



Computer Sciences Department  
Faculty of Computers and Information  
Mansoura University

# Social Networks User Modeling

---

SUBMITTED BY

**Wasan Adballah Abdlateef**

Computer Science Department, Faculty of Computers and  
Information, Mansoura University, Egypt  
Ministry of Higher Education and Scientific Research Al-Qadisiya  
University, Iraq

A Thesis

Submitted to the Computer Sciences-Department Faculty  
of Computers and Information-Mansoura University  
In Partial Fulfillment of the Requirements for the Master  
Degree in Computer Science

UNDER THE SUPERVISIONS OF

**Assist. Prof. Samir Eldesoky Elmougy**

Chair of Computer Science Department  
Faculty of Computers and Information  
Mansoura University

**Dr. Shahenda Salah El-Din Sarhan**

Lecturer of Computer Science Department  
Faculty of Computers and Information  
Mansoura University

**Mansoura University - Egypt**

**2016**

## *Acknowledgment*

First of all, I thank **Allah** for achieving this work and giving me the ability to finish it. Second, I would like to express my appreciation to Assist. **prof. Dr. Samir Eldesoky Elmougy** for his supervision on my thesis and his continuous support and encouragement during the research study in this thesis.

I would like to express my thanks to my **supervisor Dr. Shahenda Salah Sarhan** for her supportive help and advice starting from the initial ideas of thesis until finalizing it.

Special acknowledgment is given to them for continuous support during the steps of this work and their strong concern about all work details, even small ones.

I extend my sincere thanks and sincere gratitude to **Prof. Dr. Mohamed Eisa** and **Prof. Dr. Magdy Zakaria Rashad** which made an honor agreed to be in the jury venerable, and that observation will take into account and will raise the value of this research.

Thank for all the people of the Arab Republic of Egypt for good morals and high generosity, dedicate the fruit of my best to my beloved Iraq. God save Iraq and its people. The latest pretext that Praise be to Allah.



To my dear husband...

**Moath Al awsi**

To my babies

**Zenaib**

**Montather**

**Salaam, Hager**

To my family,

brothers and sisters

To all who gave a helping hand

to me .. my friends loyal



## **Abstract**

Social networks is a source of great interest to researchers because of the rapid attention of peoples, the contents of social networking and the variety of transmitted data between one site and various sites (such as photos, messages, and personal information, news, websites, scientific research, and other information).

The main objective of this study is to build a user modeling to predict user interests in Facebook site based on user activity in different locations, and the most associated attributes. This model uses direct features derived from Facebook users depending on unsupervised learning method and rough set theory. The main stages of the proposed model are data collection, data preprocessing, clustering using SOM, and reduction. The used dataset was collected from 680 Facebook users including family, friends circle, and college students from the third and fourth years at the Faculty of Computers and Information, Mansoura University, Egypt. The results showed that the proposed model reduced the number of attributes by 84%, and also gave an accuracy rate 94.28%.

## List of Figures

<b>Figure name</b>	<b>Page</b>
Figure 3.1: Structure of an Artificial Neuron	19
Figure 3.2: Overall SOM algorithm	25
Figure 3.3: SOM input and output layers	26
Figure 4.1: The general workflow of users modelling in social networks	39
Figure 4.2: The basic structure of the framework	41
Figure 4.3: The preprocessing data stage	42
Figure 4.4: A: The gender distribution of users in the dataset B: The educational distribution of users in the dataset	43
Figure 4.5: General confusion matrix	47
Figure 4.6: Classifier evaluation using 6-Fold Cross-Validation	48
Figure 5.1: Example of using SOM algorithm	51
Figure 5.2: Results of executing SOM	52
Figure 5.3: The percentage of the attributes before and after reduction	60
Figure 5.4: Description of the particular distribution in testing model	62
Figure 5.5: Description of the particular distribution in testing model (users outside Facebook site)	63

## List of Tables

Subject	Page
Table 1.1: The average active daily users of the first 10 used SNS	2
Table 2.1: Abstract SN interaction paradigms and their underlying native count	14
Table 2.2: A comparison between the number of popular sites	15
Table 3.1: Positive and negative features of neural networks	22
Table 3.2 : Type of ANN learning	23
Table 3.3: The main advantages and disadvantages of SOM	27
Table 3.4: Advantages and Disadvantages of NB	31
Table 4.1: Demographic information profiles	43
Table 4.2: Set of attributes in the used data set	46
Table 5.1: Results of the reduction algorithms.	53
Table 5.2: Result of executing NB	53
Table 5.3: Executing CV with 90% training and 10% testing	54
Table 5.4: Executing CV with 80% training and 20% testing	54
Table 5.5: Executing CV with 70% training and 30% testing	55
Table 5.6: Executing CV with 60% training and 40% testing	55
Table 5.7: Executing CV with 50% training and 50% testing	55
Table 5.8: Executing CV with 40% training and 60% testing	56
Table 5.9: Executing CV with 30% training and 70% testing	56
Table 5.10: Executing CV with 20% training and 80% testing	56
Table 5.11: Executing CV with 10% training and 90% testing	57
Table 5.12: Sensitivity, Specificity, and Accuracies (%) results for k=10 Different Participations	58
Table 5.13: Cross Validation Result for all Partition	58
Table 5.14: Evaluation measurements of reduction and rules produced by different algorithms	59
Table 5.15: The output confusion matrix	63

## Abbreviations

<b>Term</b>	<b>Description</b>
AIM	An Instant Messenger
ANN	Artificial Neural Network
AOL	America Online
ART	Adaptive Resonance Theory
CV	Cross-Validation
FN	False Negative
FP	false positive
GA	Genetic Reduction
ICQ	Internet Chat Query
JA	Johnson Algorithm
LVQ	Learning Vector Quantization
MLP	Multi-Layer Perceptron
NB	Naive Bayes
NN	Neural Network
RNB	Rough Set Naive Bayes
RS	Rough Set
RST	Rough Set Theory
SQL	Structure Query Language
SN	Social Network
SNS	Social Network Sites
SOM	Self-Organizing Map
SVA	SINE-VNTR- <i>Alus</i>
TN	True Negative
TP	True Positive
UM	User Modeling

# Contents

<b>Chapter One: Introduction</b>	<b>1-5</b>
1.1 Introduction	1-2
1.2 Problem Statement	2-3
1.3 Cold-Start Problem	3
1.4 Motivation	3
1.5 Thesis Objectives:	4
1.6 Challenges	4
1.7 Thesis Contributions	4-5
1.8 Thesis Contents	5
<b>Chapter Two: Social Networks: Overview</b>	<b>6-17</b>
2.1 Introduction	6-7
2.2 Benefits of Social Networks	7-8
2.3 Social Networks Categories:	8-12
2.4 Social Networks Technologies	12-13
2.5 Social Networks Comparison	13
2.6 User Modeling(UM)	13-16
2.6.1 Modeling of SN	16-17
<b>Chapter Three: Background and Related Works</b>	<b>18-36</b>
3.1 Introduction	18
3.2 Artificial Neural Networks	18-27
3.3 Rough Set (Theoretical Aspects)	27



## *Contents*



3.3.1 Attribute Reduction	27-30
3.3.2. Naïve Bayesian (NB) Classification	30-32
3.3.3. Cross-Validation	32-33
3.4. Related works	33-36
3.5 Summary	36
<b>Chapter Four: The Proposed Model Structure</b>	<b>37-49</b>
4.1 Introduction	37
4.2 Scope and Solution Plan	38
4.3 The Proposed Methodology	38-39
4.4 The Proposed System	40
4.4.1 Data Collection	40-41
4.4.2 Data Preprocessing	42-45
4.4.3 Clustering	45
4.4.4 Reduction	46-47
4.4.5 Classification	47
4.4.6 Cross-Validation	47-49
4.4.7 Simulation Model	49
4.5 Summary	49
<b>Chapter Five: Experimental Results and Discussion</b>	<b>50-63</b>
5.1. Experimental Results	50
5.2 SOM algorithm	50-52
5.3 Rough Set	52-60
5.4 Model Results	60

## *Contents*



5.5 Discussion	61-63
<b>Chapter Six: Conclusion and Future Work</b>	<b>64-65</b>
6-1 Conclusion	64-65
6.2 Future Work	65
<b>References</b>	<b>66-75</b>
<b>Appendixes</b>	<b>76-84</b>



# ***Chapter One***

## ***Introduction***

# Chapter One

## Introduction

### **1.4 Introduction:**

A Social Networking Sites (SNS) offer collaborative platform that enables its users to communicate with other providers to create social relations to share knowledge and information [1]. SNS can be defined as a communication between users within virtual communities and networks such as Twitter, Blogs and Facebook [2]. SNS encompass profile data like name, gender, old, and nationality [3]. SNS provide communication environment that offers its users collaborative structure online with ease and supported using instant message, offline messages, e-mail, and posting images and pictures [4]. Because the services conferred by SNS are specific to belong to users at these sites, SNS is working as a service interactive of users within a specific community.

