].

Big data can be found in social networks such as Twitter and Facebook. Social networks have become a popular means for cyber communication among the society. Since the foundation of Twitter in 2006 it has provided the ability to freely, easily, and instantaneously express, reach, and share opinions and feelings in public in an SMS style text, tweets. Twitter is a micro blogging social site, which contains tweets; each tweet has 140 characters or less [4]. There are over 1 billion Tweets every 72 hours from more than 140 million active users out of 200 million users in 2012 [5]. In Twitter, there are more than 6.5 Million Arabic users who produce more than 10.8 million tweets per day [5] out of 175 million tweets from all over the world as shown in Figure 1.

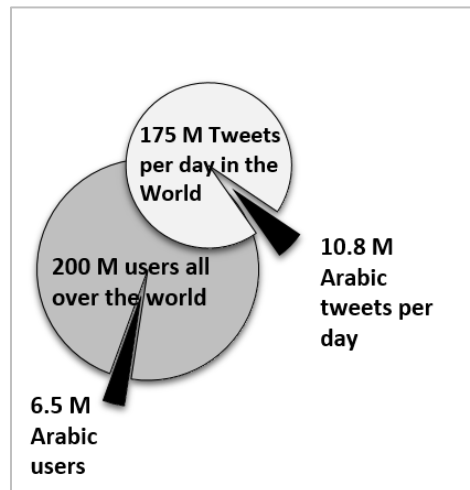


Figure 1: Arab usage in Twitter.

A combination of techniques such as Data Mining, Text Mining, Machine Learning and Natural Language Processing (NLP) are used to extract valuable information for taking corporate decisions. Data mining is the exploration and analysis of large amounts of data [6]. The process of extracting sentiments relies first on data mining techniques to find patterns then Sentiment Analysis (SA) techniques are applied on these patterns.

SA is a one of the NLP concepts, which is also called opinion mining [7]. This field of computer science is used to extract sentiment out of text giving useful information about the author and his/her tendency toward a specific topic. Two approaches can be used in SA, supervised approach, known as corpus-based approach. The supervised approach uses machine learning algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (D-Tree) and K-Nearest Neighbor (KNN). The second approach is unsupervised approach, also known as lexicon-based. The polarity of a word is based on a dictionary where each word is associated with a polarity value (+1, -1 or 0 for positive, negative or neutral, respectively) [8].

This research represents Ra'ai Analyzer (RA) that performs sentiment analysis on Arabic content in Twitter. Ra'ai (رأي) is the Arabic word that means opinion or sentiment. RA extracts the sentiment from Arabic tweets (Saudi dialect) to get the outcome of users' tendencies toward a specific topic.

This research will discuss the following sections: Related work, Project's problem definition, objectives, methodology, results and dissection; and finally the conclusion.

## 1.2 Problem Definition

Many organizations need to know clients tendencies, feedback and opinion in order to improve their services and products. SA will aid in giving the organization the ability to understand clients' tendencies toward a product or a service. Organizations usually resort to conduct interviews directly with clients or distribute questionnaires to collect clients' feedback. That drains a lot of time, effort and cost. In addition, it may not serve as a precise indication to actual costumers' behavior and preferences as the questionnaire questions may not cover all needs or it may not be answered thoughtfully and accurately. Moreover, clients may not express their immediate feedback openly in a questionnaire compared to what they do in Twitter. This hides valuable information that would benefit the organization.

From this problem the idea of this project appeared. Harvesting clients' opinions from the free public space will aid in meeting organizations needs of feedback in an easy and inexpensive way. This is done through building a classifier to analyze clients' sentiments, then classify them into positive and negative to produce the overall sentiment.

## 1.3 Objectives

The objectives of Ra'ai analyzer are:

1. Provide client's tendencies about a specific topic this enables a researcher to use Twitter's users as a sample of the society.
2. Get customers' feedback about a specific product to enable corporations to observe product's weaknesses and strengths to deal with and reinforce respectively.
3. Find the highest classification process accuracy.
4. Produce a user interface that facilitates using the RA such as:
  - Produce a bar plot for the overall sentiment that represents the number of positive and negative classified tweets.
  - Produce a word cloud that shows the most frequent words that appeared in tweets containing the entered keyword. Where the word's size is relevant to its occurrence frequency relative to other words in the cloud.
  - Give the user the ability to look at the collected tweets and show him/her how it was classified.

## 1.4 Targeted Users

Users who can benefit of Ra'ai Analyzer are:

1. Researchers.
2. Companies.
3. Public organizations.
4. In addition, anyone who have an interest for a specific topic.

## 1.5 Project Scope

RA works on Twitter's Arabic (Saudi slang) tweets through collecting tweets of football domain (Hilal team).



## **1.6 Conclusion**

This chapter has given an introduction about the project. It discussed the importance of data and how this project will make the advantage out of analyzing it. This chapter also highlighted the problem and proposed the solution. Explained the advantages of the project, for whom it could be useful and explained the scope it covers that is, Arabic scope only. Following chapters will explain more about the idea of the project and how it works.

## **CHAPTER TWO: LETERAURE REVIEW**

## **2.1 Introduction**

Sentiment analysis is a subtopic in data mining. Sentiment analysis manipulates a text dataset in order to extract the sentiment of the entire dataset. While data mining is the exploration and analysis of large amount of data [6]. This chapter introduces the related fields and related work to this project. Starting by a brief overview on data mining. Followed by a discussion on text mining and its relation to sentiment analysis concept. Then, an overview on sentiment analysis. After that, description for the general process flow of sentiment analysis in Twitter. Finally, a preview to the related work on the project's field.

## **2.2 Overview of Data Mining**

Data mining is the process of discovering the knowledge from a large collection of data. It has a number of functionalities that are needed to specify pattern types. Pattern types could be found through data mining tasks, which are classified into two methods descriptive and predictive. Descriptive mining tasks are the processes of characterizing properties of data in a target data set. Predictive mining tasks perform induction on the current data in order to make predictions [6].

### **2.2.1 History**

The concept of data mining was a demand developed over time. It was an improvement of other concepts and techniques introduced before it. Starting with the evolution of storing data on computers, and tape in 1960 [6]. That brings another demand to arrange and sort these data using relational databases and structured query languages in 1980 [6]. In 1990, the concept of data warehouse has got common among data bases' operators [6]. It was a need to manage large number of databases. With the growth of data warehousing by the contribution of online analytic processing and multidimensional databases the need of data mining was even more defined [6].

### **2.2.2 Functions**

Data mining has numbers of functionalities. Each function manipulates data for main purpose. One of the functions is data characterization, which is used for summarizing the common characteristics or features of data. Another function is data discrimination that is used to compare the common features of one data class with other data classes' features. The main goal of frequent patterns function it to discover the patterns that occur frequently in data. Classification is one of data mining functionalities where its purpose is to construct a model (or function) to find the differences among data classes. Clustering is the function of grouping data into groups based on the similarities between data and without consulting class labels (model) [6].

### **2.2.3 Techniques**

Data mining uses many techniques that helps with performing its functionalities such statistics, machine learning, pattern recognition, database, data warehouse systems, information retrieval, visualization, algorithms, high-performance computing, and many more other methods and techniques [6].

### **2.2.4 Research Trends**

The use of data mining has expanded and researchers in the field took many directions. The latest trends in data mining are varied and branched. Web data mining with

the main concern of how to extract the hidden and valuable knowledge from the *World Wide Web* (WWW) [9]. Another recent research direction on data mining is Intrusion detection which is a network security critical issue. Intrusion threatens network resources integrity, confidentiality, or availability. Authors in [10] proposed an intrusion detection system using data mining. Health care sector used data mining to tackle the process of patients' records to decide the right decision for each patient. Authors in [11] present a study where they used data mining to predict the survival percentage for each patient according to his/her preoperative info and tests results.

## 2.2.5 Applications

Applications of data mining have affected many aspects in life. For business, data mining can be applied on customer data such as in Italy; AC Nielsen Company has applied data mining on supermarkets to find the most frequent combinations of products bought by customers. That enabled them to place these products in close proximities to increase profits [12]. While in the banking sector, data mining was used by bank of America as an example. They used data mining to find the qualified low-income and minority customers to ensure the compliance with the Fair Housing Act [12]. Another application of data mining used in the fire department in New York City. The department employed data mining to predict the next building that may go up in flames. Though specifying an algorithm that chooses set of buildings that can be under risk. The algorithm picks buildings based on 60 factors that may be available in the building to be considered under risk such as Low-income neighborhoods [13]. Finally, another application of data mining is in social networks where sentiment of data is extracted to bring valuable knowledge that is used to decide better business decisions. The process of extracting sentiment out of data is called sentiment analysis, and will be described in the section 2.4.

## 2.3 Text Mining

Text mining is the process of applying data mining methods and functions on an unstructured text dataset to discover and extract useful information [14]. It encompasses everything from collecting text data whether from documents or WebPages to text classification and clustering [14]. While text mining manipulates on natural language text, it used NLP to find and extract the meaning of the text. NLP uses linguistic concepts such as *Part-Of-Speech* (POS) of the text (noun, verb, adjective...), grammatical structure and other techniques to understand the text [14]. One of the major tasks for text mining is sentiment analysis, which will be explained in section 2.4. After some research, the project team has concluded the relation between data mining and text mining which is represented in Figure 2.1. The figure shows that SA can be done by combining some of the data mining techniques along with text mining techniques.

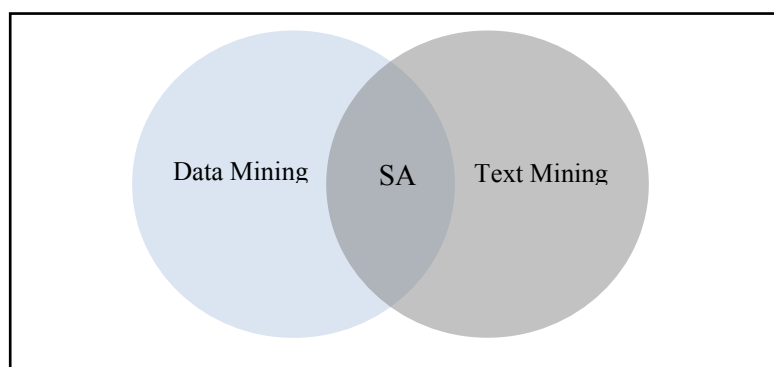


Figure 2. 1: Relation between SA, data mining and text mining

## 2.4 Overview of Sentiment Analysis

Sentiment analysis or opinion mining is the process of classifying a text dataset into positive, negative and sometimes neutral [7]. It is one of the text mining search trends and application that uses NLP to capture the sentiment or polarity of the text [7].

There were previous researches on sentiments and opinions since 1992 where researchers proposed the idea of mining text [15]. However, the term sentiment analysis has first appeared in 2003. Since the year 2000, studies on sentiment analysis has grown rapidly to become one of the most active research areas in data science [7]. There are many reasons that aided in the development of sentiment analysis research. First, the broad applications that affect every aspect in life from customer products to political elections. That makes a great motivation for researchers, individuals and corporations. Second, there is a huge volume of sentimental data in the social media such as Facebook, Twitter, and Tumblr. This huge amount of data has a scientific, business; political, national and governmental value that worth analyzing and understanding [7].

## 2.5 General Process Flow of Sentiment Analysis in Twitter

Sentiment analysis of data can be accomplished by passing through four main steps as shown in Figure 2. 2 [16]. The process starts by specifying sentiment keyword under study. After that is the collecting stage and there are different ways to collect target tweets. These collected tweets will be stored in a dataset. After collecting the tweets, the next step is preprocessing the tweets which will remove all unrelated contents and get the Arabic text only. Then, the filtering step that involves removing all the words that do not affect the meaning of the text. The next stage is to classify the content to positive and negative. The final step is to get the overall sentiment of all the collected tweets. These stages will be described in more detail in the following sections.

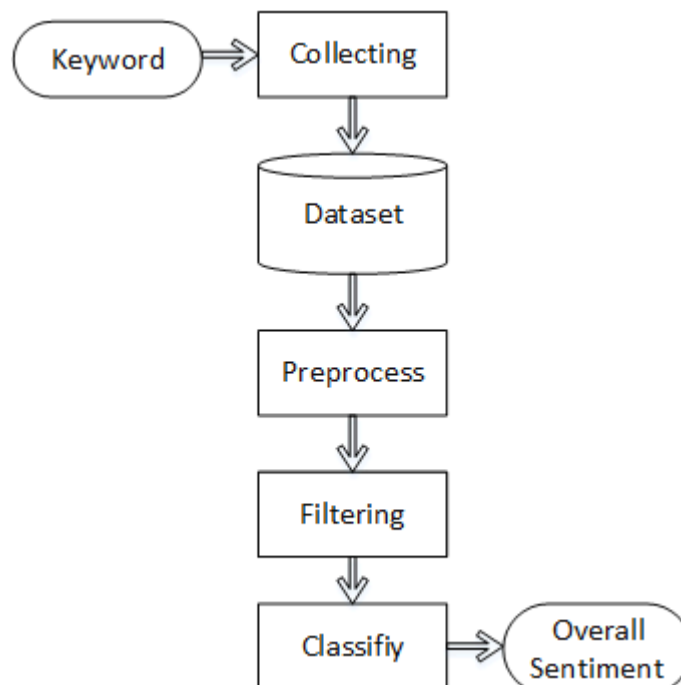


Figure 2. 2: General process flow of SA in Twitter

## 2.5.1 Collecting

The first step in the sentiment analysis process is collecting tweets by specifying a keyword to retrieve all tweets that are related to that keyword. Tweets could be collected either using tweet crawler or Twitter APIs. Tweet crawler collects a group of linked tweets by querying the Twitter web service. Twitter API is an *Application Program Interface* (API) provided by Twitter that gives developers the ability to use the Twitter's functions such as retrieving tweets with the selected keyword and language. The collected tweets will be stored in a dataset in order to classify it.

## 2.5.2 Preprocessing

Preprocess is a technique used to clean text from unrelated contents, such as user-names, images, hash tags, URLs and all non-Arabic words or sometimes gagging these content with a unified name [8], [17], [18]. This process referred to as tagging [18]. Tagging is the process of marking unrelated content in a tweet that does not have any impact on the tweet sentiment. These will differ in type and number. For example, a URL link may be replaced with a URL tag, the username, which is a word that appears after the symbol "@" in Twitter will be tagged with USERNAME and the word that appears after hash "#" and do not relate to the topic will be tagged with hashtag. Also since Twitter users use symbols such as ":" and "☺" to express their opinions these emoticons expresses valuable information to the sentiment. Therefore, in order to extract the sentiment out of the emotions they are tagged as well. For example an emotion is tagged as HAPPY if the used symbols are":)", ":)" and tagged as SAD if the used symbols are ":((", ":((" [18]. Emotions tags will affect the classify process since they hold a sentiment.

## 2.5.3 Filtering

After the preprocessing stage, the outcome will be only text. Filtering stage includes other steps that are needed to remove all the words that do not affect or relate to the meaning. Moreover, in this stage, misspellings are corrected and the repeated letters of the text are removed.

### A. Misspelling

Users may misspell some words because of their fast typing or even the weakness in spelling skills. In order to overcome this problem misspelling can be corrected manually or by using tools.

### B. Repeated Letters

Users express their feeling about something like a product and they may use a word with some repeated letters in it such as "كثييير". Removing repeated letters from the words is important for the classification process to recognize the word. For example, "كثييير" will be corrected to be "كثير" where the letter "ي" is repeated four times. The repeated letters will be replaced by one letter of those repeated letters. Using naive algorithm which simply counts the number of letters in each word if the letter repeated then the repeated letters will be removed and one letter is kept [17].

## C. Stop Words

Stop Words are a group of words that do not affect the meaning of the text, such as prepositions. The problem with Arabic language is the limitation of the built in stop words lists. One available Arabic stop word list is Khoja stemmer tool [17]. Also if the post or tweet that need to be analyzed use a *Modern Standard Arabic* (MSA) then there will be a need to add extra stop-words to the list of stop words from different Arabic dialects. Arabic dialects that have been studied so far are the Egyptian dialect [8] and Jordanian dialect [17].

## D. Normalization

Normalization is the process of replacing similar letters that are used interchangeably by one of them and removing none letters from the words, the normalization conditions are as follows:

1. Remove any punctuation from the string such as ( . “” ; ’).
2. Remove any diacritics (short vowels) such as ( َ ُ ُ )
3. Remove non-letters from the word such as (+ = ~ \$)
4. Replace “؛” , “،” , and “ آ ” with bare alif "ا" regardless of position in the word [8].
5. Remove "Tatweel" which is the process of machining the letter longer than normal by using “-”. For example, using Tatweel the word “كتب” may look like "كـتـبـ".
6. Replace final “ى” with "ي".
7. Replace final “ة” with "ه" [8].
8. If a word starts with “ء” then replace it with “ا”.
10. Replace “و” and “ؤ” with “و”.
11. Replace “ئ” and “يء” with “ي” [18].

### 2.5.4 Classifying

This stage represents the final stage where each tweet will be classified as positive and negative by the classifier. These tweets will be annotated manually to be compared with classifier results in order to examine the accuracy of the classifier.

Sentiment analysis classifies the text using different approaches as shows in Figure 2.3. Which are machine learning, lexicon based, and semi-supervised. Each approach will be explained in the following sections.