SNS have allows users to share thoughts, events, posts, events, images, and preferences between participants in their sites. Table (1.1) shows the average active daily users of the first 10 most used SNS [5].

**Table 1.1: The average active daily users of the first 10 used SNS [5].**

SNS	median active daily users
Facebook	1.184 billion users
TencentWeibo	220 Million users
Tumblr	230 Million users
Twitter	232 Million users
LinkedIn	259 Million users
Wechat	272 Million users
Google +	300 Million users
WhatsApp	400 Million users
Ozone	632 Million users
Tencent QQ	816 Million users

### **1.5 Problem Statement:**

In this study, researcher try to determine the attributes for users in Social Networks (SN) to help in building user's social profile to make more effective recommendations to the user and to enable web site developers of adoption. Despite active works in this area in recent times, the access to the results of a satisfactory implementation process is still difficult because of the shortcomings of existing recommendations on the content. And the

researches depend on the content analysis that may be efficient within a given area of knowledge but not efficient in other various fields.

### **1.3 Cold-Start Problem:**

The cold-start is one of the major SNS problems. The solve problems adopted mostly on historical data, data like ratings of users or product features, and the similarity between user's features can't give useful results in new applications.

### **1.4 Motivation:**

SNS widespread attracted a massive number of users. Facebook is a typical example of SNS being accommodate the biggest number of users as revealed in Table (1.1). It is the most popular in terms of the normal activities as it includes a large and varied amounts of information like photos, messages, video, news, events, etc. Hence, this site attracts a lot of interesting researchers.

Prompting researcher thinking to yield advantage of this data is actively used various social sites and yields advantage of automated learning of techniques to get a composition of the model to the concerns of user's social sites Facebook.

### **1.5 Thesis Objectives:**

The significance of the study is not only represent in the fact that SNS urges communication between people, but also in being the most down to the world, we may describe the significance of the study as follows:

1. Finding and discovering the concerns side of Facebook users.
2. Identifying the correlation between personal features of the users' demographic profile and possible roles which appear in their online behaviors.
3. Recognizing the discovery from dataset and converting raw data into useful ones and the utilization of modern and sophisticated algorithms for building users model on social networks using a new sample of community.

### **1.6 Challenges:**

The most important challenges facing us in the preparation of this study is to gather information as it is the Arab societies of social determinants and also due to the announcement of the users in their daily activities and our personal page on social sites information concerns.

### **1.7 Thesis Contributions:**

Building a model for users of important SN preoccupied researchers for many reasons, the most important of the speed of deployment, and the large

amount of data is very transmitted among them, for this was the main objective of our study is to build a model of the users, where Artificial Neural Networks (ANN) is applied for a compilation of features user latent and then Rough Set Theory (RST) is used to reduct the attribute and classification.

### **1.8 Thesis Contents:**

The rest of this thesis is organized in six chapters as follows:

**Chapter 2** discusses the types and benefits of SNS as well as the foundations of modeling and benefits of SNS.

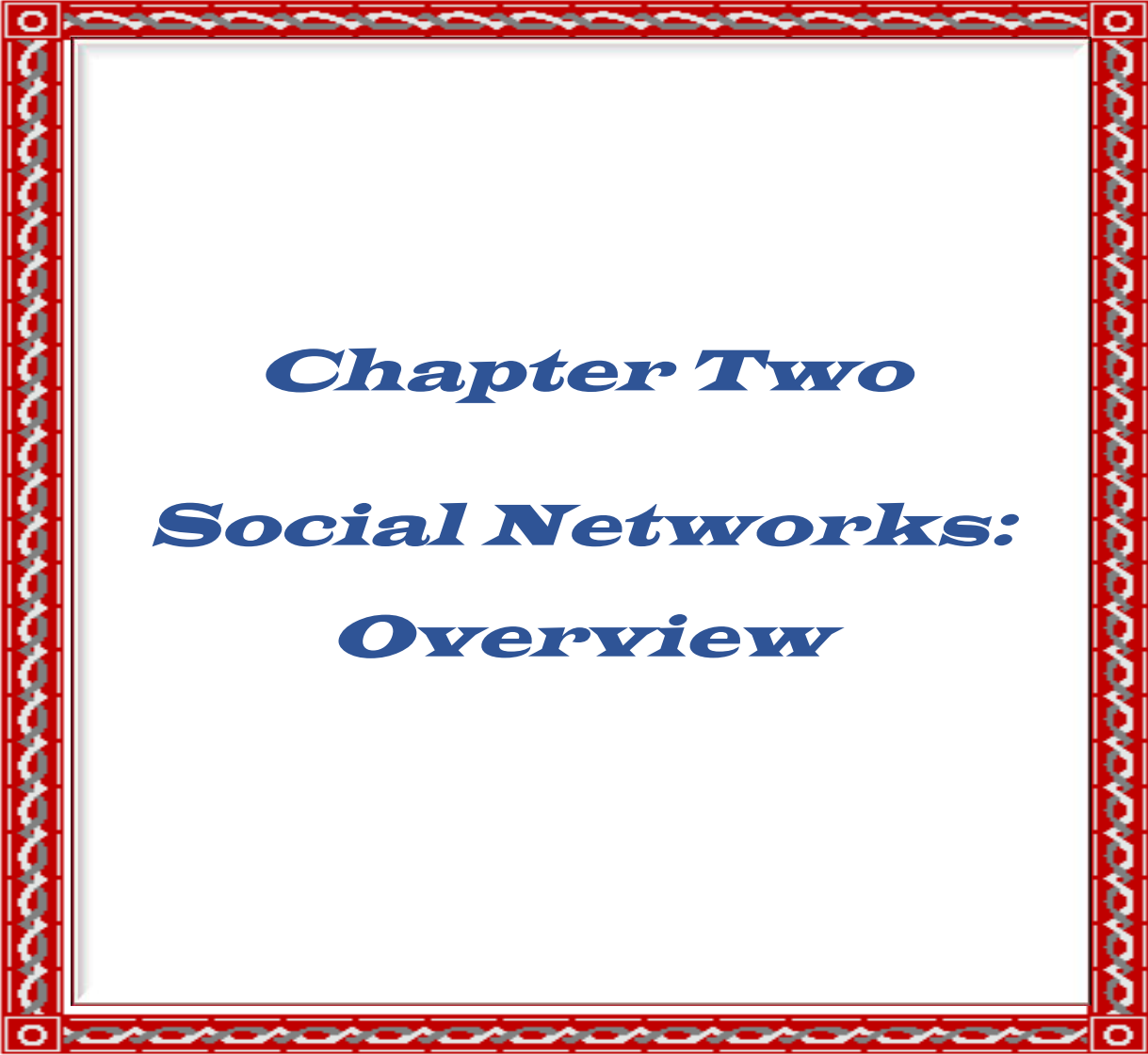
**Chapter 3** discusses ANNs with focusing on the clarification of self-regulation and RST.

**Chapter 4** presents the general structure of the proposed model and the steps that represented the used mechanism.

**Chapter 5** represents the practical experimentations, the results and discussion.

**Chapter 6** concludes the work and presents some future works.





***Chapter Two***

***Social Networks:***

***Overview***

## Chapter Two

### Social Networks: Overview

#### **2.7 Introduction:**

In the age of internet, SNS are speedy growth like Facebook, MySpace, LinkedIn, and Orkut, in recent years [6]. Users from all around the globe have subscribed on SNWS so that they could meet others with the same interests or experiences, and to share personal information with both friends and visitors.

SNS service is an internet service environment which assists in making and keeping social relations between individuals who are sharing the same interests and/or activities. SNS being unparalleled not only because it lays in giving individuals a chance to meet strangers, but also in helping users to elucidate and making their social networks apparent. Among the first SNWS, Friendster appeared in 2002, and made a huge success.

In 2004, many other sites started to appear after Friendster such as Myspace, LinkedIn, then Orkut, Facebook, YouTube, and Twitter. By 2007, Friendster had already lost space to Myspace. In 2007, MySpace won a gloss area that lost Friendster while Myspace lost space in 2010 to the fast growing Facebook. Facebook is similar to business model as Friendster and MySpace [7].

SNWS provide around 1.5 billion parts of information provided on Facebook, in which over 140 million tweets exist on Twitter, more than two million videos exist on YouTube, and Flickr nearly five million image [8].

## **2.2 Benefits of Social Networks:**

SNSs such as MySpace and Facebook have become widely used in both of the world of Internet commerce as well as in the lives of American youth. While these sites are popular with a younger demographic, there have been growing fears from religious leaders about the safety and contents of these sites [9].

SN features are depending on the number of connected and interacting people or groups of people, with patterns of connections and relations [10,11]. For example, of benefit SN the friendships between users, and business relations between corporations [12,13,14]. Online SN are virtual places that are introduced for a specific population; on such platforms people with similar interests are gathering to communicate, exchange contact details, build relations, and take and discuss ideas [13]. In the consumer-to-consumer space [15], SN is described as virtual communities of consumption, which feature characteristics like high consumer knowledge and companionship, and therefore influence consumer behavior. Some of the main uses of SNS are [9]:

- 1- Communicate with a new people.

2- Sharing media.

3- Culture about social events.

4- Entertainment.

5- Keeping friends and colleagues relationships.

### **2.3 Social Networks Categories:**

The following are the major categories of SN [16]:

#### **1. Social Connections:**

Social connections are one of the most advantages of SN, in which it focused of users interesting in connecting with family and friends. The following descriptions are the most commonly site of social connections internet:

- Facebook: It is considered the most common social media service. It is available users to build relations of friendship and exchange information with users on the Internet.
- Twitter: In Twitter, users can share ideas and retain up with others through services that enable them to refer and read. Each user can post 140-character as maximum length of each tweet.
- Google+ : Features in Google+ include "Posts" for posting status updates, and "Circles" for sharing information with different groups of people (like Facebook Groups).
- MySpace: It has provided a place for users to meet new friends and

keep in touch with people providing a venue for social connections related to movies, music, and games.

## **2. Multimedia Sharing:**

Using SN, it is simple to share photographs contents and videos online. Below are most popular sites for multimedia sharing [16].

- **Flicker:** This site provides an authoritative option for sharing and managing digital photographs online with others.
- **YouTube:** Sharing video content is main aim in this site.
- **Picasa:** This site is a product of Google in which it uses posting and sharing images. It submit integrated tagging and sharing with Google+.
- **Instagram:** It is a free online photo sharing and social network platform that was acquired by Facebook in 2012 [3].

## **3. Professional:**

Expert SNS are made for provide chances for career -related growth. Below are example of professional sites [16].

- **LinkedIn:** It is the largest professional online network. It gives a chance to participants for building relationships by making connections and joining pertinent groups.

- **Classroom 2.0:** It is a SN which is designed specifically to help teachers to share, connect and cooperation participation with profession-specific issues.
- **Nurse Connect:** It is a SN specifically planned for the aim of helping users in the nursing work communicate and connect with one another.

#### **4- Informational:**

To pursue inquiring about the daily problems of people, a user may perform a web search and find a countless number of blogs, websites, and forums that are filled with people that search for the same piece of information. The following include some examples of Informational sites [16].

- **Super Green Me:** A community in which individuals are interested in adopting green living practice that is capable of interacting online.
- **HGTV Discussion Forums:** Connect individuals interested in home design development by the HGTV message boards.

#### **5. Educational:**

SN makes it is possible for many scholars to cooperate with other students on theoretical projects in order to do research for the school, or to communicate with professors and teachers through blogs and classroom forums. The following are some examples.

- **The Student Room:** It is a UK-based site that contains subjects correlated to study and the communication between students and teachers
- **The Math Forum:** It is a large educational network planned for the aim of connecting students that are interested in math.
- **ePALS School Blog:** Through this international social network, users are devised to reside international connections for the aim of promoting world peace.

#### **6. Hobbies:**

Among the most widespread clarifications used by many people on the internet, performing a research on their preferred projects or topics of concern that have a correlation to their personal hobbies. These sites are concerned with finding individuals that have the same passion from all around the globe. It is considered to be the base of social networks work. A few examples include [16]:

- **Oh My Bloom:** A social media network specifically for gardening interest. It features groups, blogs, content, video, forums and more.
- **My Place at Scrapbook.com:** Planned specifically for scrapbooking supporters, Individual can building profiles and share information.
- **Sport Shouting:** An online purpose for sports supporter to voice their opinions and to connect with other enthusiasts.

## **7. Academic:**

These webs provide share research and display results achieved by colleagues. Some examples for the most popular communities for academics include [16]:

- **Academia.edu:** Users of this academic SN can follow research submitted by others and share their own research.
- **Connote a Collaborative Research:** These sites for scientists, researchers, and clinical practitioners to find, share and organize good information.

## **2.4 Social Networks Technologies:**

Before going to the social media techniques, we first need to clarify the difference between SNs and social media networks. SN means communications between people such as LinkedIn. Social media networks are sites for information broadcasting such as Flickr and YouTube. Some sites are both in the same type such as Facebook and Myspace. From here, it became a clear to us that social media techniques takes many types including blogs, a blog and social networks for businesses, social games, social networks, virtual world, and share photos [17].

Ixdegrees.com was the first online business that was created for real people using user names. Unlike instant messaging clients such as ICQ (Internet Chat Query) and AOL(America Online), AIM (An Instant



Messenger), or chat [18,19]. But these networks have lost interest quickly, which led to the emergence of networking sites meeting.

## **2-5 Social Networks Comparison:**

Comparing SN is different measures of network structure that we have encountered so far allowing us mainly to understand the structure of a single particular network. Table (2.1) explains different SNs based on four properties between six sites [16, 20]. There are general patterns in network structure that distinguish different entire classes or groups of networks, as shown in Table (2.2) that explains a comparison between a number of popular sites including description site, initially, communication, number of user under sites, and communication type [20 ,21 ,22].

## **2.6 User Modeling(UM):**

UM is the process of designing a user model. A user model comprises the assumptions of systems concerning all features of the users that are considered appropriate for altering the dialog activities of the system to the user [23]. The main aim of UM is to adapt and customize systems to specific needs of the user. UM is a set of informational structures planned to characterize one or more of the following elements [24]:

- (1) Representing presumptions concerning the aims, strategies, favorites, tasks and abilities, and the knowledge about one or more categories of users.

**Table 2.1 Abstract SN interaction paraigms and their underlying native count [20, 16]**

Likes Property	Shares Property	Comments Property	Views Property
Facebook Like	Facebook Share	Facebook Comments	Flickr Views
Google++1	Twitter native	Twitter manual RT,	Twitpic Views
Instagram Like	Retweet	@Replies Twitpic	YouTube Views
Flickr Favorite	Google+ Share	Comments	Moby Picture
Twitter Favorite		Google+ Comments	Views
YouTube Like		Instagram Comments	
YouTube Favorite		Flickr Comments	
		Moby Picture Comments	

- (2) The classification of a user in one or more of these subgroups.
- (3) Representation of relevant common characteristics of users pertaining to specific user subgroups (stereotypes).
- (4) The recording of user behavior.
- (5) The formation of assumptions about the user based on the interaction history.
- (6) The generalization of the interaction histories of many users into stereotypes.

The information that a user model has is directly convenient with the personalization content and presentation which is to be personalized. A user model is designed by a procedure of UM where data about a user that are not observed are inferred from data that are observed from that user [23,25].