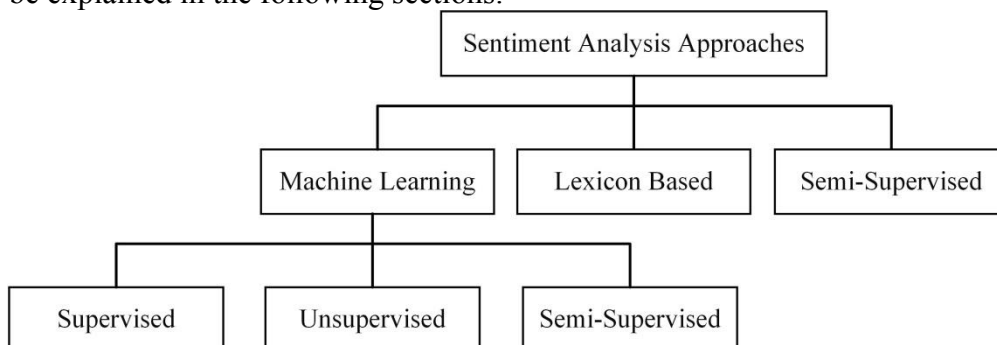


Figure 2. 3: SA approaches

### 2.5.4.1 Machine Learning Approach

Machine learning could be categorized into three approaches that are Supervised, unsupervised and semi supervised. The supervised approach apply the classification data mining function and unsupervised apply the cluster data mining function while the semi-supervised combines both the supervised and the unsupervised approaches to define the boundaries between the data classes [6] The figure 2.4, 2.5, and 2.6 show the difference between the three approaches.

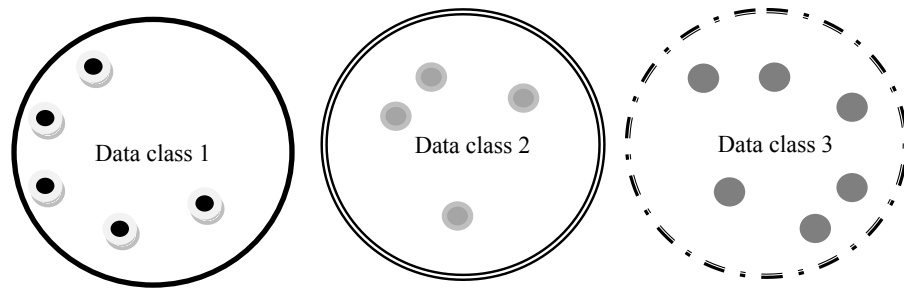


Figure 2. 4: Classification of data (supervised approach)

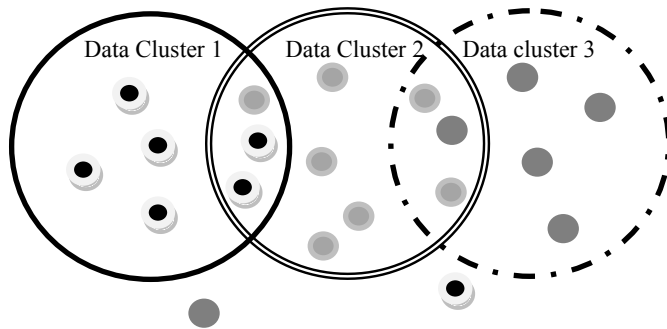


Figure 2. 5: Clustering of data (unsupervised approach)

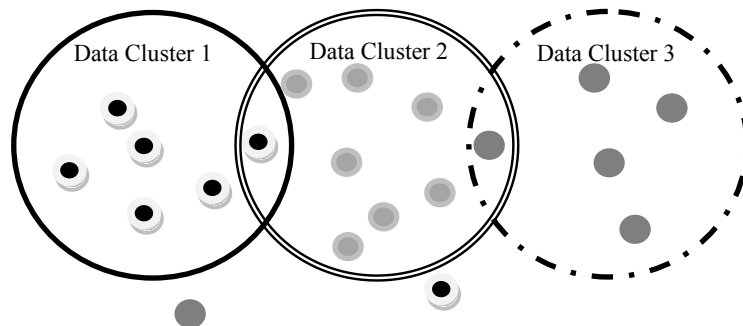


Figure 2. 6: Defining the boundaries between data (semi supervised approach)



Supervised or corpus-based approach works by first training the classifier using a training dataset. This dataset contains tweets with positive and negative labels. It teaches the algorithms which tweet is positive and which is negative. After training the classifier it will be able to build a general model to be used in classifying a new dataset. Classification using feature vector [8] is considered under supervised approach. For this method, features must be extracted first and this can be done using UniGram, BiGram, or TriGram. Unigram technique deals with each word as a single unit without considering what is surrounding [17]. The extracted feature must exceed a threshold value to be extracted. The supervised approach tends to have more accuracy results when it is tested on a dataset that have the same domain of the training dataset. Accuracy can be increased by training with huge dataset and extracting multiple features.

Machine learning algorithms such as *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *Decision Tree* (D-Tree), *K-Nearest Neighbor* (KNN) are used in the supervised approach. SVM algorithm is widely used for binary classification. It uses linear conditions in order to classify data into classes. The idea behind using the linear condition is to separate the two classes from each other as well as possible [19].

NB is a probabilistic algorithm. It first counts the elements of each class in the training dataset. Then when a new data arrive it set a pre-assumption that the probability of its class will be of the class that has more elements in the training dataset. After that it finds the common features between the new data and each class on the training dataset. Finally the posterior possibility of a class is calculated to decide the class of the new data [19].

D-Tree algorithm builds a hierarchical decomposition from the training dataset by setting a condition on the attribute value to divide the dataset. The condition or predicate is the presence or absence of one or more words. The training dataset will continue to be divided recursively until the leaf nodes contain certain minimum numbers of records which are used to classify the aim of them [19].

KNN algorithm does not set any assumptions on the training dataset. The data will be classified by the class election by its neighbors. The data will have the class that has got the most votes [19]. KNN and IBK are used interchangeably to refer to the KNN algorithm by the papers' authors.

#### **2.5.4.2 Lexicon Based Approach**

Lexicon based and also called unsupervised is the second type of classification. It uses a lexicon or dictionary where there is no training step. The classifier will classify a dataset directly using a dictionary of words. Each word has a polarity +1, -1 or 0 for positive, negative or neutral, respectively [8]. The dictionaries could be built manually or it can be built beforehand. Since there is a lack for Arabic dictionaries researchers often built their own dictionary. The dictionary (lexicon) may contain more than the polarity of the word according to the classifier whether to consider the POS of the word within the tweet such as verb, noun, adjective, adverb, and others [20], or just use the polarity inside the dictionary (lexicon). Table 2.1 shows the main differences between the performing sentiment analysis using supervised and lexicon based approaches. In this project, both approaches will be examined.

### 2.5.4.3 Semi Supervised Approach

This approach uses both the supervised machine learning approach along with lexicon based approach. Where there will be a training dataset, test Dataset, and word dictionary.

Factors	Supervised (Machine Learning)	Unsupervised (Lexicon Based)
Ability to classify different topics	Restricted to limited number of topics.	No limitation for the topics that can be classified.
Accuracy depends on	<ul style="list-style-type: none"> <li>- The classified topic is known.</li> <li>- Size of the training dataset.</li> <li>- The features are varied.</li> <li>- Effective features</li> </ul>	The size of the data dictionary (lexicon).
The features that can be used	<ul style="list-style-type: none"> <li>- POS.</li> <li>- Word Frequency.</li> <li>- Sentiment Words.</li> </ul>	POS

Table 2. 1: Comparison between supervised and unsupervised approach

## 2.6 Related Work

In this section, a list of related works to sentiment analysis field on both Arabic and English languages will be discussed. Section 2.6.1 reviewed the related work on sentiment analysis in general while Section 2.6.2 reviewed the papers that have use sentiment analysis to analyze Arabic content on Twitter. The related work has also covered some of the available application in Section 2.6.3.

### 2.6.1 Related Work on Sentiment Analysis

The proposed work in research [21], [22], [23] have been done over English web content. As in [21] they started by extracting snippets from the web contents. Then scoring the sentiment of each snippet based on different categories. To evaluate the importance of a word in a category they used point wise mutual information and mutual support. After that, the word with highest frequent value and highest point wise mutual information value is chosen as the topic.

Paper [22] presents sentiment phrase classification vector an algorithm that compares the similarity between document vectors. Then mines the theme of the document and judges the document theme attributes. The algorithm shows better effectiveness and practicality.

SenseNet is a tool developed and presented in paper [23]. The purpose is to assign each sentence a numerical valence values and output sense value. Paragraph will be the input of SenseNet which will be divided into a set of sentences and each sentence is further divided into three parts. A Valence value will be assigned to each word in the three parts of sentence. These triplets are then processed to calculate the sentence level sentiment valence.

Research that has been done in paper [24] and [25] concentrated on analyzing the sentiment of Twitter English content. Both projects have collected Tweets automatically using Twitter API. Paper [24] have annotated the tweets to positive and negative manually. They preprocessed them by removing URL, avoiding misspellings and slang words. They used the supervised approach through using Naive Bayes, SVM, and *Maximum Entropy* (MaxEnt) algorithm.

While research in [25] used tagging technique to remove the unrelated content. They used different Dataset sizes to figure the best classifier among the available classifiers. They started by dividing data into neutral, polar and irrelevant. The polar data will be classified to positive or negative. They used WEKA tool for data mining and machine learning algorithms to perform the test.

In research [26] the system consists of two components. First component is an online game to build the lexicon and make it up to date. This is done by making the players responsible to give each word a positive or negative polarity. The second component is the sentiment application where they used the unsupervised approach. They tested three algorithms to analyze sentiment which are pattern matching approach, majority approach with entities, and majority approach without taking entities into account. Their precision measurements shows that pattern matching approach has the least rate while both , majority approach with entities, and majority approach without taking entities into account have convergent rates of precision.

Another work by paper [27] where they introduced a new way to perform the sentiment analysis. Through associating with each, term a limited number of grammars that express sentiments according to the domain. They used statistical models to extract the financial terms, and using the term’s popular collocations, they build a grammar associated with each term.

The effort of research [28] conducted the sentiment in addition to subjectivity analysis over MSA. They used two tiers of classification. First tier; classify the subjective and objective text using binary classifier. Second tier; classify each sentence to positive and negative using SVM classifier. They manually mark the dataset using the Penn Arabic Treebank which constructs a new Arabic polarity lexicon.

A survey based on IEEE, ACM and ScienceDirect has been done to find the most used data sources for Arabic SA with a total of 40 papers. Figure 2.7 shows that social media with Twitter and Facebook have been used more for SA. Table 2.2 reviews the number of the scanned papers on each data source and in Appendix E a list of the scanned papers are listed. The most used Arabic dialect in SA show in Figure 2.8.

Number of Published Papers

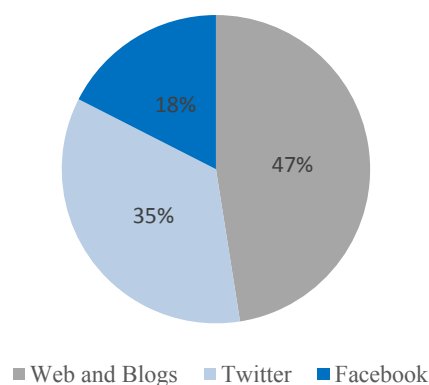


Figure 2. 7: Most used data sources for Arabic SA

Data Sources	Web and Blogs	Twitter	Facebook
Number of Published Papers	19	14	7

Table 2. 2: Data sources for published papers on Arabic SA

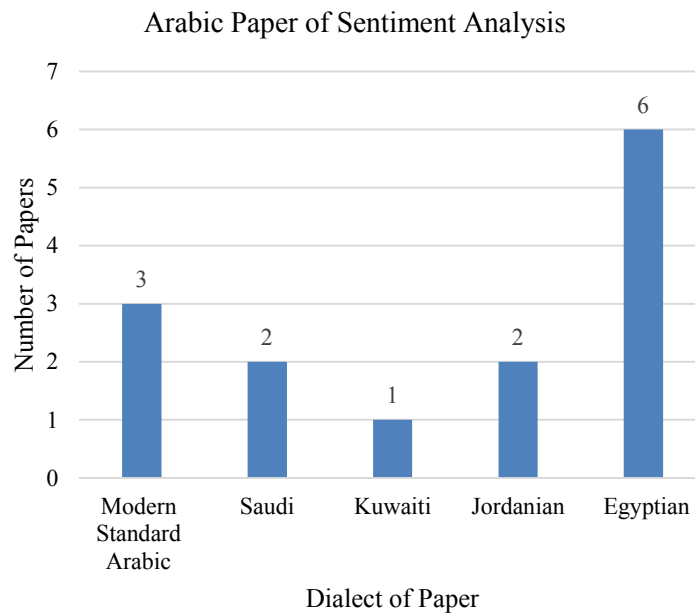


Figure 2. 8: Most used Arabic dialect in SA

## 2.6.2 Related Work on Twitter Sentiment Analysis

The work that has been done regarding Arabic SA in Twitter is limited. Papers could be categorized based on the used classification approach. Each effort that was done in researches [8], [17], [18], [29] used a supervised approach.

Where paper [8] has used a supervised approach. They examined two algorithms: SVM, and NB. Their dataset where trained in two ways. First, with the frequency of the unigrams and second combining unigrams and bigrams. They have used a tool that uses Twitter APIs to collect tweets.

While Paper [30] proposed a system for sentiment analysis using a SVM with presence vector. Their experiment shows that SA accuracy was not affected by adding Part Of Speech (POS).

Paper [18] that used five algorithms of supervised approach. Which are NB, SVM, Maximum Entropy, Bayes Net, J48 decision tree. With the help of Twitter filter stream API they were able to collect Arabic tweets.

Researchers in [29] have built *Kuwaiti Dialect Opinion Extraction System from Twitter* (KDOEST). They used a supervised approach using SVM and decision Tree algorithms to classify the tweets. They extracted features for the machine by dividing the Kuwaiti terms into classes such as happiness class. These classes are categorized under positive or negative.

Another supervised approach was used in paper [31] where they used RapidMiner to classify their collected tweets using both NB and D-Tree algorithms. They studied the impact of considering the emotion faces (emoji) that are used widely by Twitter's users. Their approach showed that classification with emotion faces has raised the accuracy from 58.28% to 63.79%.

Another work has been done for the Jordanian dialect was by paper [32]. They used RapidMiner for preprocessing and filtering stages. To annotate their tweet they hired the

CrowdSource website that displays the tweets on users and they do the annotation process. For the classification process they also used RapidMiner to examine three algorithms which are SVM, NB, and KNN.

While in paper [17] they examined both approaches the supervised and the unsupervised. In the unsupervised, they built their lexicon manually using SentiStrength website. And enhance it through adding the synonyms of the word. The supervised examination has been done using RapidMiner software. Their dataset consist of both MSA and Jordanian dialect. They have been collected using tweets crawler.

Authors of [33] have used a lexicon of 200 words. That size has affected their results. They found that the stemmer is useful to reduce the size of the lexicon since multiple words will have the same root at the lexicon. They used a small dataset of 100 tweets. Their accuracy was 73%.

Finally in this literature the effort of paper [20]. They used unsupervised approach, and built their lexicon from a seed lexicon of 380 words and extended it. After that, they used two algorithms to annotate these words. Finally, they hired two methods to calculate the sentiment of the tweets collected by Twitter search API. The first one is the sum which sum the polarity of each word in the text and the second one is the double polarity method where each word in the tweet has positive and negative polarity.

Comparison between all the scanned related works that used sentiment analysis to analyze the Arabic content on twitter shown in Table 2.3.

Reference	Collecting	#Tweets	Preprocessing & Filtering	Classifying
[8]	Tool to get tweets using Twitter's APIs.	- 1000 Tweets. - 500 Positive. - 500 Negative.	Removing: - Usernames. - Images. - Hash tags. - URLs. - Stop words.	Supervised approach - Used algorithms: SVM and NB. - Trained using the frequency of the unigrams. - Trained using a combination unigrams and bigrams.
[17]	Tweet crawler.	- 2000 Tweets. - 1000 Positive. - 1000 Negative.	- MS Word for misspelling. - Naive algorithm for repeated letters. - Normalization. - The Khoja stemmer tool for stop words.	Unsupervised approach - Building lexicon manually from SentiStrength website. - Add the synonyms of each word to enhance the lexicon. - Use unigram technique for feature extraction. Supervised approach - Used algorithms: SVM, NB, D-Tree, and KNN - Using RapidMiner software.
[30]	TAGREED	- 3015 Tweets.	- Tokenize the text automatically. - POS.	Supervised approach - Used algorithms: SVM.
[18]	Twitter filter stream API.	- 2861 Tweets. - 612 Positive. - 513 Negative. - 848 Neutral.	- Tag adding. - Normalization.	Supervised approach - Used algorithms: SVM, NB, MaxEnt, Bayes Net, and J48.
[32]	Twitter API	- 1000 Tweets.	- Stemming - Tokenizing. - Filtering - Stop words - Using RapidMiner	Supervised approach - Used algorithms: NB, KNN, SVM.
[20]	Twitter Search API	- 500 Tweets. - 155 Positive. - 310 Negative. - 35 Neutral.		Unsupervised Approach - Building lexicon using a seed 380 words manually. - The lexicon words tagged with their POS. - Set the polar sentiment for each word on the lexicon. - Used algorithm: Sum method, double polarity method
[29]	Twitter API	- 340,000 Tweets.	- Tokenization using Stanford Arabic tokenize.	Supervised approach - Used algorithm: SVM, D-Tree.
[31]	Twitter API	- 3000 Tweets.	- Normalization - Convert the emotion to synonym sentiment word. - Removed: Username, URL, hashtags, and stop words.	Supervised approach. - Used algorithm: NB, KNN. - Using RapidMiner software.
[33]	NODEXL tool (Microsoft)	- 100 Tweets. - 40 Positive. - 69 Negative.	- Tokenization. - Remove stop words. - Stemming using Khoja stemmer.	Unsupervised approach - Lexicon of 200 words.