**Table 2.2: A comparison between the number of popular sites [20,21,22].**

<b>Social Network</b>	<b>Description</b>	<b>Number of Users</b>	<b>B2B versus B2C</b>	<b>Communication</b>	<b>Initially</b>
<b>Facebook</b>	General	1,280,000,000	B2C	Posting to much-spam	2004
<b>Twitter</b>	Micro-blogging updates	645,750,000	B2B & B2C	Possibility of being overlooked	2006
<b>YouTube</b>	video sharing	1000,000,000	B2C	allow you to connections	2005
<b>Google+</b>	General	1,600,000,000	B2B	"circles" allow you to group connections	2011
<b>LinkedIn</b>	Business& professional networking	200,000,000	B2B	Less personalized	2003
<b>Instagram</b>	A photo and video sharing site.	300,000,000	B2C	group connections	2010

Where P2P : peer to peer      P2C: peer to companies

A user model determines the characteristics, requirements, favorites and aims of end users. It could differ in the basis of the domain or environment. Each domain model has a high level of concept which signifies the core of a problem, knowledge, or activity of the real world termed as Domain Model [26].

### **2.6.1 Modeling of SN:**

There are huge and massive developments in the field of Internet and modeling including the social programs which is a new model for the Web. This model provides the capability to share content on the Internet in many ways and locations, as well as the potential of stimulating the marking process easily (and sources of information, described by the word), and the publication, and the inclusion of new content and exchange of information (as Facebook) and provide comments and supports relations between the whole world and this makes the world as a small villages. It is known that these cooperative activities and social participation [27].

There are two more different types of relationships in user model. The first one is evidence links, which describe the results of user activity. They links learning objects for users and groups (because users interact with learning objects as members of some group). Evidence links are assigned time stamps and contain results of such interaction. Usually, the result is expressed in the form of a decimal value between 0 and 1, with 0 denoting an unsuccessful result and 1, the opposite.

The second special type of link, assertions about user knowledge – represents the user model’s probabilistic hypotheses about the user knowledge level of some knowledge components. Assertions are modeled with respect to the cognitive levels of Bloom’s Taxonomy [2].

Definition of modeling SN is to release relationships that a user has within the established network, in the direction of the center of that network, to help Community-driven. It is a tool for communicating and expressing the different shades of the society tool for SN and take experiences, cultures and Interests. The SN models and web services allow individuals to:-

1. Establish a public profile or semi-public as part of the system is bounded.
2. Express the list of other users who can phone them.
3. View and traverse other users' list of personal contacts within the system [28].



***Chapter Three***

***Background and***

***Related Works***

## Chapter Three

### Background and Related Works

#### **3.1 Introduction:**

Many recent researches these days are working on SN using different technologies to deal with various challenges concerning by data analysis. Some of these technologies include intelligent system of fuzzy logic, Neural Networks (NNs), genetic, and rough set. As each of these techniques has a distinct style, it allows collaborative work, ensuring production smarter system.

In the previous chapter, we reviewed the types of SN. In this chapter, the tools that will be used for the purpose of building the proposed model are presented and some previous studies of types of modeling used on social sites will be discussed.

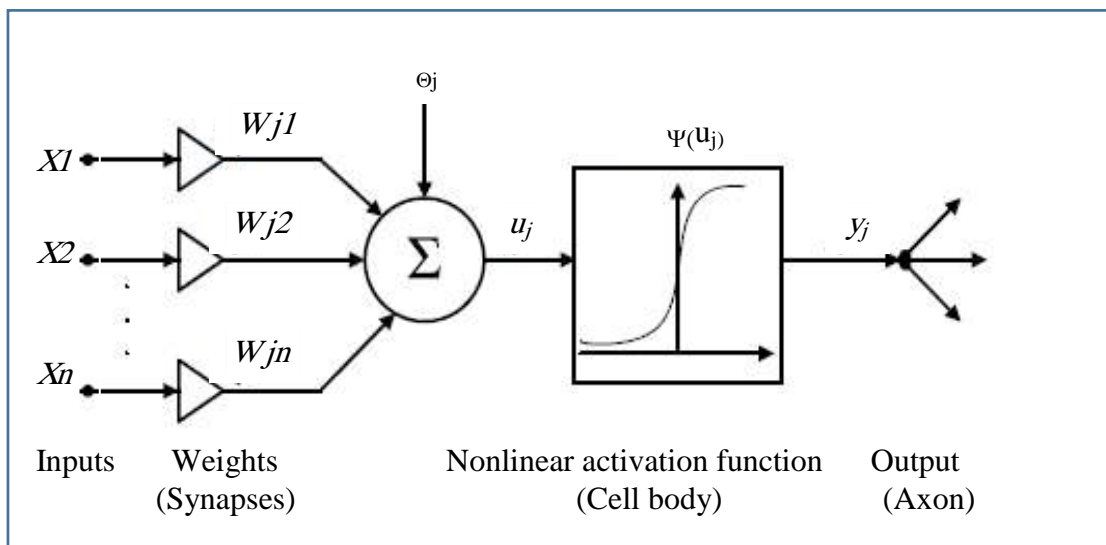
#### **3.2 Artificial Neural Networks(ANNs):**

The basic idea in constructing biological neural systems in neural networks are taken from human brain, where the neurons in the human brain is a dense network in which each neuron receives signals through synapses which control the effects of the signals on the neuron. These

synaptic connections play a strict role in the action of the brain, which takes an amazing skill to learn, save and generalize [29].

Artificial neuron includes three basic layers as shown in Figure (3.1) as follow:

1. Inputs layer to represent the inputs of network, where  $n$  represent number of input, known as  $x_i$ ,  $i = 1, \dots, n$ , for provide weights ( $W_{ji}$ ).
2. An aggregation function (hidden layer): to collect the weighted inputs to compute the input to the activation function.
3. Output layer: to represent the result of the neuron and consists of neurons that communicate to the external environment [29, 30].



**Figure 3.1: Structure of an Artificial Neuron [30]**



**The main learning types in ANNs are:**

- 1- Supervised Learning
  - › Pattern recognition
  - › Regression.
- 2- Unsupervised Learning
  - › Clustering
  - › Compression
  - › filtering.
- 3- Reinforcement Learning
  - › Games
  - › Coupled with dynamic programming [31].

Each neuron is linked to other layers means of interconnections or connected with a related weight. Layers can take different forms depending on the interconnections of neuronal link between them [32,33,34] as:

- ❖ Fully Connected: Each neuron in the first layer is linked to every neuron in the second layer.
- ❖ Partially Connected: Each neuron in the first layer does not have to be linked to all neurons on the second layer.
- ❖ Feed-Forward: In the first layer, the neurons send their outputs to the neurons of the second layer, but they do not extradite any input back from the neurons in the second layer.

- ❖ Feed-Backward: The outcome signals from the neurons on a layer directly feed the neurons in the same of previous layer.
- ❖ Bi-Directional: Feed (forward and backward)
- ❖ Hierarchical connections: For neural network with more than two layers, the neurons of lower communicate only with those of the next layer [35].

The behavior of an ANN depends on both the interconnections and the input-output function (transfer function) that is specified for the units [36].

ANN relies on upon both of the interconnections and the information yield capacity (exchange work) that is indicated for the units. Neural networks are trained rather than programmed [37]. Table (3.1) shows some of the positive and negative features of neural networks.

ANN has fabulous the ability of building the relationship between information yield mapping from a given dataset with no learning or suspicion about the measurable conveyance of information. This ability of gaining from information with no earlier learning makes neural systems especially reasonable for classification and regression errands in practical situations. In most financial and manufacturing applications, classification and regression constitute integral parts.

**Table 3.1: Positive and negative features of neural networks [31].**

Positive Features	Negative Features
<ul style="list-style-type: none"> <li>- ANN can help to solve the difficult problems that are difficult to solve with common rule.</li> <li>- Based programming. Possibility of learning.</li> <li>- ANN capacity finds the better capacity from a class of capacities for settling an undertaking by means of using a cost capacity. Multilayer Neural with at least one hidden layer are universal approximates i.e., they can be used to approximate any target function. Data can be classified rapidly.</li> </ul>	<ul style="list-style-type: none"> <li>- Too numerous weights can yield overfitting.</li> <li>- Required a major verity of training for true operation, training a NN is a period exhaustion process delicate to the presence of noise.</li> <li>- Not efficient where the goal is to characterize the physical process and the roles of the individual inputs.</li> </ul>

NNs are also inherently nonlinear which makes them more practical and accurate in modeling complex data patterns as opposed to many traditional methods which are linear [35,38]. There are two distinct types of learning in ANN. Table (3.2) shows the important methods used in each type of learning.

- A- **Supervised Learning:** The framework designer takes in the system a right answer, and decides the weight for information. The thought in regulated learning is that the system is over and over given facts about the different cases, alongside the expected outputs [36,37]. The system utilizes the learning technique to change the weights with a specific end goal to deliver the yields near what is prospective.
- B- **Unsupervised Learning:** The system gets just the inputs without unsurprising data about the output. In these systems, the frameworks learn to produce the pattern of what it has been exposed to [36,37].

**Table 3.2: Types of ANNs learning [38].**

Type learning	Important Method
<b>Supervised</b>	<ul style="list-style-type: none"> <li>o Multi-Layer Perceptron neural net (MLP) – Backpropagation</li> <li>o MLP Levenberg-Marquardt</li> <li>o Resilient Propagation</li> <li>o MLP Cascade Correlation neural net</li> <li>o Learning Vector Quantization (LVQ) neural net</li> </ul>
<b>Unsupervised</b>	<ul style="list-style-type: none"> <li>o Binary Adaptive Resonance Theory (ART)</li> <li>o Kohonen Self-organizing Map (SOM)</li> </ul>

## **Self-Organizing Map (SOM):**

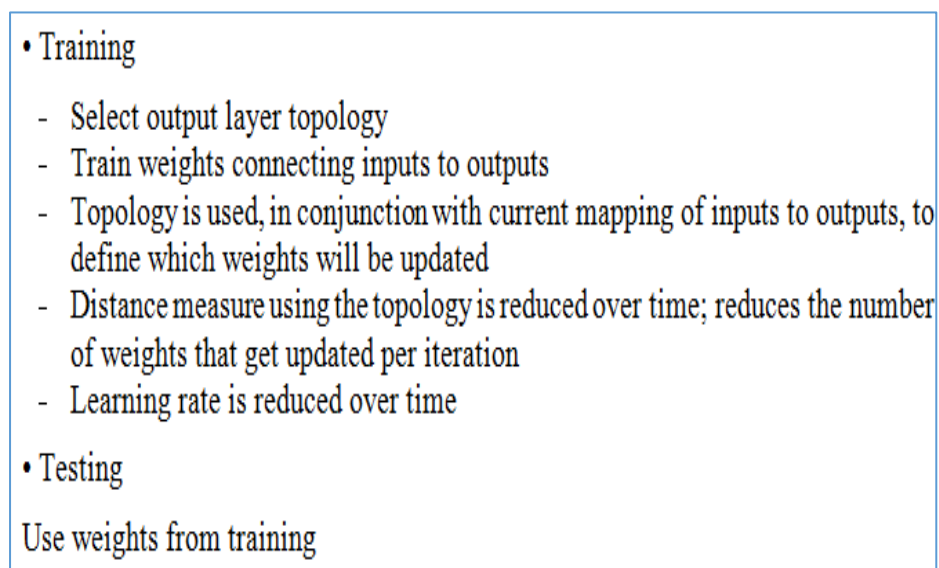
SOM, also known as Kohonen map is a type of the ANN and is based on unsupervised learning. The structure of SOMs is composed of two layers fully attached to each other: input layer and Kohonen layer [39]. Kohonen layer is also the layer where the map is formed that will ensure the observation of clustering in the data set. The number of neurons in the input layer is equal to the number of variables used. Each neuron in the input layer is connected to each neuron in Kohonen layer as feed forward [42].

SOM algorithm firstly assigns small random values to the connections between input layer and Kohonen layer. Then, the algorithm undergoes three essential processes: competition, cooperation, and adaptation [41].

- **Competition Process:** A random observation is selected from the data set. The criteria of finding the best match is based on the selection of the largest one of the scalar products that is equal to the mathematical maximization of Euclidian distance between  $w_j$  and  $x$ . For each input in the competition process, neurons in the model are in competition with each other.
- **Cooperation Process:** In the cooperation process, a topological neighborhood is determined, and the cooperating neurons will settle according to the topological neighborhood such that the winner neuron will be at the center. The winner neuron determines the topological value

of the neurons affected by competition; therefore, cooperation is ensured between nodes.

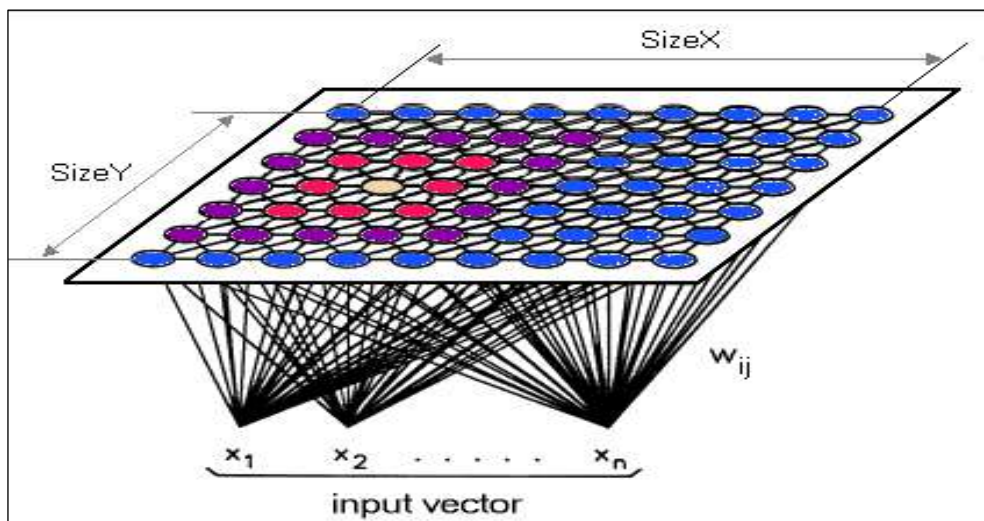
- **Adaptation (Synaptic Compatibility):** Neurons affected by competition arrange their synaptic weights. The overall SOM algorithm is as shown in Figure (3.2):



**Figure 3.2: Overall SOM algorithm [43].**

Neurons are connected to each other by neighborhood relation [42]. This neighborhood relation determines the structure or topology of the map. Figure (3.3) represents a simple SOM. Kohonen layer consisting of  $9 \times 7$  neurons in which the input vector is represented by letter  $x$  ( $n$  number of variables exist in the expression of  $x_n$ ). The weight vector is represented by  $w$  and it contains the weights outgoing from each variable to the each

neuron. The yellow-colored neuron represents the winner neuron and the surrounding neurons are its neighbors. The main advantages and disadvantages of SOM are discussed in Table (3.3).



**Figure 3.3: SOM input and output layers [42].**

### **3.3 Rough Set (Theoretical Aspects):**

Rough Set Theory (RST) is a mathematical structure for analyzing tabular data. An information system is a table with annotations (named objects) as rows, features (named attributes) as columns and discrete values as entries. The theory gets the data in terms of equivalence classes, i.e. sets of objects that are invisible (in distinguishable) with respect to the attributes. RS is a set of objects that cannot be uniquely signified by these

equivalence classes since the set only partly correspondences with at least one of them [45].

**Table 3.3: The main advantages and disadvantages of SOM [43,44]**

<b>Advantages</b>	<b>Disadvantages</b>
<ol style="list-style-type: none"><li>1. Extremely powerful mechanism for automatic mathematical characterization of acceptable system activity.</li><li>2. Simple and easy-to-understand algorithm that works.</li><li>3. Topological clustering.</li><li>4. Unsupervised algorithm that works with nonlinear data set.</li><li>5. The excellent capability to visualize high-dimensional data onto 1 or 2 dimensional space makes it unique especially for dimensionality reduction.</li></ol>	<ol style="list-style-type: none"><li>1. Difficult to determine what input weights to use.</li><li>2. Mapping can result in divided clusters.</li><li>3. Requires that nearby points behave similarly.</li></ol>

### **3.3.1 Attribute Reduction**

A large datasets may be contain complex due to the existence of conditional attributes, which may generate cost in the decision-making process, so they a process of remove these features a provide (reduce the complexity and reduce the cost) [47,48]. However, finding all reduces is NP-complete, it is generally not necessary to find all of them but one or



few of them are sufficient. RST provides helpful techniques to reduce immaterial and excessive attributes from a big dataset with a many of attributes [48,49,50]. In RST, the dependency degree (or approximation quality, classification quality) and the information entropy are two most popular attributes for reduction measures.