**Table 2. 3: Comparison between techniques that were used for Arabic SA in terms of SA stages.**

### 2.6.3 Related Applications

This section shows the scanned similar applications that implement SA, which are:

- **Sentiment140:** It is a third party application that uses Twitter for analyzing tweets written in English and Spanish Languages.
- **Crowd Analyzer:** It is a paid application for Arabic sentiment analysis on social networks. It serves the companies or any organization to help them understand their clients.

## 2.7 Conclusion

This chapter has introduced background information on the sentiment analysis and the related fields. Starting by explaining the term data mining and an overview of its functions, applications, and other topics. Then an introduction on text mining and what its relation with sentiment analysis and NLP. After that, a detailed description on analyzing sentiments was introduced. Last in this chapter was presenting related work.

## **CHAPTER THREE: ANALYSIS AND METHODOLOGY**



## 3.1 Introduction

Analysis and methodology chapter will give a deep view on the requirements of *Ra'ai analyzer* (RA). The word *Ra'ai* represents the Arabic word (رأي) which means opinion. This chapter will describe the user and system requirements and specify which one will be applied by the supervised classifier and which will be applied by the unsupervised classifier. The chapter divided as follow: Section 3.2 explains user requirements where both functional and non-functional requirements are discussed. Section 3.3 for RA system requirements from a functional and non-functional view and the characteristic of analyzer. Section 3.4 spotlights on the methodology used for RA. Section 3.5 displays the project's timeline by time and the interval that took each stage of the RA. Section 3.6 explains functional modeling that shows the use case diagram along with use cases' tables and the activity diagram of the system. Finally, section 3.7 concludes and summarizes what will be discussed in the chapter.

## 3.2 User Requirements

User requirements are all the requirements that the user expects the system to provide. These requirements were stated by project's team.

### 3.2.1 Functional Requirements

Functional requirements are the services that the system will provide. Following are the requirements of the RA.

- User shall access the system online for unsupervised classifier.
- User shall enter a keyword for unsupervised classifier.
- System shall calculate the overall sentiment of the collected tweets.
- User shall get the overall sentiment toward collected tweets for unsupervised classifier.
- System should view the information about the collected tweets.
- User should view the policy of the system for unsupervised classifier.
- User should view the manual page for unsupervised classifier.

### 3.2.2 Non Functional Requirements

- System shall response to user click within 10 seconds.

## 3.3 System Requirements

System requirements are the requirements that needed by the system in order to be able to achieve its goals. The requirements are divided onto two categories: functional and non-functional.

### 3.3.1 Functional Requirements

The functional requirements of the system describe the services that system provides in details. RA function requirements are:

- User shall have an internet connection for unsupervised classifier.
- System shall view the RA web page to the user when he/she hits the system URL for unsupervised classifier.

- System shall provide a text field for the user to enter the keyword for unsupervised classifier.
- System shall validate the keyword entered by the user for unsupervised classifier.
- System shall provide an error message if the user entered invalid keyword for unsupervised classifier.
- System shall provide an error message if the user did not enter keyword for unsupervised classifier.
- System shall retrieve the stored tweets that relate to the keyword entered by the user.
- System shall extract the text out of the tweets to prepare it for the preprocess stage.
- System shall preprocess the extracted text to prepare it for the filter stage.
- System shall filter the extracted text to prepare it for the classification stage.
- System shall classify each tweet to positive and negative for unsupervised classifier.
- System shall provide the tab (analyze sentiments) "تحليل الآراء" to be pressed by the user to command the system to view the overall sentiment as a figure for unsupervised classifier.
- System shall provide the tab (words cloud) "سحابة الكلمات" to be pressed by the user to order the system to view the sentimental words as a cloud of words where the size of a word represents its occurrence frequency relative to other words in the cloud.
- System shall count number of the collected tweets to be viewed to the user.
- System shall provide the tab (show tweets) "عرض التغريدات" to be pressed by the user to command the system to view the information about the collected for unsupervised classifier.
- System shall provide a policy tab "سياستنا" in the RA web page that contains the RA's policy in the unsupervised classifier.
- System shall provide a user manual tab "دليل المستخدم" in the RA web page that contains RA manual for unsupervised classifier.

### 3.3.2 Non Functional Requirements

Non-functional requirements place constraints on *how* the system will do its functional requirements. The RA non-functional requirements are:

1. Usability
  - Simple graphical user interface to simplify the interaction between user and RA.
  - Use Arabic interface since the target users are Arabian.
2. Availability
  - The system is available free.
  - The system is available in public domain online.
  - The system works 7 days a week and 24 hours a day.

### 3.3.3 System Characteristic

- This system will be built for Arabic tweets.
- This system will classify tweets in domain (football) in Arabic.
- System will provide two approaches to classify the tweets. First approach is supervised and second approach is unsupervised.

### 3.4 Methodology

The process model that was used in the project is the modified waterfall model. The model consists of six stages. First stage is the requirement analysis where the requirements of the project are collected. Design stage is next and focuses on designing and specifying the process of building the project. Third stage is the implementation stage where the project requirements are implemented. Forth stage is the testing where functions of the system are tested. Final stage is the operation and maintenance for the system. The following figure represents the used modified waterfall model Figure 3.1 [16].

The modified waterfall model was used in the system because of the flexibility provided by the model. The system needs to be reviewed and changed after each stage. This leads the team to return to earlier stages for modifications which is allowable by the modified water model.

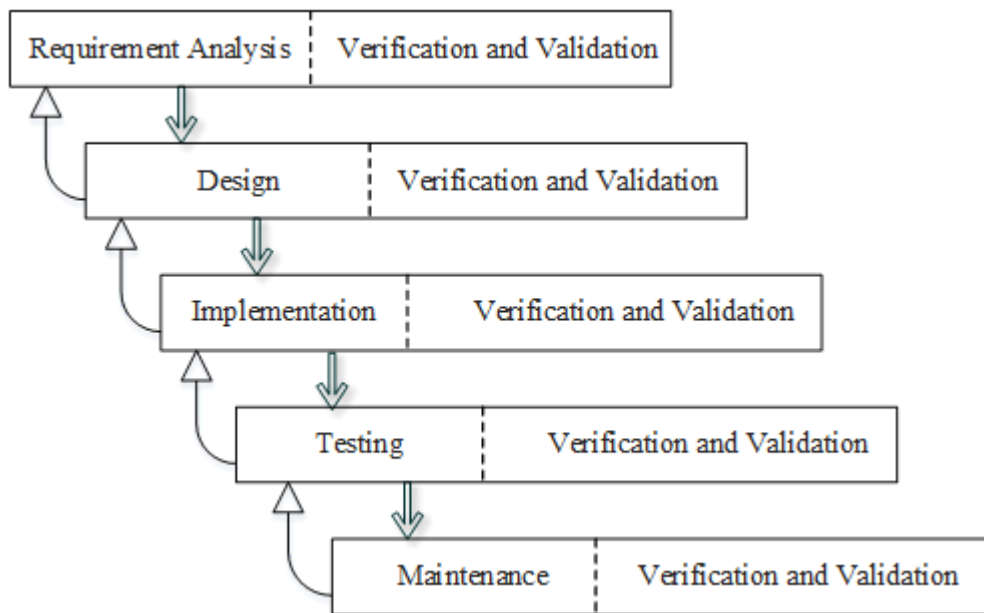


Figure 3.1 : Modified water fall model.

### 3.5 Timeline

The project passes through sequence of stages. Some stages may pass in parallel with each other. The figure below show the timeline of the project Figure 3. 2.

Task	5 Mar	1 Apr	1 May	1 Jun	13 Jul	12 Aug	1 Sep	1 Oct	1 Nov	15 Dec
Literature Review										
System Requirements and Analysis										
System Design										
System Implementation										
System Testing										

Figure 3. 2: Project timeline.

### 3.5.1 Description of Project Stages

#### Stage 1: Literature Review

This stage is about reading and analyzing related work in the project fields which took 7 months.

#### Stage 2: Application Requirements and Analysis

This stage is about determining the system services and what services should be provided by the system it took three months.

#### Stage 3: Application Design

At the design stage, the members will design the system architecture, the system interface, and designing the system's logo. It took the team two months for this task.

#### Stage 4: Implementation

At this stage, the members will be implemented the system according to elected requirements.

#### Stage 5: Testing

In the testing stage the system will be tested to check wither the promised services are done as expected or not.

#### 3.5.1.1 Roles and Responsibilities

The following Table 3.1 describes the role that assigned for each member and the responsibilities related to each role.

Role	Member	Responsibilities
Team Leader	Wejdan Alohaideb	<ol style="list-style-type: none"> <li>1. Plan the meetings.</li> <li>2. Ensure every member knows her task and working on it.</li> <li>3. Divide the tasks between the members.</li> </ol>
Technical Writer	Wejdan Alohaideb	<ol style="list-style-type: none"> <li>1. Correct the Grammar.</li> <li>2. Collect all parts of the documents.</li> </ol>
Literature Reviewing	All team members	<ol style="list-style-type: none"> <li>1. Review literature for all the project parts.</li> </ol>
Communication with Supervisor	All team members by email, google sites, and DropBox	<ol style="list-style-type: none"> <li>1. Conform meetings.</li> <li>2. Discuss some issues.</li> </ol>
Document Formatting	Tarfa Albuhairei	<ol style="list-style-type: none"> <li>1. Check the format of the document.</li> <li>2. Correct the format.</li> </ol>
Drawing Diagrams	Mawaheb Altuwijri	<ol style="list-style-type: none"> <li>1. Drawing all project diagrams.</li> </ol>
Analysis	All team members.	<ol style="list-style-type: none"> <li>1. Write overall description and design system interface.</li> <li>2. Specify functional and non-functional requirements.</li> <li>3. Writing the use cases.</li> </ol>
Design	All team members.	<ol style="list-style-type: none"> <li>1. Design the interfaces.</li> </ol>
Implementation	All team members.	<ol style="list-style-type: none"> <li>1. Implementing all system components.</li> <li>2. Integrating all the components.</li> </ol>
Testing	All team members	<ol style="list-style-type: none"> <li>1. Ensuring that the system functions are working.</li> </ol>

Table 3. 1: Team's roles and responsibilities.

## 3.6 Functional modeling

This section demonstrates all the functions provided by the RA and the interaction between the actors of the RA.

### 3.6.1 Use Cases

The diagram in Figure 3.3 shows all the use cases and assigns them to the appropriate user [16].

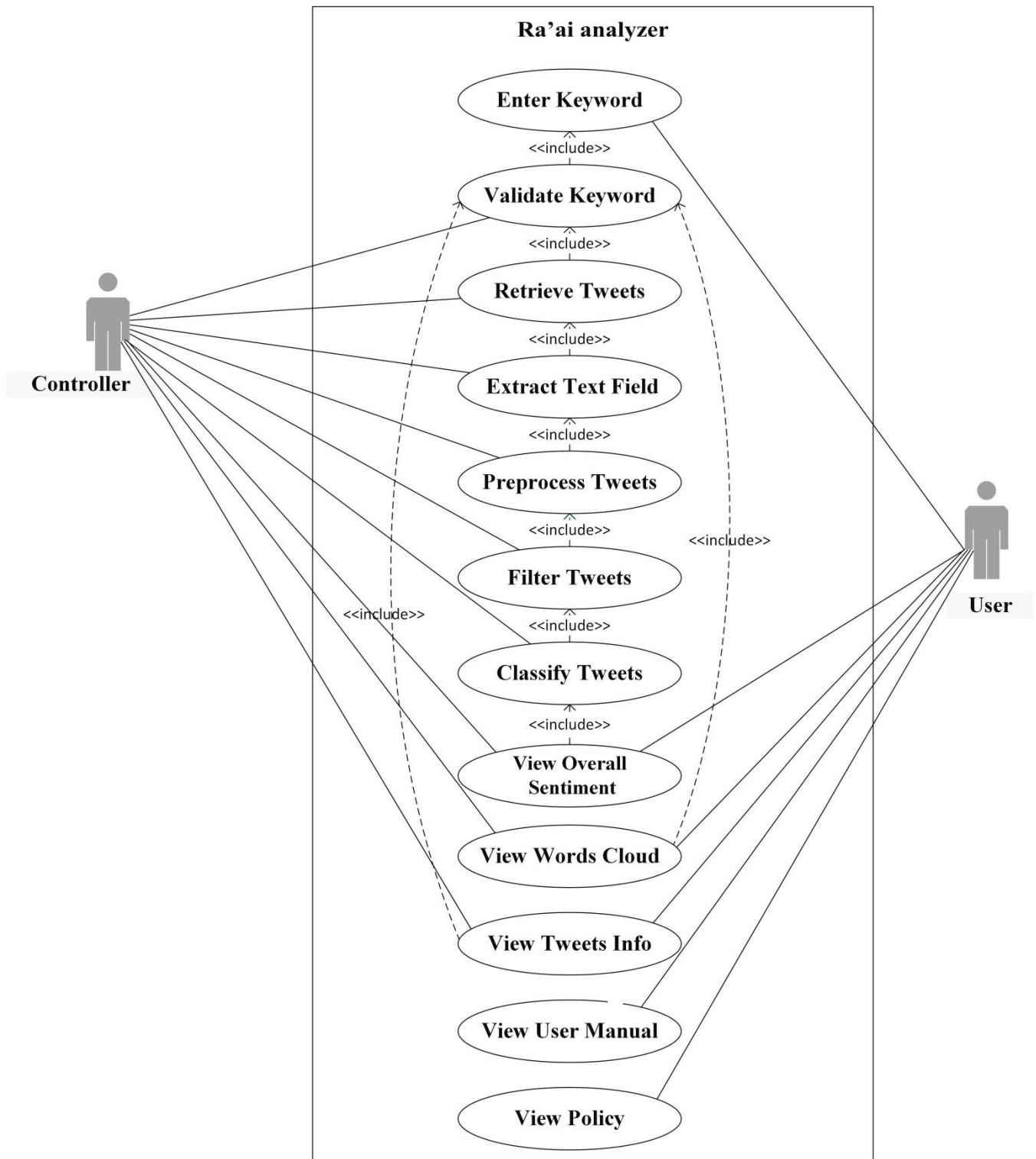


Figure 3.3: Use cases diagram.

**Actors:**

There are two main actors:

- 1- User of the RA.
- 2- Controller.

**Use cases:**

1. **Enter Keyword:** This function gets the keyword from the user.
2. **Validate Keyword:** This function verifies the validity of the keyword entered by the user.
3. **Retrieve Tweets:** This function retrieve the collected tweets from the dataset.
4. **Extract Text Field:** This function extracts the text field out of the collected tweets.
5. **Preprocess Tweets:** This function removes any unrelated content form the text such as images, URLs, hashtags, and all non-Arabic letters.
6. **Filter Tweets:** This function filters the tweets from stop words, and normalizes the word.
7. **Classify Tweets:** This function classifies each tweets into positive or negative.
8. **View Overall Sentiment:** This function displays the overall sentiment as bar plot.
9. **View Words Cloud:** This unction display a word cloud for the most frequent words in the retrieved tweets.
10. **View Tweets Info:** This function displays information about the collected tweets.
11. **View User Manual:** The function displays the manual of the system.
12. **View Policy:** The function displays the policy of the system.

Use case:	Enter Keyword
Goal:	Entering keyword by user to get the overall sentiment about it.
Actors:	User.
Precondition:	None.
Triggers:	User clicks on text field.
Main success scenario:	1. User enters an Arabic keyword in the text field.
Alternative Scenario:	None.
Post condition:	System passes the keyword to the server.

Table 3. 2: Enter keyword use case.

Use case:	Validate keyword
Goal:	Checking the validity of the entered keyword.
Actors:	Controller.
Precondition:	User entered the keyword.
Triggers:	User clicks on start button “ابدأ”.
Main success scenario:	1. System validates that the entered keyword is written in Arabic.
Alternative Scenario:	a. The keyword was not entered. 1.1.1. System displays message “قم بإدخال كلمة من فضلك.” 1.1.2. System backs to Enter Keyword use case. 1.2. The keyword is invalid. 1.2.1. System displays message “قم بإدخال كلمة مكونة من حروف عربية فقط.” 1.2.2. System backs to Enter Keyword use case.
Post condition:	None.

Table 3. 3: Validate keyword use case.

Use case:	Retrieve Tweets
Goal:	Retrieving the collected tweets from the dataset.
Actors:	Controller.
Precondition:	User entered valid keyword.
Triggers:	None.
Main success scenario:	1. Controller retrieves the collected tweets from the dataset.
Alternative Scenario:	None.
Post condition:	None.

**Table 3. 4: Retrieve tweets use case.**

Use case:	Extract Text Field
Goal:	Extracting the text field out of the tweet fields.
Actors:	Controller.
Precondition:	Controller retrieved the tweets from the dataset.
Triggers:	None.
Main success scenario:	1. Controller extracts text field by removing unrelated fields such as the user account names, location and date.
Alternative Scenario:	None.
Post condition:	System updates the dataset with the extracted text field.