In this study, choosing best reduct is an important process. The selection depends on the optimality criterion associated with the attributes. If a cost function could be assigned to attributes, then the selection can be based on the collective minimum cost criteria. In this thesis, we adopt the following three criteria [51, 52]:

- ❖ **Cardinality:** It means the number of attributes of the reduct. The lesser the cardinality, the better the reduct.
- ❖ **Number of generated rules:** represent all rule generated, in this study the less number of rules is better.
- ❖ **Support:** It is a total number of correctly classified objects divided by total number of objects to be classified. Hence, the higher the support, the better the reduct [27].

Some of the research directions on RS are classification, dimensionality reduction, feature selection, RSs and noisy data, RS based clustering, RS, and relational datasets articular, missing value problems, RS and inductive reasoning, Boolean reasoning and approximate Boolean reasoning strategies as the basis for efficient heuristics for RS methods, In particular,

variable precision rough set model, and RS based approach based on neighborhood (uncertainty) functions and inclusion relation [53].

**A-Johnson's Algorithm (JA):**

This algorithm considers more heuristic. It has special kind of techniques. The main objective of this algorithm is choosing the attribute that occurs in clause. It is important to remember that the Johnson's algorithm starts with setting S, which is candidate for the current reduction for the empty set. The next step for implementing this algorithm is counting the appearance for each one of attributes into clause. The highest count of the attribute has added into S, and the whole clauses within f have removed from discernibility function. Then, the algorithm return S as a reduction [54]. For example, the step by step procedures for obtaining the reduction of  $k = (f1 \vee f3) \wedge (f2 \vee f5 \vee f4) \wedge (f3 \vee f5) \wedge (f1) \wedge (f2 \vee f5 \vee f4)$  are:

1. Count the appearance of the attributes,  $f1 = 2, f2 = 2, f3 = 2, f4 = 2, f5 = 3$ .
2.  $f5$  represents occurring attributes therefore, it will be added into S. All clauses that contained  $f5$  from  $k$  will be removed by the classifier. So,  $k = (f1 \vee f3) \wedge (f1)$  and  $S = \{f5\}$ .
3. Counting the appearance for each attribute in  $k$ . It finds that  $f1$  is usually representing occurring attribute. Then, the whole clauses of  $f1$  are removed. Thus,  $k$  becomes  $\emptyset$  and  $S = \{f5, f1\}$ .
4. End the algorithm as a result of  $k = \emptyset$ , and we obtain reduction  $f5 \wedge f1$

In concerning of this algorithm, the appearance attribute is greatly significant and that is not true all the times. However, this algorithm can find a solution close to the optimal [55].

**B- Genetic Algorithm (GA):**

GA is an evolutionary computing global search method simulating the biological evolution. It can tackle the optimization problems of complexity, nonlinear and notwithstanding including space. GA has the following accompanying elements:

- (1) It is a sort of clever calculation that self-learning, self-organization and adaptive ability.
- (2) It can manage the parameters of code set straightforwardly, as opposed to the parameters of the issue itself.
- (3) It uses fitness function to evaluate the intermediate individuals and guide the search direction in the process of search.
- (4) It is a kind of parallel algorithm. It is based on the population rather than an individual in each iteration to complete the search process in the solution space.
- (5) Its appearance is straightforward. Its essential thought is basic, and its operation mode and usage steps are standard [56].

So it is a good idea to use both of GA and RST to perform attributes reduction. This comparison can lead to the optimal or semi-optimal attribute reduction result [57].

**3.3.2. Naïve Bayesian (NB) Classification:**

Compared to other classifiers, NB has many advantages such as being simple, computationally efficient, requires relatively little data for training, do not have lot of parameters, and can deal with incomplete and noisy data. All these advantages made it an attractive classifier in many research areas. But, NB may suffer from some difficulties when applied to real life domains violating its main assumptions [57]. Table (3.4) presents the advantages and disadvantages of NB.

**Table 3.4: Advantages and Disadvantages of NB**

Advantages	Disadvantages
-Fast to train (single scan). -Fast to classify -Not sensitive to irrelevant features -Handles real and discrete data -Handles streaming data well	-Assumes independence of features

NB classifier is one of the most effective and efficient classification algorithms. It is based on applying Bayes' theorem with strong (naive) independence assumptions. Rough set Naive Bayes (RNB) addresses two main problems faced with NB classifier; firstly is its main assumption of

attributes' independency and secondly is the attributes' equality importance assumption. The presented approach tackled the former problem with rough set dependency measure while the latter is tackled with rough set significance measure.

The attribute weighting mechanism is based on the following two main propositions:

- (1) The more dependency detected with other attributes, the less weight is assigned to the corresponding attributes.
- (2) The higher significant attributes, the more chance it will improve the final classification, and therefore the more weight is assigned to the corresponding attributes.

Therefore, incorporating weighting of attributes with rough set measures could help in improving the performance of the produced model [58].

### **3.3.3. Cross-Validation (C-V):**

C-V denotes an untried testing process that is broadly utilized. C-V procedure is a method to get estimates that are more reliable and more mileage by a way of probable uncommon data. In the method of k-fold C-V is a random division of dataset into k disjoint blocks of objects, regularly have the same size. Then, there is training for the algorithm of data mining using k-1 blocks. The block which remains is utilized for testing the

algorithm performance. There is a repetition for this process for each of the  $k$  blocks where there is a record of a measure for all iterations [55].

The measure relies on the utilized task of data mining. For task of classification, there is a common use of the measure of classification. By the completion, the measures that have been recorded are averaged. Through this procedure, there is a guarantee for every object to be once in the test set and  $k-1$  times in the size of training. Choosing  $k=10$  or any other size is common relying on the original dataset size.

A great variant of selecting  $k$  is choosing  $k = |U|$ , i.e., letting every test set consists of one example. This is known as leave-one-out C-V, and, even though possibly enormously computer intensive can be instinctively fair as it best mimics the actual training set size. The chosen objects for training are not required to be head-to-head [55].

### **3.4. Related works:**

Stages of construction and development of models for social networks have the attention of many researchers in the scientific field by using different algorithms and technology. RST is a mathematical tool that can be used to deal with inexactness, uncertainty, and vagueness that may appear in datasets. It can be seen as improvement for the set theory for building intelligent system that based on incomplete data [59]. This section is dedicated to address some of previous work that based on RST.

In [60], RST is used to identify the most influencing features that identify the e-mail usage habits. Also RST is used to discover the decision rules from real dataset pertaining 266 academic staff. Each record contains 13 conditions and one decision features. The discovered rules were used later for classification taking into account that the dataset contains uncertain information with accuracy 96.3%.

In [61], study of news articles in twitter, and build a multi-dimensional component space got from properties of an article and assess the adequacy of these elements to serve as indicators of online prominence. and examine both regression and classification algorithms and regardless of randomness in human conduct, it is anticipate scopes of ubiquity on twitter with a general 84% exactness. In [62], a hybrid technique is proposed to classify web object either to cash or not to increase the access speed in mobile environment. The proposed technique is based on ANN and Particle Swarm Optimization. Also, it generates rules from log data using RST.

In [63], a computationally efficient method based on rough set is used for ranking the documents to be used in content based retrieval. In [64], the rough set is to classify and cluster in social network. The paper provides a critique that covers the limitation of rough set and suggested that the use of Covering Based Rough Set would be a better alternative.

In [65], an effective method is proposed to enrich tweets representation based web search engine and RST by adding synonyms for

the original terms. The proposed method integrated and tested for Arabic tweets categorization using NB and support vector machine classifiers. The results showed that the performance of the categorization system is greatly enhanced.

In [66], an enhanced algorithm is proposed based on rough set and K-means clustering to find overlapping communities in social networks. The rough set is used to handle several disadvantages that caused by K-means clustering method such as determining the value of K and the relations between the community node and the community.

In [67], RST is applied to predict links over the Facebook based on hemophilic features. Other classifiers are used and compared to the rough set classifier.