**Table 3. 5: Extract text field use case.**

Use case:	Preprocess Tweets
Goal:	Cleaning collected tweets.
Actors:	Controller.
Precondition:	Controller retrieved tweets from the dataset.
Triggers:	None.
Main success scenario:	1. Controller removes from tweet unrelated content such as images, hash tags, URLs, and all non-Arabic words or letters.
Alternative Scenario:	None.
Post condition:	System updates the dataset with the text after preprocessing tweets.

**Table 3. 6: Preprocess tweets use case.**

Use case:	Filter Tweets
Goal:	Filtering the tweet from stop words, and normalize the words.
Actors:	Controller.
Precondition:	Controller preprocessed the tweets.
Triggers:	None.
Main success scenario:	1. Controller retrieves the tweets from the dataset. 2. Controller deletes stop words if any. 3. Controller normalizes the word.
Alternative Scenario:	None.
Post condition:	System updates the dataset with the text after filtering tweets.

**Table 3. 7: Filter tweets use case.**

Use case:	Classify Tweets
Goal:	Classifying each tweet into positive and negative. Then produce the overall sentiment.
Actors:	Controller.
Precondition:	Controller filtered the tweets.
Triggers:	None.
Main success scenario:	<ol style="list-style-type: none"> <li>1. Controller retrieves the tweet from the dataset.</li> <li>2. Controller classifies each word to positive or negative.</li> <li>3. Controller classifies the whole tweet.</li> <li>4. Controller produces the overall sentiment of the tweets.</li> </ol>
Alternative Scenario:	None.
Post condition:	System saves the overall sentiment on buffer memory.

**Table 3. 8: Classify tweets use case.**

Use case:	View Overall Sentiment
Goal:	Displaying the overall sentiment as a bar plot.
Actors:	User.
Precondition:	Controller Classified all tweets to positive or negative.
Triggers:	None.
Main success scenario:	1. Controller displays the result that is generated after classification.
Alternative Scenario:	None.
Post condition:	None.

**Table 3. 9: View overall sentiment use case.**

Use case:	View the Words' Cloud
Goal:	Displaying the words' frequency in the tweets as a cloud of where the size of each word is relative to its frequency.
Actors:	User.
Precondition:	Controller Classified all tweets to positive or negative.
Triggers:	User press on the (word cloud) "سحابة الكلمات" tab.
Main success scenario:	<ol style="list-style-type: none"> <li>1. Controller calculates the frequency of each word in the collected tweets.</li> <li>2. Controller omits the words least frequent.</li> <li>3. Controller displays the words' cloud.</li> </ol>
Alternative Scenario:	None.
Post condition:	None.

**Table 3. 10: View the words cloud use case.**

Use case:	View Tweets' Info
Goal:	Display information about the collected tweets.
Actors:	User.
Precondition:	Controller collected tweets.
Triggers:	User press on (show tweets) "عرض التغريدات" tab.
Main success scenario:	<ol style="list-style-type: none"> <li>1. System displays the collected tweets.</li> <li>2. System displays the polarity that the classifier gave for each tweet</li> </ol>
Alternative	None.



Scenario:	
Post condition:	None.

Table 3. 11: View tweets info use case.

Use case:	View User Manual
Goal:	View the help page of the system.
Actors:	User.
Precondition:	None.
Triggers:	Press “دليل المستخدم” tab.
Main success scenario:	1. System displays the manual of the system.
Alternative Scenario:	None.
Post condition:	None.

Table 3. 12: View user manual use case.

Use case:	View Policy
Goal:	View the policy of the system.
Actors:	User.
Precondition:	None
Triggers:	User press the “سياستنا” tab.
Main success scenario:	1. System shows the policy of this system.
Alternative Scenario:	None
Post condition:	None

Table 3. 13: View policy use case.

### 3.6.2 Activity diagram

Activity diagram is a graphical representation of workflows of the system activity and actions. Activity diagram shows the sequence of the system, from the start point until the end point. Figure 3. 4 represents the activity diagram [16].

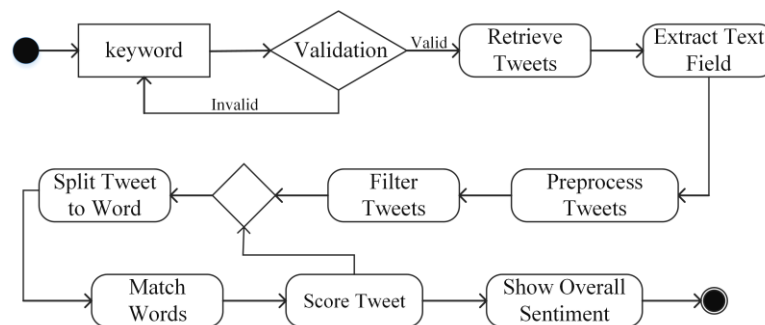


Figure 3. 4: Activity diagram

### 3.7 Conclusion

The chapter has described the analysis of the system from different aspects. First, both the user and system requirements were listed. Then the methodology of the system and the system timeline were presented. Finally, the functional modeling was discussed in the last section.

## **CHAPTER FOUR: DESIGN AND IMPLEMENTATION**

## 4.1 Introduction

This chapter discusses the design and implementation phases of the project. It starts by explaining how the system was designed on different levels which are the architectural design and the interface design that will be covered in section 4.2. After designing the system, the implementation phase started. Section 4.3 will highlight all implementation aspects from the used programming language until the used tools to bring an idea on how the system was implemented.

## 4.2 System Design

This system applies two sentiment analysis approaches which are supervised and unsupervised. The supervised and unsupervised approaches have been trained on one domain dataset. In general, the system contains four main components: collecting tweets, preprocessing, filtering and classifying as shown in Figure 4. 1 [16]. Each component will be explained in section 4.2.2 later in this chapter.

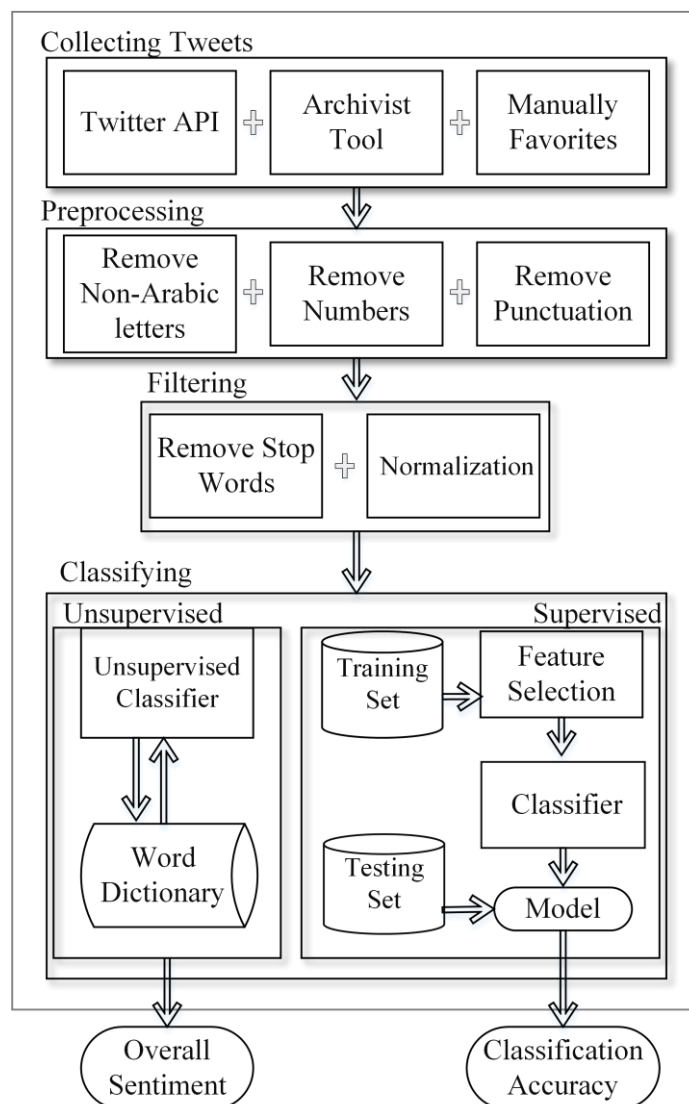


Figure 4. 1: System architecture.

### 4.2.1 Modular Decomposition

The system uses both functional and object oriented design models. For the supervised approach, the system uses an object oriented design model since it uses WEKA package which are set of classes implemented in Java. While the unsupervised

approach uses a functional oriented design model. Since the unsupervised components consists of functions where each function takes an input and produce an output.

## 4.2.2 Architectural Design

As explained above the system contains four main components. Collecting tweets, preprocessing, filtering and classifying. Each component takes an input and produces an output to the component that comes below it. Followed sections will explain each component.

### 3.2.2.1 Collecting Tweets Component:

Tweets were collected on a specific topic using three approaches, which are Twitter Application Program Interface (API) by using search function, Twitter Archivist and manually by favoring tweets then retrieving them using TwitteR package. The number of collected tweets was more than 30,000 that contains advertisements, retweets, and unrelated tweets. After removing the advertisements, retweets, and unrelated tweets manually the remaining tweets was 1000. The number of positive tweets is 600, and negative is 400.

TwitteR package is an R interface to Twitter API that deals with Twitter to retrieve tweets and do other functions. The project team found that the package works only if the entered keyword was in the most trended hash tags and with more than almost 300,000 tweets. For that reason, the football domain was selected as it was the most Saudi trending topic in the region at the time of this study, coinciding with the Asian and Gulf cups.

### 3.2.2.2 Preprocessing Component:

Preprocessing component is responsible for removing all unrelated content such as URLs, user names, non-Arabic letters, numbers, and punctuations.

### 3.2.2.3 Filtering Component:

The third component has filtered the tweets by doing three processes. First is removing all stop words from the collected tweets. The available stop words list [34] was not suitable for SA because of two reasons. The first reason is that it contained negation words such as (لا، لم، لن) that may change the sentiment of the tweets. The second reason is that the used tweets are written mostly in Saudi dialect which has different stop words than the available one. For SA purposes the stop words list was modified to combine both the Modern Standard Arabic (MSA) stop words along with Saudi dialect stop word. In addition, negation words were removed from the stop word list. The modified stop words list is displayed in appendix A. The second process is normalizing each of (أ، إ، آ) letters to (ا), (ة) to (هـ), and remove any diacritics (short vowels) such as (أَ أُ إَ إِ آَ آِ). This normalizes all words to hold the same letters shape as the one in the dictionary for the final step which is classification. The reason for applying this process is that many Arabic Twitter users often mistake between these similar letters and use them interchangeably. Third process is correcting misspellings and removing the repeated letters manually such as (gooooood = زبيبيين) to (good = زين).

### 3.2.2.4 Classifying Component:

This component includes two subcomponents supervised approach and unsupervised approach.

#### 3.2.2.4.1 Supervised Approach

The supervised approach used a training dataset which contains 1000 tweets that was treated as unigram and bigram. Unigram treats each word as an independent unit, while bigram treats a pair of consecutive words as one unit. The frequency of the presence of a word affects the classifier's accuracy. For that reason, the feature selection was applied to remove the word (هلال) from all the tweets; since it has the highest frequency and it is not sentimental word.

The dataset is transformed into feature vector which transform it to numeric to be understandable by the classifier.

The feature extraction is used as *Term Frequency – Inverse Document Frequency* (TF-IDF) transformer. Then the feature vector used different machine learning algorithms; SVM, NB, and KNN. KNN and IBK are used interchangeably to refer to the KNN algorithm by the papers' authors and the term IBK will be used in chapter 5 instead of KNN. Each algorithm extracts a model (classifier) which will be used in the evaluation step. The evaluation step took the training dataset as testing dataset and uses the produced model to provide the accuracy percentage of each classifier.

Java language with WEKA packages [35] was used to implement the machine learning algorithms.

#### 3.2.2.4.2 Unsupervised Approach

The unsupervised approach has two words dictionaries. Positive words dictionary and negative words dictionary. The dictionaries contain both MSA words and Saudi dialect words. The MSA words dictionaries are built beforehand [36]. While the Saudi dialect words were built with the help of a part of the collected dataset about 840 tweets. Each word was labeled as positive or negative based on two human experts. If the experts did not agree on the label of a certain word, a third expert is consulted to break the tie. The negation words that were removed from the Stop words list were added to the negative words dictionary. Finally, positive words dictionary was expanded from 1294 to 1451. While negative words dictionary was expanded from 2245 to 2460. A sample of the expanded words dictionaries available at appendix B.

Each tweet was split into a list of words. A match between the tweet's words and the dictionaries' words is done to score each tweet according to the number of words that were found in both dictionaries.

To increase the accuracy, two dictionaries were added to the classifier which contain the positive and negative phrases such as ( كما ) which indicate negative sentiment, and ( ما شاء الله ) which indicate positive sentiment. The following algorithm explains the process. It has a time complexity of  $O(N^2)$ . R language was used to implement the lexicon based algorithm.

**INPUT:** Tweets  $T$ , Positive words Lexicon  $PL$ , Negative words Lexicon  $NL$ , Positive Phrases lexicon  $PP$ , Negative Phrases lexicon  $NP$ , Words of the tweets  $W$ .

**OUTPUT:**  $P = \{Pos, Neg, \text{ or } Net\}$ , where **Pos:** Positive, **Neg:** Negative, **Net:** Neutral.

**INITIALIZATION:** Score = 0,  $P = 0$ ,  $N = 0$ , where  $P$ : accumulate the positive words,  $N$ : accumulate the negative words, Score: subtract  $N$  from  $P$  to get the tweet score.

**Begin**

1. For each  $T_i \in T$ 
  1. For each  $PP_i \in PP$ 
    1. If  $PP_i \in T_i$  then
      1.  $P = P + 1$
    2. End if
    3. If  $NP_i \in T_i$  then
      1.  $N = N + 1$
    4. End if
  2. End for
  3. For each  $W_i \in T_i$ 
    1. If  $W_i \in PL$  then
      1.  $P = P + 1$
    2. End if
    3. If  $W_i \in NL$ 
      1.  $N = N + 1$
    4. End if
  4. End for
  5. Score =  $P - N$
  6. If Score > 0 then
    1.  $P \leftarrow Pos$
  7. End If
  8. If Score < 0 then
    1.  $P \leftarrow Neg$
  9. End If
  10. If Score == 0 then
    1.  $P \leftarrow Net$
  11. End If
2. End for

**End**

### 4.2.3 Logo Design

The logo represents the RA system and it gets the user's first impression. For that reason the team members have designed it carefully to make it represent able. Since the system related to Twitter and with people's sentiment form these relations the logo was made. It is a simple Twitter bird with chat bubble contains plus and minus signs was designed using Photoshop software. Figure 4. 2 displays the designed logo.



Figure 4. 2: RA logo.

### 4.2.4 Interface Design

As the interface make a difference on the ability of the user to use the system properly. The team members focus on how to increase successful interaction between the

system and the user. In order to achieve these goals the team has scanned the similar systems and tried to find the best features to add to the system.

## 4.3 Implementation

The implementation section discusses the used programming languages in the RA and the reasons behind this decision. Then the used tool will be listed and the purpose of using it. Finally, a list of the used packages along with a description of how they work was listed.

### 4.3.1 Programming Languages

Two programming languages used for this project the Java language along with the R language. Java was used to implement the supervised approach while R was used to implement the unsupervised approach. The reasons of choosing these two languages are:

- 1- R language is a language built for data analyzing. R used because of these reasons:-
  - It was designed in a way that makes the process of analyzing data more efficient.
  - The plentifully and the variation of the R packages.
  - The ability to connect R code with other languages code.
  - R is an open source language making the ability to modify on some of its functions a legal process.
- 2- Java language was a replacement for R language to implement the supervised approach. R's machine learning packages do not support Arabic language and it would took time for the team to make R able to understand Arabic. Consequently, the best and fast solution was Java.

### 4.3.2 Tools

These tools were essential to complete the implementation process. Starting from the programming languages to the hardware that is needed to implement RA. Here is a list of the used tools:

- **Eclipse:** It is a framework programming languages development such ad C++, C, PHP and Java. This tool was used to build the supervised approach in Java language.
- **RStudio:** It is a framework for R language. This tool was used to build the application interface and unsupervised approach.
- **Twitter App:** It is a service provided by Twitter for developers that enables connecting their application with Twitter. This will make the developer able to get tweets from Twitter, post tweet, and other services. Twitter App was used to collect tweets for RA.
- **Repository CRAN and CRANextra:** It is the Comprehensive R Archive Network where all the packages are published by R developers. This site was used to install the needed packages for the RA.
- **GitHub:** It is a web based platform for source codes. It was used to download some R packages that are not available at the CRAN repository.
- **Shiny Apps.io:** Is a service by RStudio that facilitates the deployment for R web applications.
- **WEKA Packages:** A WEKA package provides the supervised approach classes.

- **WEKA GUI:** Which is a graphical user interface provided by WEKA.

### 4.3.4 Package and Classes Description

Packages are a set of functions that are built in R it is a synonym for libraries that is used for other languages. This section lists the used packages and classes for implementing the supervised and unsupervised approach. Table 4.1 lists the set of the used R packages to implement the unsupervised classifier. While Table 4.2 shows the used Java classes to implement the supervised classifier which are provided by WEKA.

Package	Description
Twitter	It is an interface written in R for the Twitter web API. Used for collecting stage.
Stringr	It is package to handle the string manipulation in R. used in preprocess stage.
Tm	Package to handle the filtering phase of the system.
Plyr	A set of tools that solves a common set of problems: such as breaking a big problem down into manageable pieces.
Ggplot2	An implementation of the grammar of graphics in R.
Wordcloud	Creates an expressive word clouds.
Rcolorbrewer	The package provides palettes for drawing expressive maps.
Shiny	A package for implementing web based applications. It provides HTML, CSS, and javaScript functions that can be used to form the interface. Used to implement the system interface.
Shinyincubator	An extended package for shiny.

**Table 4. 1: List of the used packages.**

Classes	Description
NGramTokenizer	Used to split a tweet in reading to n items from a given tokenize by using setNGramMinSize() and setNGramMaxSize() functions.
StringToWordVector	Used to convert string attributes into a set of attributes representing word occurrence depending on the tokenize by using setTFTransform() and setIDFTransform() functions.
Classifier	Used to generate a classifier model of algorithm (SVM, NB, IBK) by using buildClassifier() function.
Evaluation	Used to evaluate a classifier via: <ul style="list-style-type: none"> <li>- CrossValidateModel() function to perform a cross-validation.</li> <li>- EvaluateModel() function to perform several test options (using training set as testing set, supplied test set which used new dataset, and percentage spilt which splits the dataset into training set and testing set)</li> <li>- ToSummaryString() function to output the accuracy.</li> <li>- ToClassDetailsString() function to output the details of accuracy for each class such as true/false positive rate, precision/recall/F-Measure.</li> <li>- ToMatrixString() function to output the confusion matrix. Confusion matrix shows the distribution of predicted polarity.</li> </ul>

**Table 4. 2: List of the used classes.**



### 4.3.5 Procedures Description

The following is a list of the implemented procedures in the unsupervised approach along with a description of their functions shown in table 3.4.

Procedure	Description
Collect_tweets()	Retrieve the tweets from twitter according to the entered keyword.
Validate()	Checks if the entered keyword is valid or not.
CleanTweets()	Remove URLs, user names, punctuation, Arabic punctuation, numbers and non-Arabic letter.
Normalize()	Normalize each of (أ, إ, آ) letters to (ا), (ة) to (ه).
ReadTweets()	Retrieve the collected tweets.
Wordcloudentity()	Built the words cloud for the most frequently words in the collected tweets.
Score.sentiment()	Classify the tweets into positive and negative and return a data frame with the polarity of each tweets
Sentiplot()	Plot the sentiment of the collected tweets as bar plot.

Table 4. 3: List of the unsupervised classifier procedures.

## 4.4 Conclusion

This chapter showed the RA system design and implementation. The system design was explained through viewing system architecture, interface design, and logo design. Then the implementation process was discussed through the used programming languages and the tools that were needed for the system. Finally, the used packages and classes and implemented function were listed along with the description for each one.

## **CHAPTER FIVE: TESTING**

## 5.1 Introduction

One of the important phases that any project must pass through is testing phase. This phase checks the validation of the RA functions and whether they work well or not. For the sake of finding, the system weaknesses and strengths points this chapter displays all the tests that have been done over the system.

The evaluation of accuracy for the supervised approach used the precision, recall and accuracy equations. While the unsupervised approach used only the accuracy equation. Their mathematical forms are in equations 1, 2 and 3:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (3)$$

Where TP, FP, TN, and FN are true positive, false positive, true negative and false negative, respectively.

The chapter is classified as follows: section 5.2 displays the results of the tests that were done on the supervised approach, section 5.3 displays the unsupervised approach testing, section 5.4 displays a comparison between the process of the supervised and unsupervised, finally section 5.5 concludes the chapter results.

## 5.2 Supervised Testing

This section reviews the tests that have been done over the supervised classifier. Three different experiments were conducted using three classifiers: SVM, NB and IBK. The motivation of selecting these classifiers was the superior performance they showed in previous related studies [37], [8], [17]. The test has been repeated ten times and each different dataset is used.

### 5.2.1 N-Gram Comparison Test

The goal of this test is to show the accuracy effect with the use of the UniGram and BiGram. For testing and validation purposes, the 10-fold cross validation technique was used; since cross validation is more suitable for small datasets. It starts with 100 tweets, 250, 400, 650 and 1000 tweets. Table 5.1 shows the average of the experiment's outcomes which clearly confirms SVM has better accuracy than other classifiers.

As appeared, the classifiers' accuracies are more than 90%. The reason of high results is that cross validation has evaluated the classifiers on the same training dataset containing tweets with common sentimental words.

The unigram has achieved higher accuracy than bigram. The reason behind these results is the presence of negation words in negative tweets. In bigram negation, words will appear differently for the classifier based on how the tweet was divided. Suppose the tweet (أحمد ما يحب الكرة) which means (Ahmed do not like football). Bigram technique breaks the tweet into units of two words such that the tweets' units are (أحمد ما) (يحب كرة). The sentimental words (ما) and (يحب) were separated, the unit (أحمد ما) will have no meaning while the unit (يحب الكرة) will have a positive meaning which will classify the tweet to positive rather than negative and therefore decrease the accuracy of the results. Figures 5.1, 5.2 and 5.3 show the test results on SVM, NB and IBK respectively. In addition, figures 5.4 and 5.5 show the significant change on the algorithms' accuracy when using UniGram and BiGram.

NGram	UniGram			BiGram		
	SVM	NB	IBK	SVM	NB	IBK
100 Tweets	91	92	86	73	73	71
250 Tweets	96	95	89	79	78	74
400 Tweets	96	96	90	83	77	78
650 Tweets	98	97	96	95	90	92
1000 Tweets	98	95	94	98	79	82

Table 5. 1: Classifier accuracy.

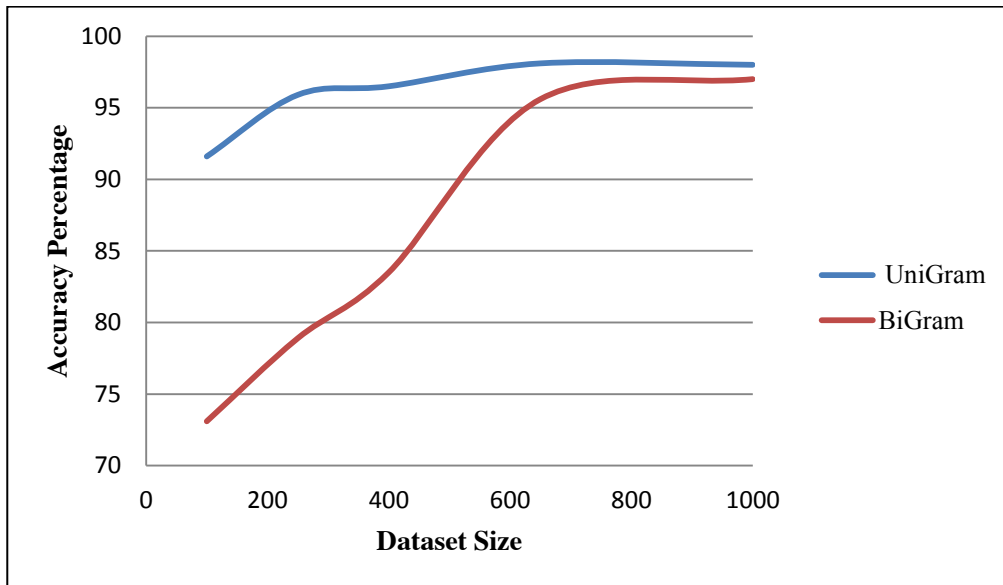


Figure 5. 1: The effect of use UniGram and BiGram on SVM

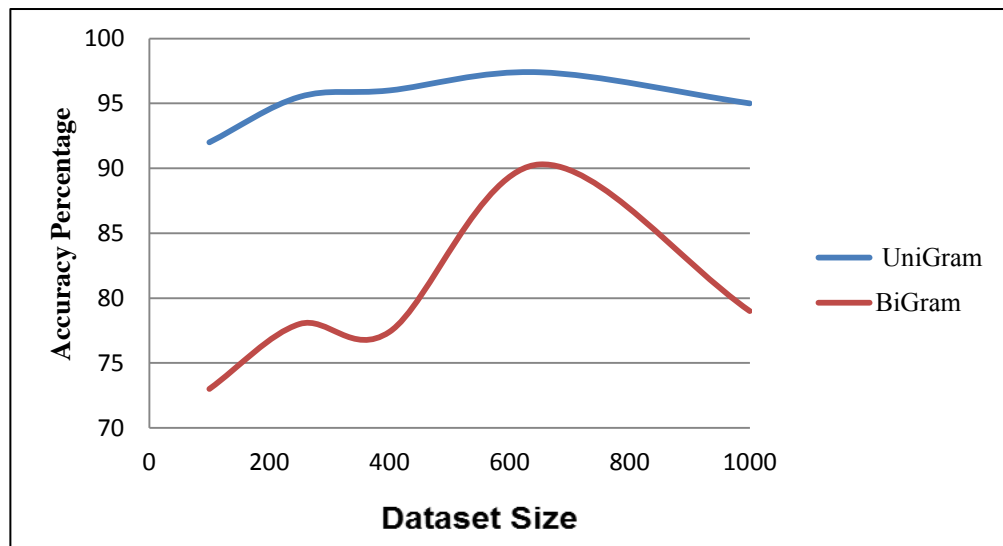


Figure 5. 2: The effect of use UniGram and BiGram on NB

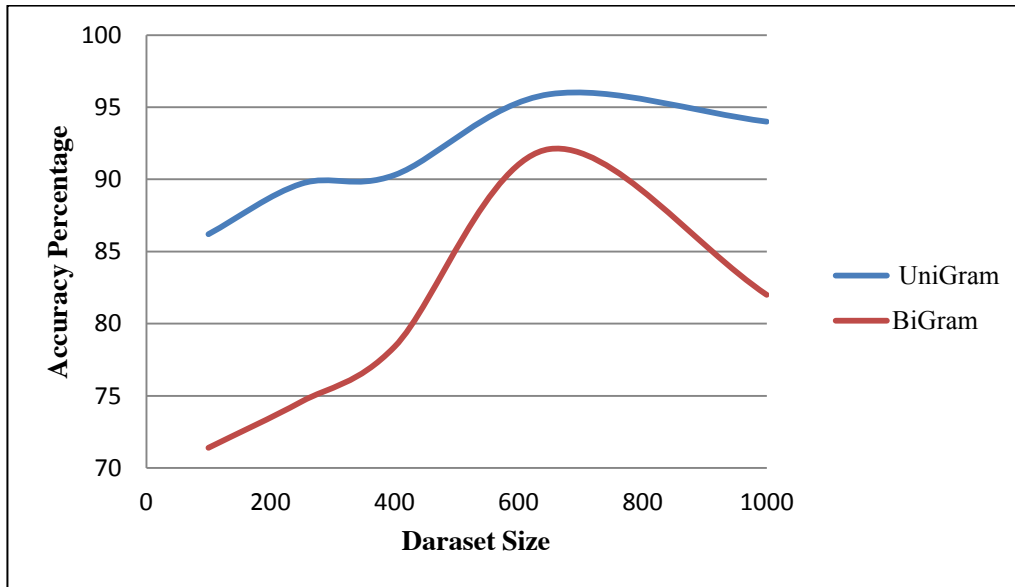


Figure 5.3: The effect of use UniGram and BiGram on IBK

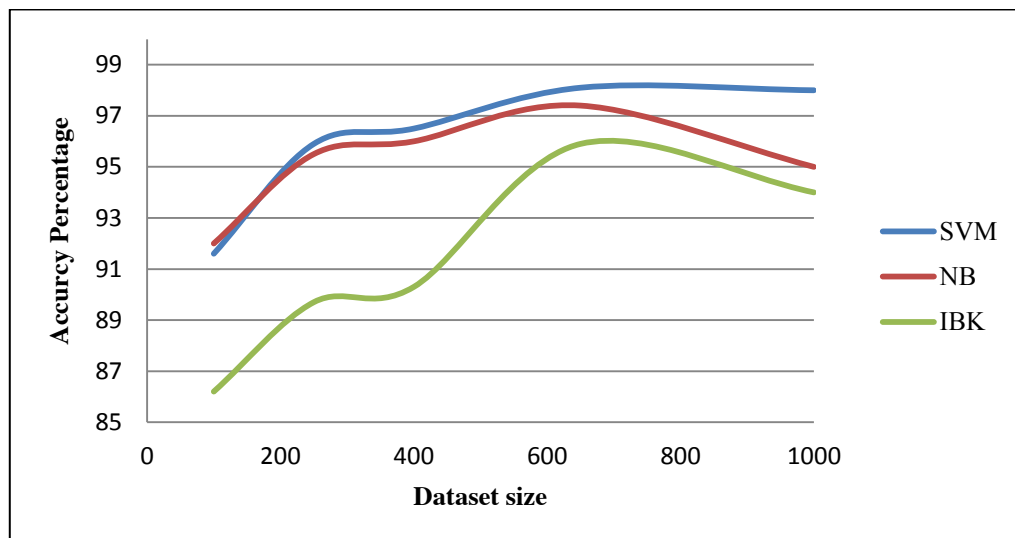


Figure 5.4: The effect of UniGram on classifiers accuracy

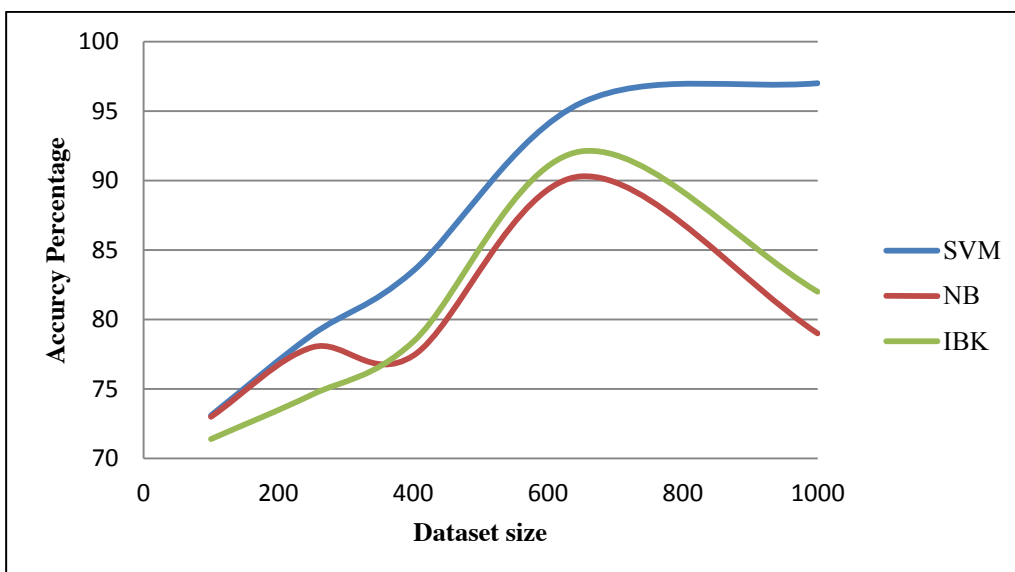


Figure 5.5: The effect of BiGram on classifiers accuracy.

The accuracy of the supervised approach's algorithms can be improved by increasing the size of the training dataset, and by using a words' stemmer.

## 5.2.2 Test Options Comparison

A comparison was done between the accuracy results based on different tests options. Test has run over three test options which are cross validation; use training set and percentage split. The test shows the effect of the selected test option over the size of data using UniGram. The test shows that the results of use training set have the best result with 100%. While the percentage split come in the second rank, and cross validation has got the least accuracy result since the dataset is small. Table 5.2 displays the accuracy results and Figure 5.6 plots the results.

Test Option	Cross Validation Folds=10			Use Training set			Percentage Spilt (70% Training, 30% Testing)			
	Classifier	SVM	NB	IBK	SVM	NB	IBK	SVM	NB	IBK
100 Tweets		91	92	86	100	99	100	91	93	86
250 Tweets		96	95	89	100	99	100	96	95	92
400 Tweets		96	96	90	100	98	100	95	96	92
650 Tweets		98	97	96	100	99	100	97	96	94
1000 Tweets		98	95	94	100	97	100	99	95	92

Table 5. 2: Test options comparison.

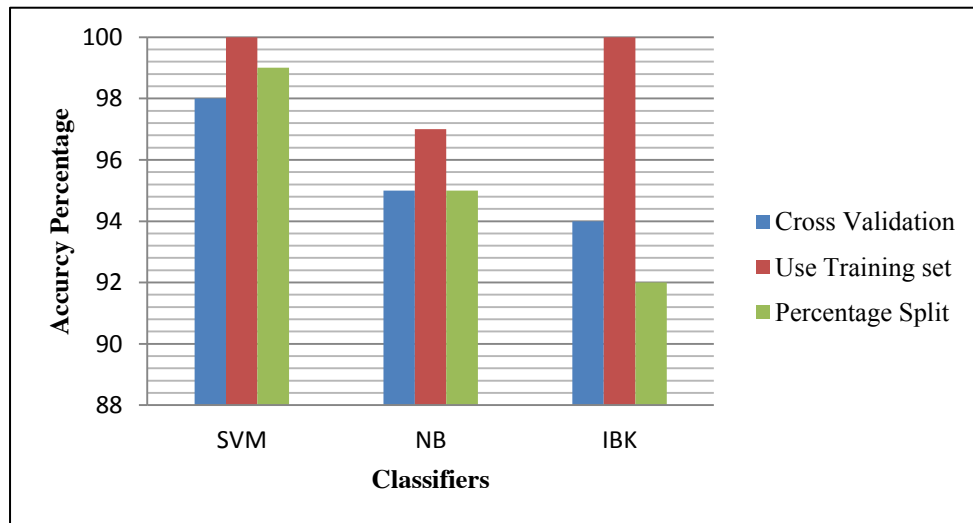


Figure 5. 6: Accuracy of classifiers based on the used test options.