In [68] Present a study of 17 social applications, comprehensive user model especially fitted to the needs of the Social Web, furthermore, we present a WordNet for the user modeling domain as part of the user model to support user model aggregation.

In [3] Explored the factors that drive students to use online social networks (Facebook), examine the relative impact of social influence, social presence, five key values from the uses and gratification paradigm, an empirical study of Facebook users (n=182).

In [69] Employ and share information by studying a common user pool that use six OSNs – Flickr, Google+, Instagram, Tumblr, Twitter, and



YouTube. aim of discovering behavioral patterns in their multiple network use. the limitation of this work, it have only studied largely networks and their user's public sharing activities.

In [70] Scraped data from consenting Facebook users demographic and psychological profiles, collected data by using survey, 1327 participants, 12 user features, uses (K-means, analysis), present five clusters of users with common observed online behaviors.

### **3.5. Summary:**

In this chapter we reviewed the tools that will be applied in the construction of our proposed model such as neural networks (self-organization map) for the work of the clusters of data, and RST that represents a mathematical model to deal with datasets to be used in reducing the number of features of the data that define the properties that will be relied upon in the proposed model, and finally the use of classification algorithm NB for the classification stage and to give a clear indication of the accuracy of the proposed model.



***Chapter Four***

***Proposed Model***

***Structure***

## Chapter Four

### The Proposed Model Structure

#### **4.1 Introduction:**

User profiles component is one of the main components in social networks sites. These profiles provide a great variety of user-provided data. However, studying and analyzing the effects of user profiles on their online connections still at the infancy stage and there is a big room for research to be done at this field. We believe that a lot of benefits can be gained from analyzing the user profiles and figuring out the relationship between the user profile and the interests of users and thus we need to know what type of included information matters, and how do elements in a profile affecting the outcomes of using an online SN?

To test the role of online interactions of profile, Facebook.com is chosen in this work as the biggest online social network. Facebook allows a user to create full profiles to describe himself and then able to build obvious connections with others. On the other hand, this largeness of data comes with its demerits and we need to focus only on the most effective attributes that have a relation with the users' behavior and filtering the irrelevant ones. For this purpose, RS is used in this work.

## **4.2 Scope and Solution Plan:**

Obtaining the user's concerns within the daily activity on the internet is a difficult task because of the complexity of the data, the fact that the user's activity may vary from one to another location, and transmitted large amount of data in the social networks, which adds considerable complexity in the process of building a model for users of social networking sites. These difficulties encouraged us in this work to use NNs with RST to get the lowest possible number of attribute, which keeps the dependency of the subject.

## **4.3 The Proposed Methodology:**

The observation and analysis of user behavior on the web is usually a preliminary stage to infer information about user interests and preferences. Adaptation and personalization is an important part of the modeling to learn about user behavior. But the complex part is the user behavior that varies depending on the type of network that can be cared the same users on Twitter to specific event, in YouTube to music or action film, and in Wikipedia for literature subject. Also, user interests are different within a certain period of time that may be radically from another time period. So in this thesis, a common general model SN user is introduced and implemented.

Figure (4.1) illustrates the general workflow of users modeling that deals with the user in the web taking into account the user's activity on other sites.

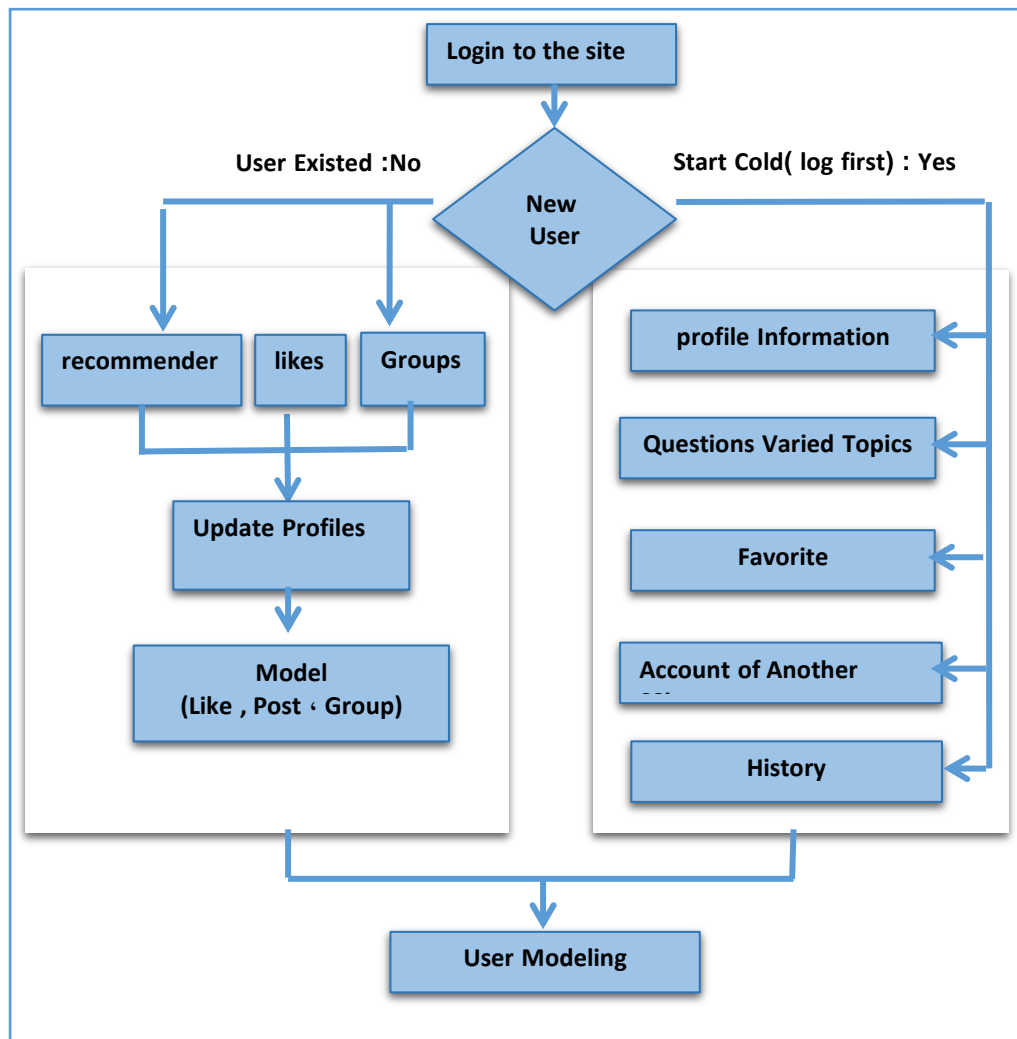


Figure 4.1: The general workflow of users modelling in social networks

#### **4.4 The Proposed System:**

User information in Facebook can be used to enhance the social network activities and interests such as clips and movies daily and concerns, as well as the common interests of classmates and work or organizations in addition to users like other pages on the network [71]. The goal is to build a general model for the user social networks using the least number of properties which can be more effective in identifying user's interests depending on the characteristics in which they can reach the best predicted and the most accurate user recommendations of the social network users.

The main framework of the proposed modeling system is shown in Figure (4.2) in which it is composed of eight stages (data collection, data preprocessing, clustering using SOM, reduction, classification, cross-validation, model, and evaluation) which are discussed below.

##### **4.4.1 Data Collection:**

The used dataset was collected from 680 Facebook users through making information form (Appendix 1) family, friends circle, and college students from the third and fourth years at the Faculty of Computers and Information, Mansoura University, Egypt.