## 5.2.3 Detailed Accuracy by Class

Third test was done to show the TP rate, FP rate, precision and F-Measure of the result of each classifier. The F-Measure is combined measure for precision and recall as this equation:

$$F\_Measure = 2 * Precision * Recall / (Precision + Recall) \quad (4)$$

Table 5.3 shows the result for each classifier, while Figure 5.7 shows the results as bar plot. Figure 5.8, shows the programmed code results when using SVM with UniGram and Figure 5.9 shows the WEKA GUI results and they match.

Class \ Classifier	TP Rate	FP Rate	Precision	F-Measure
SVM	0.98	0.01	0.98	0.98
NB	0.95	0.05	0.95	0.95
IBK	0.94	0.08	0.95	0.94

Table 5. 3: Detailed accuracy by class.

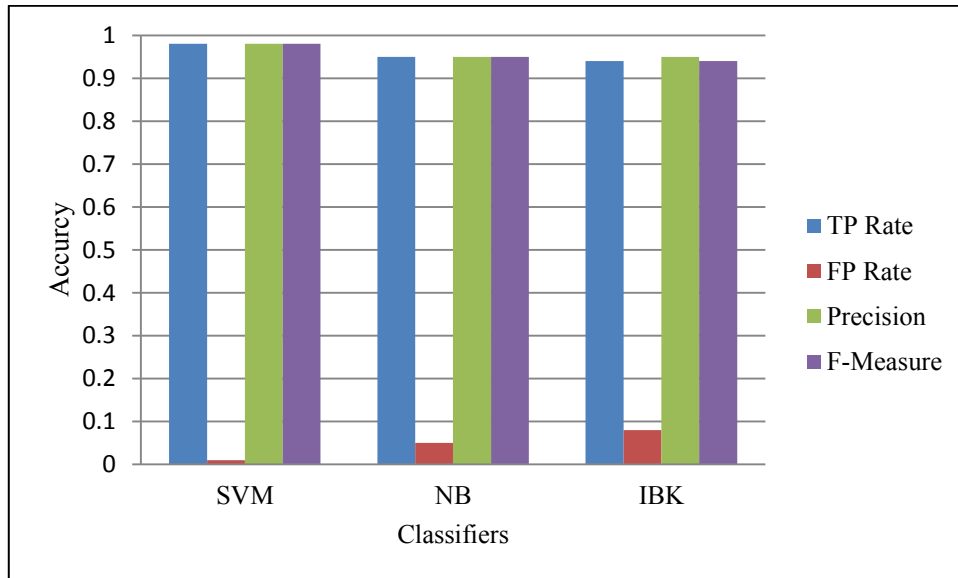


Figure 5. 7: Detailed accuracy by class.

```

***** The Classifier: SVM Classifier *****

=== The Test Option: Cross Validation ===

Correctly Classified Instances      986          98.6  %
Incorrectly Classified Instances    14           1.4  %
Kappa statistic                    0.9708
Mean absolute error                 0.014
Root mean squared error            0.1183
Relative absolute error             2.9164 %
Root relative squared error        24.1523 %
Total Number of Instances          1000

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
          0.988   0.018   0.988     0.988   0.988     0.985    Positive
          0.983   0.012   0.983     0.983   0.983     0.985    Negative
Weighted Avg.  0.986   0.015   0.986     0.986   0.986     0.985

=== Confusion Matrix ===

  a  b  <-- classified as
593  7  |  a = Positive
  7 393 |  b = Negative

```

Figure 5. 8: Code reults

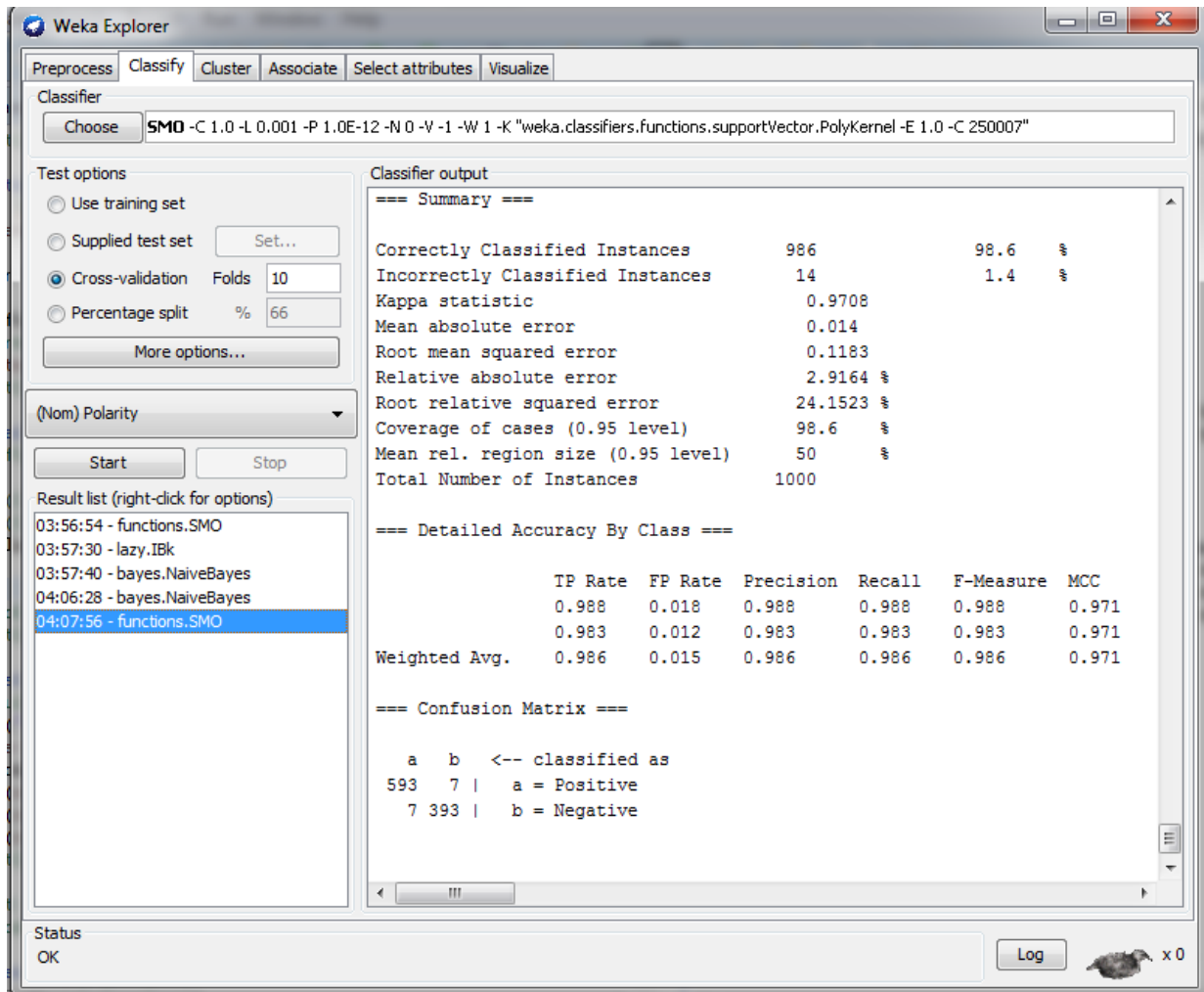


Figure 5. 9: WEKA GUI results.

## 5.2.4 Time Cost Comparison between Classifiers

Forth test shows that IBK has the zero time cost regarding the size of the dataset. Three dataset sizes was used which are 200, 650 and 1000. Table 5.4 shows the time for each dataset size has taken for each classifier to classify the whole dataset. Figure 5.10 plots the result which shows that after 650 tweets dataset size the curve has got a significant change between the SVM and NB. Appendix C shows the results for each test.

#Tweets \ Classifier	200 Tweets	650 Tweets	1000 Tweets
SVM	0.077	0.322	0.52
NB	0.083	0.242	0.61
IBK	0	0	0

Table 5. 4: Time cost over dataset size



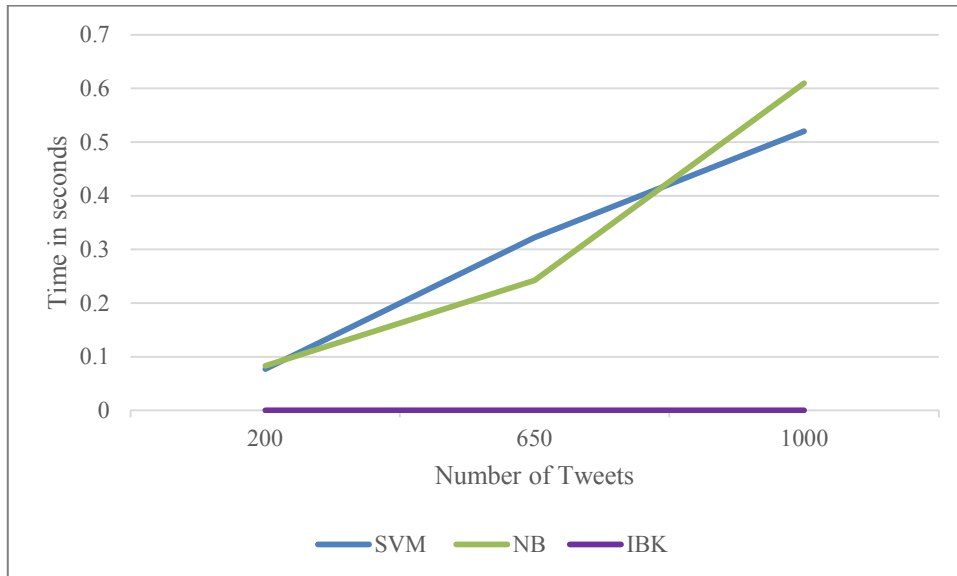


Figure 5. 10: Dataset size effect on time cost in seconds

### 5.2.5 Stop Words Effect over Classifiers Time

The final test was run over the supervised classifier shows the effect of removing stop words over the time that taken by SVM and NB. The test shows that NB benefits from removing the stop words more than SVM. Table 5.5 shows the time results for each classifier for five data set size. The test ran two times one test with removing stop words from the dataset and the other one without removing the stop words. Figure 5. 11 show the results. The time results of each classifier are presented in appendix C.

Classifier \ #Tweets	SVM without stop word	NB without stop word	SVM with stop word	NB with stop word
100 Tweets	0.036	0.032	0.047	0.032
250 Tweets	0.087	0.202	0.098	0.052
400 Tweets	0.213	0.143	0.149	0.079
650 Tweets	0.35	0.319	0.272	0.147
1000 Tweets	0.59	1.175	0.585	0.172

Table 5. 5: Stop words effect on classification time in seconds.

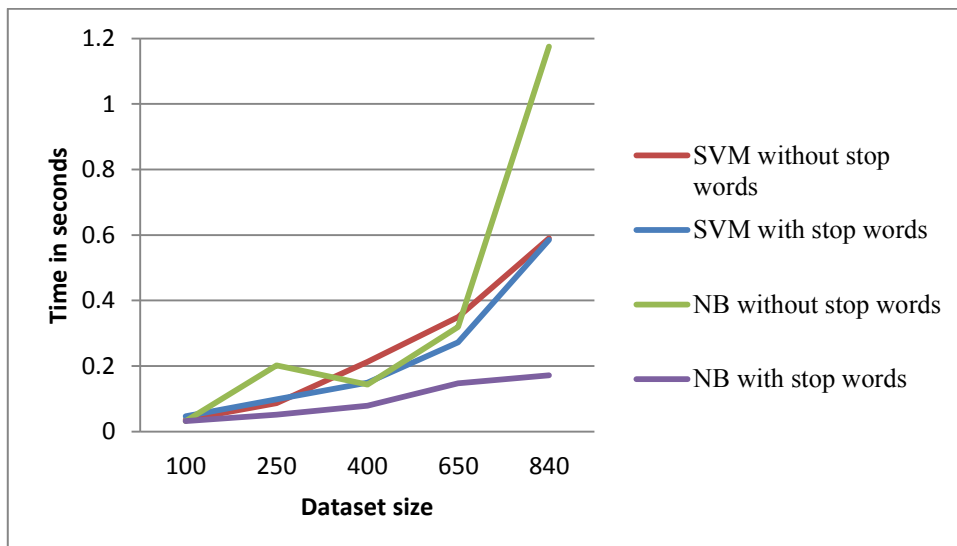


Figure 5. 11: Stop words effect on classification time in seconds.

## 5.3 Unsupervised Testing

The lexicon based test was applied on tweets. The results show that lexicon-based approach gives much lower accuracy compared to the corpus-based. The achieved accuracy was 78%. The accuracy of the unsupervised approach relies on the containment of the dictionary matched words to the tweets' words. Table 5.6 shows an example of classified tweets by the unsupervised classifier.

Original Tweet	Tweet after Prepressing and Filtering	Tweet's Polarity
جميبييل لعُهم اليوم، أعجبنى مرّة. The team has played weeeell today, I liked it much.	جميل لعهم اعجبنى مره The team has played well I liked it much	Positive
المدرّب شين، وكان اللاعبين أداؤهم ضعيف !!! Bad coach, and the players played poorly !!!	المدرّب شين اللاعبين اداؤهم ضعيف Bad coach, and the players played poorly	Negative

Table 5. 6: Example of classifying tweets.

The accuracy of the lexicon-based tool can be improved by expanding the dictionaries to include more words, and using words' stemmer.

### 5.3.1 Unsupervised Functions Testing

This section examines the system success of the system requirements. The tests have been done on the RA to validate the system functionalities and wither it works properly or not. The test was done over the user interface for the unsupervised classifier. Table 5.7 shows the requirement results of requirements. Figures 5.12, 5. 13 and 5. 14 display what will happen when the user enter a keyword. Figure 5.12 shows a bar plot for the number of the tweets that were classified as positive, negative and neutral. Words cloud will be displayed when the user clicks on the "سحابة الكلمات" tab as shown in Figure 5.13 and when he/ she clicks on "عرض التغريدات" tab a table of the classified tweets along with classification class for each tweet will be displayed as in Figure 5.14.

Test Case	Results
The user access the system online	Done as expected.
The system accepts the input from the user	Done as expected.
The system validate the entered keyword	Not success.
The system calculate the SA	Done as expected.
The system show the overall sentiment	Done as expected.
The system view the words cloud	Done as expected.
The system view the collected tweets	Done as expected.
The system view the user manual	Done as expected.
The system view the policy	Done as expected.

Table 5. 7: Requirements test results

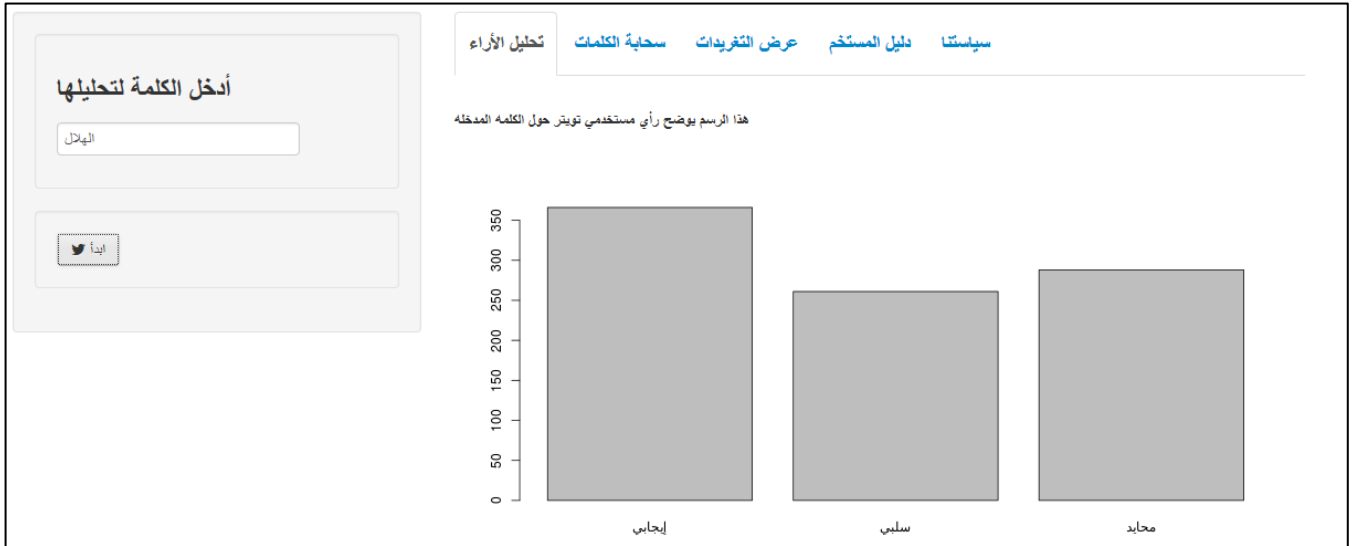


Figure 5.12: Overall sentiment bar plot.

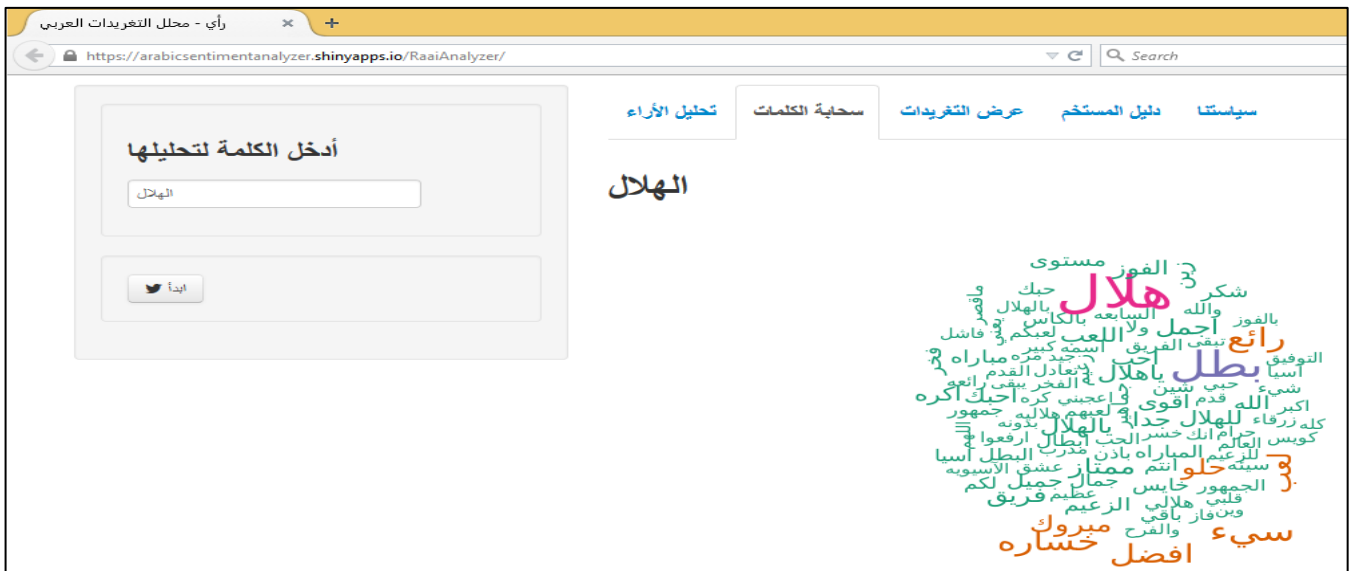


Figure 5.13: Words cloud function.



Figure 5.14: View tweets function

## 5.4 Comparing between Supervised and Unsupervised

Figure 5.15 shows how the system works in the supervised and unsupervised classifiers from two points of view. Which are the results of classification and the process flow of each one with superiority of the supervised over the unsupervised.

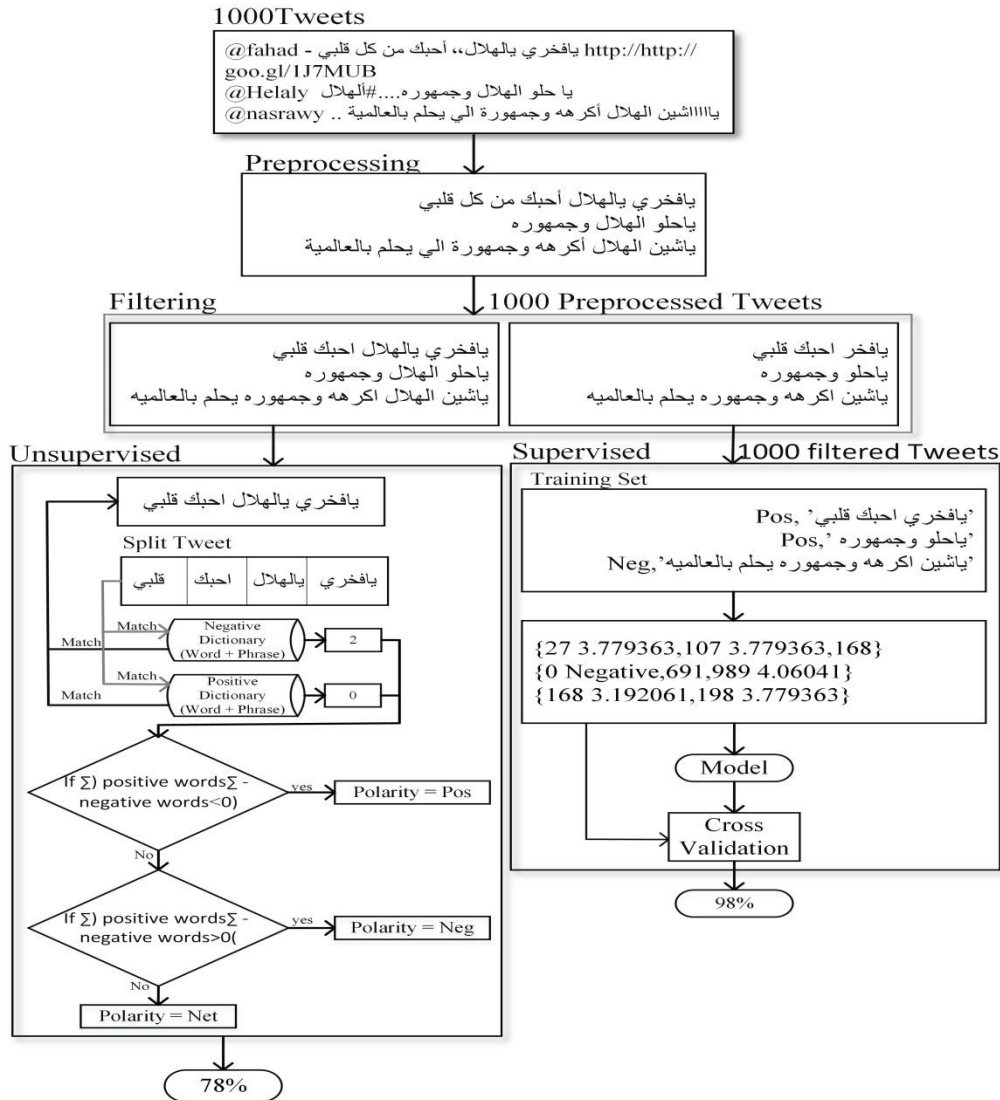


Figure 5. 1: Comparing the results and process flow of supervised and unsupervised classifiers

## 5.5 Conclusion

The chapter have displayed and reviewed the tests that were done over the Ra'ai Analyzer. The tests show that the user interface for the unsupervised approach works properly. The classifiers were tested extensively. The tests that were made over the supervised classifier show that SVM was the best for sentiment analysis. While the unsupervised classifier classify 73% of the collected tweets correctly.

**CHAPTER SIX: CONCLUSION AND FUTURE WORK**

## 6.1 Introduction

Sentiment analysis still in its early steps and it is considered as one of the main research trends among data scientists. Arabic language is one of the active languages on the web especially in social networks. This content could be useful therefore for the purpose of contributing and taking the advantages of it this project is needed. The project manipulate on one of the Arabic content sources which is Twitter to discover the public sentiment over a special topic.

This chapter will discuss the following topics: Review for the challenges that the project has faced in Section 6.2. The team's acquired skills will be reviewed in section 6.3. A list of the future directions and works that would be added to the project to improve its functionality will be discussed on Section 6.4. Finally, Section 2.5 summarizes the chapter.

## 6.2 Challenges

Many challenges have faced this project. Some challenges formed an obstacle to complete some functions and some the team was able to overcome them. The challenges that the project passed through are listed in Table 6.1.

Challenge	Description	Reflection
Limitation on Arabic support	The available development tools provide little support on recognizing Arabic letters and hence collecting Arabic tweets. This challenge was the biggest reason behind the project domain limitation.	Collect using three ways: <ol style="list-style-type: none"> <li>1. TwitteR package.</li> <li>2. Twitter Archivist tool.</li> <li>3. Manually.</li> </ol>
Difficulties and complexity of the Arabic language structure.	Arabic language has difficult structure.	Limit the system with classifying the simple structure sentences.
Concatenation of words.	Twitter users tend to concatenate two or more words to utilize the 140 characters they have such as typing (“هذا بلدمعطاء”) rather than typing (“هذا بلد معطاء”). That made it more difficult for the classifier to recognize the words as different separate words. And affect the classifier ability to classify if the concatenated words contain a sentimental word such as (حُبَّالك) rather than (حُبَّ الك).	The team has scan the tweets and found that the sentimental words that tend to be concatenated by the users are limited. Therefore the team has added these words to the positive and negative words dictionaries to enhance the classifier performance.
Absence of Arabic support in R.	While R is one of the best languages for Data analysis, Arabic was not well supported in libraries and packages that are available in R.	The team tried to overcome this challenge for the supervised approach but for time restrictions reasons the team decided to use Java as replacement for implementing the supervised. While the unsupervised was done completely in R.

Presence of Advertisement in the collected tweets.	The amount of unrelated tweets in the collected dataset using TwitteR package that have the needed keyword is overwhelmingly high. Around 90% of the targeted tweets are unrelated tweets with advertisement.	The team came up with scripts to eliminate most of the ads tweets and it was successful.
Absence of the word stemmer in R.	The absence of the word stemmer in R language, it could be used for the unsupervised approach.	The team decided to build the classifier without using stemmer.
Lack of resources.	A considerable number of papers were published in ACM library and the team did not have the access to this library.	The team used the abstract that were provided to understand the basic idea about some papers and their methodology.
Lack of papers on published Arabic SA field.	---	The team was searching on three libraries which are IEEE, ACM and ScienceDirect every period to find a new publication.

Table 6. 1: Challenges and reflections

## 6.3 Acquired Skills

This project was a great opportunity for the team to learn new skills. The skills varied between programming skills and communication skills. The skills are listed below:

- Learning R language which is one of the most important and power languages for data analysis.
- Learning Java language which is the most widely used for application.
- Team work was a big skill that the team learned from the project.

## 6.4 Future Work

The system can be enhanced and improved by adding more functions. Some functions were set as a future work because of the limitation on Arabic language processing. A list of future directions are:

- Classify live streams of tweets
- Provide an approach to deal with word negation.
- Use word stemmer to enhance the supervised and unsupervised classifier performance.
- Enhance the dictionaries of the unsupervised classifier to cover more domains.

## 6.5 Conclusion

This chapter reviewed summery of what have been done in the project. A list of the challenges that the project team passed through was listed along with a description for each challenge and how it was reflected on the team to yield a solution or a reaction. The future work was discussed by listing the possible enhancements on the system. Finally, a reviewing of the acquired skills that the team has gained through the system development was discussed.

## References

- [1] J. Gantz and D. Reinsel, "Digital Universe Study: Extracting Value from Chaos," EMC2, June 2011. [Online]. Available: Internet: <http://www.emc.com/leadership/programs/digital-universe.htm>. [Accessed 6 Nov 2014].
- [2] "The 2011 IDC Digital Universe study sponsored by EMC," [Online]. Available: "[Interhttp://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf](http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf). [Accessed 6 Nov 2014].
- [3] S. Sagioglu and a. D.Sinanc, "Big data: A review," in *Proc. CTS*, 2013, pp. 42 - 47.
- [4] "About," Twitter, [Online]. Available: <https://about.twitter.com/what-is-twitter>. [Accessed 6 Nov 2014].
- [5] "Social Media Usage in Middle East – Statistics and Trends [Infographic]," Go-Gulf, 4 Jun 2013. [Online]. Available: <http://www.go-gulf.com/blog/social-media-middle-east>. [Accessed 6 Nov 2014].
- [6] J.Han, M.Kamber and J.Pei., *Data Mining: Concepts and Techniques*, Waltham: Morgan kaufmann, 2012, pp. 24-16.
- [7] B.Liu, *Sentiment Analysis and Opinion Mining.*, Toronto: Morgan,Claypool, 2012, pp. 168-1.
- [8] A. Shoukry and a. A. Rafea, "Sentence Level Arabic Sentiment Analysis," in *Proc. CTS*, 2012, pp. 546 – 550.
- [9] U.Patil and J. Patil., "Web data mining trends and techniques," in *Porc. ICACCI '12*, 2012, pp. 961-965.
- [10] L. Wenjun, "An Security Model: Data Mining and Intrusion Detection," in *Porc. IIS*, 2010, pp 448 – 450.
- [11] B. Taati, J. Snoek, D. Aleman and A. Ghavamzadeh., "Data Mining in Bone Marrow Transplant Records to Identify Patients With High Odds of Surviva," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATI*C, vol. 18, p. 21 – 27, Jan 2014.
- [12] P.Giudici and S.Figini, *Applied Data Mining for Business and Industry*, 2nd ed., Chichester: A John Wiley and Sons, Ltd., 2009, pp. 176-178.
- [13] E. Dwoskin, "How New York’s Fire Department Uses Data Mining," Dow Jones and Company, Inc., 24 January 2014. [Online]. Available: <http://blogs.wsj.com/digits/2014/01/24/how-new-yorks-fire-department-uses-data-mining/>. [Accessed 21 November 2014].
- [14] A. Kao and S. Poteet, *Natural Language Processing and Text Mining*, Bellevue: springer, 2007.
- [15] R. Alhajj and J. Rokne., *Encyclopedia of Social Network Analysis and Mining*, New York: Springer New York, 2014, pp. 1688.
- [16] VISIO. [Online]. Available: <http://office.microsoft.com/en-gb/visio/>.
- [17] N.Abdulla1, N. Ahmed, M.Shehab and M.Al-Ayyoub, "Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based," in *Proc. AEECT*, 2013, pp. 1 – 6.



- [18] S. Ahmed and G. Qadah., "Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text," in *Porc. IIT*, 2013, pp. 72 – 77.
- [19] C. Aggarwal, *Data classification Algorithms and Applications*, New York: Taylor & Francis Group, 2015.
- [20] S. El-Beltagy and A. Ali, "Open Issues in the Sentiment Analysis of Arabic," in *Porc. IIT*, 2013, pp. 215-220.
- [21] K. Cai, Spangler, S. Y. Chen and L. Zhang., "Leveraging Sentiment Analysis for Topic Detection," in *Porc. WI-IAT*, 2008, pp.265-27.
- [22] P. Han, J. Du and L. Chen, "Web opinion mining based on sentiment phrase classification vector," in *Porc. ICNIDC*, 2010, pp. 308-312.
- [23] M. Shaikh, Prendinger, H. Ishizuka and M., "An analytical approach to assess sentiment of text," in *Porc. ICCIT*, 2007, pp.1-6.
- [24] M. Neethu and R. Rajasree, "Sentiment Analysis in Twitter using Machine Learning Techniques," in *Proc. ICCCNT*, 2013, pp. 1-5.
- [25] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perera, "Opinion Mining and Sentiment Analysis on a Twitter Data Stream," in *Proc. ICTer*, 2012, pp. 182-188.
- [26] H. Al-Khalifa and A. Al-Subaihin, "A System for Sentiment Analysis of Colloquial Arabic Using Human Computation," *The Scientific World Journal*, vol. 2014, p. 8, 2014.
- [27] K. Ahmad, D. Cheng and Y. Almas, "Multi-lingual Sentiment Analysis of Financial News Streams," in *Porc. GRID2006*, 2006. .
- [28] M. Abdul-Mageed, M. T. Diab and M. Korayem, "Subjectivity and sentiment analysis of modern standard Arabi0063," *The Scientific World Journal*, vol. 2014, 2011, pp. 587–591.
- [29] J. Salamah and A. Elkhilfi, "Microblogging Opinion Mining Approach for Kuwaiti Dialect," in *Proc. ICCTIM*, Dubai, 2014.
- [30] M. Abdul-Mageed, S. K"ubler and a. M. Diab, "SAMAR: A System for Subjectivity and Sentiment Analysis of Arabic Social Media," in *Proc. WASSA*, 2012, pp. 19-28.
- [31] S. Al-Osaimi and K. Badruddin, "Role of Emotion icons in Sentiment classification of Arabic Tweets," in *Porc. MEDES '14*, 2014, pp. 167-171.
- [32] R. Duwairi, R. Marji, N. Sha'ban and S. Rushaidat, "Sentiment Analysis in Arabic Tweets," in *Porc. ICICS*, 2014, pp. 1 - 6.
- [33] L. Albraheem and H. Al-Khalifa, "Exploring the problems of Sentiment Analysis in Informal," in *Proc. IIWAS '12*, 2012, pp. 415-418.
- [34] "Stop words list," [Online]. Available: <https://code.google.com/p/stop-words/>. [Accessed 25 Des 2014].
- [35] "Weka 3: Data Mining Software in Java," WEKA The Univeristy of Waikato, [Online]. Available:

<http://www.cs.waikato.ac.nz/ml/weka/>. [Accessed 25 12 2014].

[36] " Arabic MPQA subjective lexicon & Arabic opinion holder corpus," 23 May 2012. [Online]. Available: <http://nlp4arabic.blogspot.com/2012/05/arabic-mpqa-subjective-lexicon-arabic.html>. [Accessed 25 Dec 2014].

[37] R. Khasawneh, H. Wahsheh, M. Al Kabi and I. Aismadi, "Sentiment analysis of arabic social media content: a comparative study," in *Proc. ICITST*, 2013, pp. 101 - 106.

# **Appendix**

## Appendix A: Stop Words

يا	راح	وفي
اللي	مثل	انها
له	ي	لأنهم
لها	بك	وقف
واذا	انت	ومن
شي	مافيه	وهو
أنتم	مافي	صفر
ع	انا	ايضا
ياخي	وش	ذا
هاذا		

TABLE A. I: Saudi dialect Stop words

لم
لن
ولا
ولم
لا
ما

TABLE A. II: Removed Stop Words (Negation)

الا	اما	يشكل	حتى	غدا	كما	الماضي	هذا	الوقت
ا	امس	بعد	حيث	غير	لان	مافي	هذه	وقد
اثر	ان	بعض	الذاتي	ف	لأنهم	مافيه	هناك	وقف
اجل	انا	بك	الذي	فان	لدى	مايو	هو	وكان
احد	انت	بل	راح	فى	لقاء	مثل	هي	وكانت
اخرى	أنتم	بن	زيارة	فيه	لكل	مساء	و	ومن
الاخيرة	انها	به	السابق	فيها	لكن	مع	واحد	وهو
اذا	او	بها	سنة	قال	لمن	مقابل	واذا	وهي
اربعة	اي	تم	سنوات	قبل	له	مليار	واضاف	ي
اطار	ايضا	التي	شخصا	قد	لها	من	واضافت	يا
اعادة	ب	الثاني	شي	قوة	لهذا	منذ	واكد	يكون
اعلنت	باسم	الثانية	صباح	كان	كانت	منها	وان	يمكن
اف	بان	ثلاثة	صفر	كانت	لهذا	نحو	واوضح	يوم
اكثر	بدون	ثم	ضد	كل	لو	نفسه	وفي	
اكذ	برس	جميع	ضمن	كلم	لوكالة	نهاية	وقال	
ام	بسبب	حاليا	ع	كم	ما	هاذا	وقالت	

TABLE A. III: Stop Words List After Modification

## Appendix B: Saudi Positive and Negative Word Dictionaries

ابكي	تضيعون	خسرت	ظلمونا	لا	مستهتر	وبلا
احباط	تفشلتوا	خسرتم	عقدتكم	للاسف	مشكله	وجع
اردى	جنون	خسرتو	عييب	لم	معاناه	وخذلان
ازعجونا	حسره	خسرو	غلط	لن	مغرور	وخسونا
اضاع	حاقده	خطا	غير	ما	مغرورين	وعدانتيه
اططقت	حرام	دجه	فاشل	مخيس	مكروه	ولا
افا	حرم	دموع	فاشلين	ماتتشلون	منتهم	ولم
اكره	حزن	سخيّف	قتلت	ماحالفهم	منتوب	ومتكبر
اوجاع	حزني	سوداء	قتلنا	مافرت	مو	ويقهري
بكي	حسافه	سيء	قدر	ماقدرت	موب	يخسر
بلا	خاب	شين	كراهيه	مايغثوني	نكي	يكرهكم
تبطون	خايس	صعبه	كرهتكم	مايمديكم	نكبه	ينرفز
تبكون	يخسر	ضده	كرهتونا	محبطه	نكره	ينهمزم
تخرن	خذلت	ضدهم	كرهك	محرومين	هاردلك	
تخسرون	خساره	ضعيف	كريه	مر	هريمتكم	
تخلف	خسرو	ظالمه	كفو	مرعبه	وابكي	
تخلف	خسر	ظل	كلب	مستحيله	ندامه	

TABLE B. I: Saudi Dialect Negative Words

ابدعوا	اهنيكم	حبا	شكرا	فرحه	ممتاز	وز عيما
اجتهد	باهر	حبك	صحيح	فزت	ناجح	وسنظل
احب	بتوفيق	حبنا	صداره	فن	نحب	وشرفتنا
احبك	بطل	حبيبي	صدق	فوز	نشكر	وشكر
ادعم	بطلا	حبي	عز	قوي	نفرح	وعاطفه
اسهل	بفن	حبيبي	عزك	لعشاق	نؤازوهم	وعشقك
اشكركم	بفوز	حبيبتك	عزنا	للجم	هزم	وفخر
اعجبني	تبقى	حظ	عشق	للفخر	هيبه	وفق
اعشق	تستاهلون	حمدالله	عشقك	لمعشوقني	واجمل	ومتفوق
افتخر	تستحقها	دعم	عشقنا	ماقصر	واطربتنا	ومحبه
افرح	تفداك	راضين	غي	ماقصرنوا	وافضل	ونفرح
افرحتها	تفرحهم	راقي	فاز	ميروك	واقوى	ورفاء
افضلها	تفز	رائع	فازوا	فخر	ميروك	يافخر
امتعتنا	توفيق	رائعه	فخر	ميروك	وجيد	يحبونه
امتعني	توفيق	رضينا	فخرا	متعته	وحب	يرضي
ابدعوا	جميل	زين	فرح	متفانله	وحيبت	يستحق
افضل	حب	شكر	فرحتنا	متقن	ورائع	وز عيما
يستطع	يفتخر	يفرح	يفوز	يهيل		

TABLE B. II: Saudi Dialect Positive Words

## Appendix C: Testing Results

Classifier #tweets	SVM with stop word									
100	0.02	0.02	0.03	0.11	0.02	0.02	0.03	0.02	0.02	0.03
250	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06
400	0.09	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.06
650	0.14	0.14	0.14	0.14	0.16	0.16	0.13	0.13	0.2	0.13
840	0.17	0.19	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
Classifier #tweets	NB with stop word									
100	0.05	0.03	0.02	0.02	0.03	0.03	0.13	0.11	0.03	0.02
250	0.14	0.08	0.08	0.07	0.07	0.08	0.06	0.13	0.08	0.19
400	0.2	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.19	0.12
650	0.28	0.23	0.25	0.23	0.25	0.5	0.23	0.23	0.22	0.3
840	0.44	0.46	0.48	0.61	0.44	0.47	0.45	0.81	0.88	0.81

**TABLE C. I: Stop words Affication on classifir time**

Test Option	Cross Validation with UniGram in Java Code																																
Classifier	SVM										Average	NB										Average	IBK										Average
100 Tweets	89	93	93	89	94	89	90	91	93	95	91	99	92	92	91	93	90	89	91	91	92	92	88	80	82	83	88	85	86	87	89	94	86
250 Tweets	95	95	95	99	95	93	97	97	97	96	96	96	94	97	96	95	93	95	96	96	97	95	92	88	88	93	89	87	99	86	86	89	89
400 Tweets	96	96	96	97	98	96	98	97	94	97	96	96	95	97	96	97	95	96	96	97	95	96	91	89	90	90	89	89	94	90	91	90	90
650 Tweets	97	99	97	98	98	99	98	99	98	98	98	96	99	97	98	98	97	96	97	98	98	97	86	99	96	97	98	96	94	98	97	98	96

Test Option	Cross Validation with UniGram in WEKA GUI																																
Classifier	SVM										Average	NB										Average	IBK										Average
100 Tweets	89	93	93	89	94	89	90	91	93	95	91	99	92	92	91	93	90	89	91	91	92	92	88	80	82	83	88	85	86	87	89	94	86
250 Tweets	95	95	95	99	95	93	97	97	97	96	96	96	94	97	96	95	93	95	96	96	97	95	92	88	88	93	89	87	99	86	86	89	89
400 Tweets	96	96	96	97	98	96	98	97	94	97	96	96	95	97	96	97	95	96	96	97	95	96	91	89	90	90	89	89	94	90	91	90	90
650 Tweets	97	99	97	98	98	99	98	99	98	98	98	96	99	97	98	98	97	96	97	98	98	97	86	99	96	97	98	96	94	98	97	98	96

Test Option	Cross Validation with BiGram in Java Code																																
Classifier	SVM										Average	NB										Average	IBK										Average
100 Tweets	72	71	73	76	83	71	78	67	69	71	73	72	71	73	75	81	71	76	67	72	72	73	73	69	72	75	80	69	74	64	67	71	71
250 Tweets	77	78	75	91	77	76	88	79	72	76	79	74	83	73	87	77	75	87	78	72	74	78	77	74	69	86	74	74	80	74	68	70	74
400 Tweets	83	78	81	82	80	83	95	90	83	80	83	80	74	78	75	76	80	79	73	79	80	77	80	75	76	82	76	77	82	80	77	79	78
650 Tweets	99	99	94	95	96	96	93	98	92	94	95	99	92	87	92	89	87	86	97	87	87	90	80	95	93	94	94	92	90	98	93	92	92

Test Option	Cross Validation with BiGram in WEKA GUI																																
Classifier	SVM										Average	NB										Average	IBK										Average
100 Tweets	72	71	73	76	83	71	78	67	69	71	73	72	71	73	75	81	71	76	67	72	72	73	73	69	72	75	80	69	74	64	67	71	71
250 Tweets	77	78	75	91	77	76	88	79	72	76	79	74	83	73	87	77	75	87	78	72	74	78	77	74	69	86	74	74	80	74	68	70	74
400 Tweets	83	78	81	82	80	83	95	90	83	80	83	80	74	78	75	76	80	79	73	79	80	77	80	75	76	82	76	77	82	80	77	79	78
650 Tweets	99	99	94	95	96	96	93	98	92	94	95	99	92	87	92	89	87	86	97	87	87	90	80	95	93	94	94	92	90	98	93	92	92

Test Option	Percentage Split with UniGram in Java Code																																
Classifier	SVM											Average	NB								Average	IBK								Average			
100 Tweets	96	86	96	93	96	93	80	90	96	90	91	96	73	96	93	90	94	95	96	95	98	93	96	86	83	76	86	93	76	73	86	100	86
250 Tweets	94	92	93	100	97	93	98	100	97	96	100	92	98	89	96	88	93	97	98	97	95	92	93	88	92	89	92	94	92	97	89	92	
400 Tweets	91	98	96	99	97	94	95	96	95	98	95	97	98	97	95	96	97	95	96	97	94	96	87	88	90	97	96	91	92	88	95	87	92
650 Tweets	96	99	96	95	98	98	96	100	98	98	97	96	98	94	96	94	97	91	98	98	96	96	90	96	94	94	97	94	86	98	97	98	94

Test Option	Use Training Set with UniGram in Java Code																																	
Classifier	SVM											Average	NB								Average	IBK								Average				
100 Tweets	100	100	100	100	100	100	100	100	100	100	100	100	99	100	100	99	100	100	100	99	99	100	99	100	100	100	100	100	100	100	100	100	100	100
250 Tweets	100	100	100	100	100	100	100	100	100	100	100	100	99	97	100	98	99	99	98	98	99	99	99	100	100	100	100	100	100	100	100	100	100	100
400 Tweets	100	100	100	100	100	100	100	100	100	100	100	98	98	99	98	98	99	99	98	99	97	98	100	100	100	100	100	100	100	100	100	100	100	
650 Tweets	100	100	100	100	100	100	100	100	100	100	100	98	99	98	99	99	98	98	99	99	99	99	100	100	100	100	100	100	100	100	100	100	100	

Classifier	Percentage Split BiGram in Java Code																																
Classifier	SVM											Average	NB								Average	IBK								Average			
100 Tweets	80	76	83	86	90	80	80	50	56	70	75	80	76	83	86	86	83	80	50	56	73	75	80	76	56	86	86	80	80	46	50	63	70
250 Tweets	73	76	80	85	77	80	81	80	76	77	79	77	76	80	80	81	84	72	81	76	78	79	69	76	76	85	72	77	70	76	81	73	75
400 Tweets	78	77	79	78	80	82	88	88	85	86	82	75	75	77	77	80	80	75	72	83	88	78	76	74	75	82	83	76	86	72	78	84	78
650 Tweets	92	96	91	91	92	94	88	99	90	91	92	76	93	78	90	83	81	76	90	88	85	84	78	96	88	90	90	90	84	98	90	89	89



## Appendix D: Reflections and Observations

1. With regard to **analyzing the local and global impact of your project on individuals, organizations, and society**, indicate how and where you demonstrated the extent by which you:

- Understood the impact of computing solutions on society and the environment in a local context. [(g)1a]

The project gives the organizations (whether they were governmental or special sector) the ability to understand their audience (clients or public), what is the overall sentiment of the audience towards their services.

Page: 8

- Considered a variety of available options in computing design and make a proper choice based on their impact. [(g)1b]

The team has scanned the literature review to find the best approaches to implement the SA. The comparison that was done between the supervised and unsupervised in page 18 shows that each one can Succeed under certain circumstances. So the team have used to the approach to implement the SA

Page: 18

- Understood the impact of computing solutions on society and the environment in a global context. [(g)2a]

The project gives the organizations (whether they were governmental or special sector) the ability to understand their audience (clients or public), what is the overall sentiment of the audience towards their services.

Page: 8

2. With regard to use of **techniques, skills, and tools, indicate how and where** you demonstrated the extent by which you:

- Wrote programs in a modern computer language. [(i)1a]

The project used two of the most power languages which are Java and R. Java is the most used language for programming where R is one of the most important languages for data analysis.

Page: 37,

- Used appropriate software or hardware design tool applications. [(i)2a]

The project diagrams were built with the help of VISIO software and the project logo was designed using Photoshop.

Page: 33

- Applied appropriate modern computer-based analysis and design tools. [(i)2b]

- Utilized problem solving skills and techniques to complete a task. [(i)2c]

- 1- Machine learning packages in R do not support Arabic language. This make the team decides to WEKA packages in Java to implement the supervised approach. Page: 37
- 2- One of the main problems that have faced the team is collecting the tweets. The team used many ways to collect that are Twitter API, Twitter archivist tool and manually. The collecting have took one month to collect a suitable size of dataset. Page: 35, 52

- Demonstrated knowledge of programming tools. [(i)3a]

The project was built by different software tools such as Rstudio to develop R programs, Eclipse for Java development, ShinyApp.IO for uploading the RA on server and GitHub for downloading R packages.

Page: 37-38

- Demonstrated knowledge of design tools. [(i)3b]

3. With regard to **applying foundations of mathematics, algorithms, and computer science in modeling and design**, indicate how and where you demonstrated the extent by which you:

- Calculated time and space complexity for a program. [(j)1a]

The algorithms that was constructed for implementing the unsupervised approach has  $O(N^2)$  time complexity.

Page: 36

- Described trade-offs between performance and cost for algorithms used. [(j)2a]

Not applicable.

- Described the computing theory underlying an assignment. [(j)3a]

The models for the supervised approach were built by the help of three algorithms (SVM, NB, IBK). The team chooses these algorithms based on the literature review which shows that the most used algorithms for SA are the chosen on.

Page: 15, 37

4. With regard to **applying design and development principles**, indicate how and where you demonstrated the extent by which you:

- Described a system life-cycle. [(k)1a]

The system for the supervised and unsupervised pass through four main steps. Starts first by accepting a keyword then retrieve the tweets related to it. After that the dataset will be preprocessed then filtered. Finally the clean tweets enter the classifier to classify it as positive and negative. For the unsupervised the user will be able to view bar plot that represent the overall sentiment, word clouds and tweets information.

Page: 24, 25

- Wrote documentation for different phases for the development cycle. [(k)2a]

The document Demonstrate the phases that the project passed though starting from the literature review till testing the final product

Page: 10, 23, 34, 42, 50

## Appendix E: Name of Published Papers in Arabic Sentiment Analysis

	Web and Blogs	Twitter	Facebook
1	Some methods to address the problem of unbalanced sentiment classification in an Arabic context	Open issues in the sentiment analysis of Arabic social media: A case study	Social Networks' Facebook' Statutes Updates Mining for Sentiment Classification
2	Visualising sentiments in financial texts?	Sentiment Analysis in Arabic tweets	Classifying sentiment in Arabic social networks: Naïve search versus Naïve Bayes
3	Mining Arabic Business Reviews	Sentence-level Arabic sentiment analysis	Text mining Facebook status updates for sentiment classification
4	An empirical study to address the problem of Unbalanced Datasets in sentiment classification	Topic extraction in social media	Sentiment Analysis on Social Media
5	An analytical study of Arabic sentiments: Maktoob case study	SAMAR A System for Subjectivity and Sentiment Analysis of Arabic Social	Arabic Sentiment Analysis using Supervised Classification
6	Identification of opinions in Arabic newspapers	A Comparative Study of Social Media and Traditional Polling in the Egyptian Uprising of 2011	Sentiment analysis of Arabic social media content: a comparative study
7	Mining opinions in Arabic text using an improved “Semantic Orientation using Pointwise Mutual Information” Algorithm	Role of Emotion icons in Sentiment classification of Arabic Tweets	An Opinion Analysis Tool for Colloquial and Standard Arabic
8	A Program for Opinion Survey in Arabic for	Exploring the problems of sentiment analysis	

	Servicing the Media and Researchers in the Middle East*	in informal Arabic	
9	Agile Sentiment Analysis of Social Media Content for Security Informatics Applications	Arabic Sentiment Analysis using Supervised Classification	
10	Estimating the sentiment of social media content for security informatics applications	Social media evolution of the Egyptian revolution	
11	A proposed sentiment analysis tool for modern Arabic using human-based computing.	Key issues in conducting sentiment analysis on Arabic social media text	
12	Arabic opinion mining using combined classification approach	Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based	
13	Subjectivity and Sentiment Analysis of Modern Standard Arabic	Microblogging Opinion Mining Approach for Kuwaiti Dialect	
14	An extended analytical study of Arabic sentiments	Sentiment analysis of Arabic social media content: a comparative study	
15	opinion Mining and Analysis for Arabic Language		
16	Ensemble of Classification Algorithms for Subjectivity and Sentiment Analysis of Arabic Customers' Reviews		
17	Finding Opinion Strength Using Rule-Based Parsing for Arabic Sentiment Analysis		

18	Opinion Mining from Arabic Quotations		
19	Aara'- a system for mining the polarity of Saudi public opinion through e-newspaper comments		

**TABLE E. I: Published papers on Arabic SA**