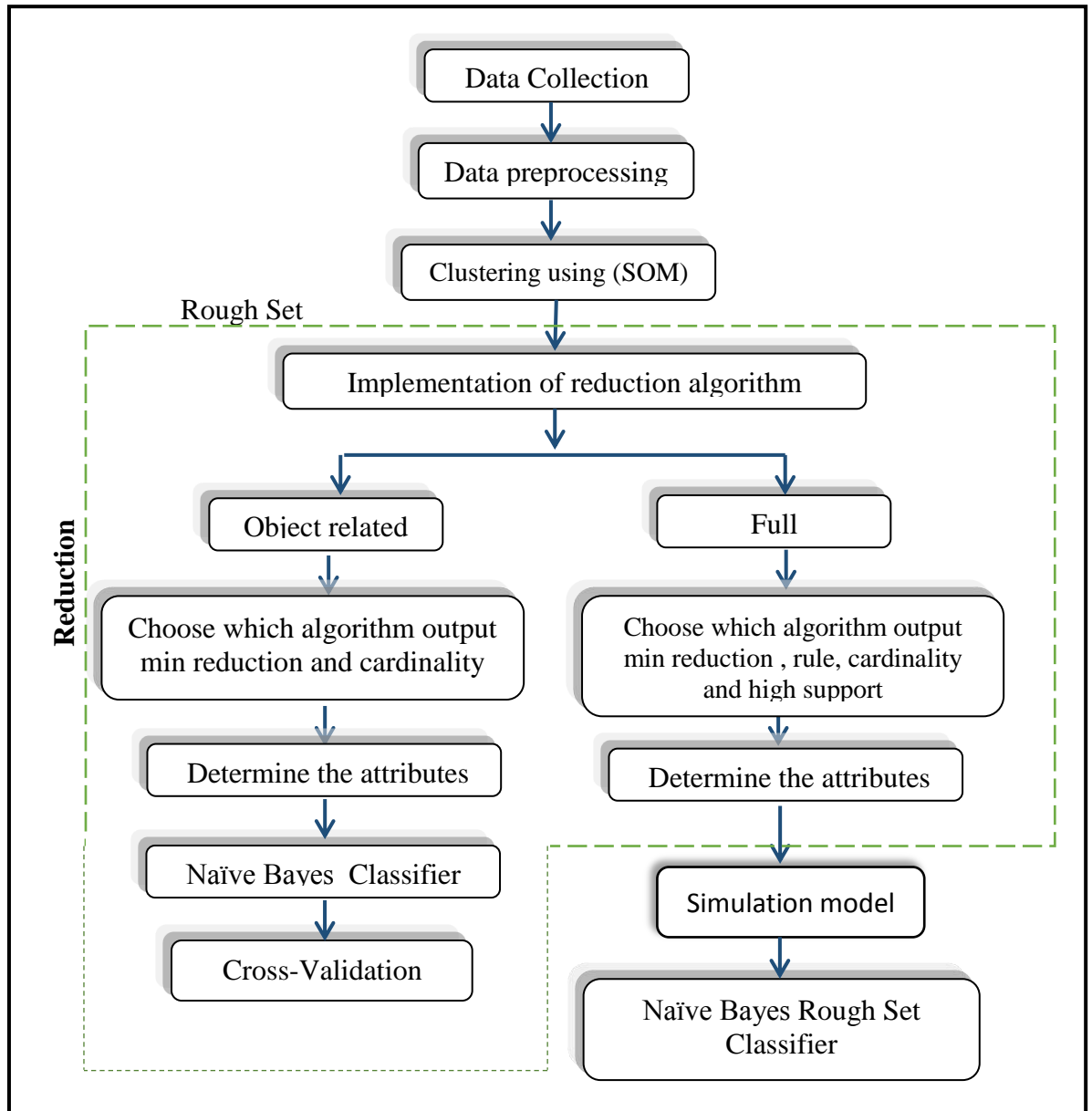


Figure 4.2: The basic structure of the framework

#### 4.4.2 Data Preprocessing:

The preprocessing stage steps are divided into removing blank and uncorrected forms, translating character data to digital value, storing in excel file, and completion (in Rosetta) as shown in Figure (4.3).

To predict the user behavior from dataset, we had selected (16) attributes (age, gender, education, post average, average Facebook time, average internet, time, friends number, friend like sharing, find new friends, interest play games in Facebook, proposal page, interaction, research, like, preference, post). Figure (4.4 A) shows the gender distribution of users in the dataset. (4.4 B) shows the educational distribution of users in the dataset. Table (4.1) describes the description statistic values of each attribute used in this work.

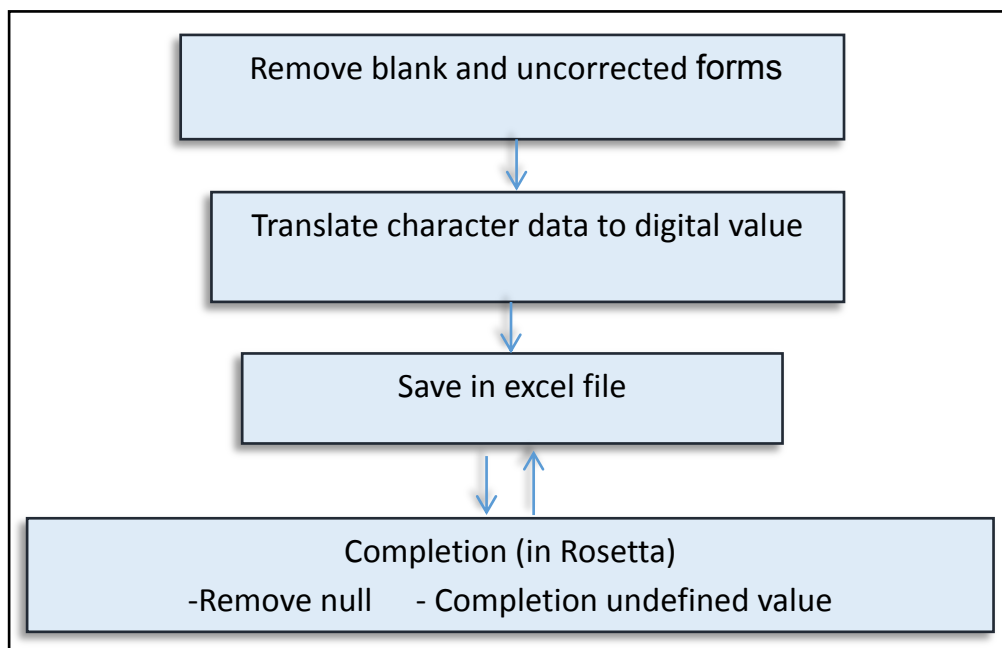
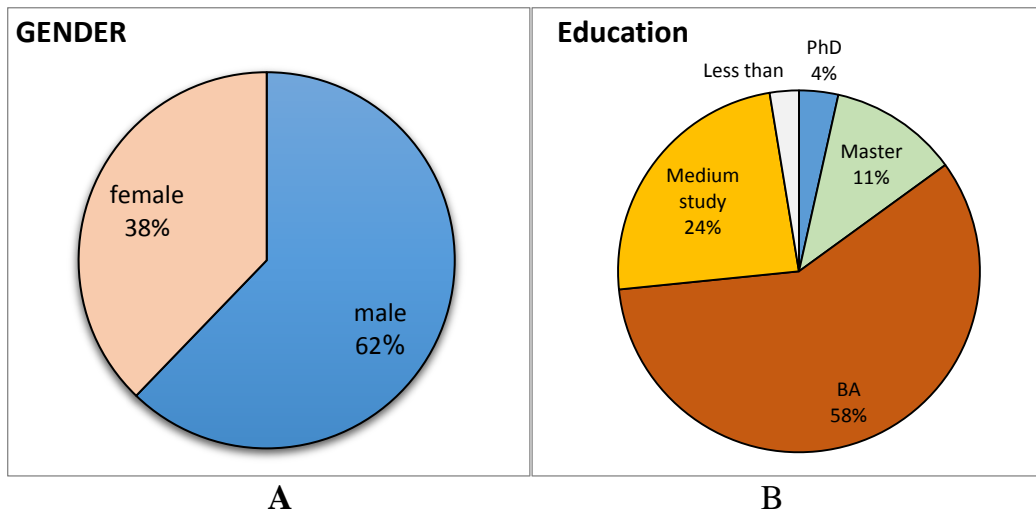


Figure 4.3: The preprocessing data stage





**Figure 4.4: A: The gender distribution of users in the dataset  
B: The educational distribution of users in the dataset**

**Table 4.1: Demographic information profiles**

Attribute	Percentage	Value
Age		13 – 52 years, Mean 39 years
Gender	62.2 % 37.8 %	Male Female
Marital status	64.8 % 35.2 %	Single Married
Nationality	66.5 % 32.7 % 0.8 %	Iraq Egypt Other
Education	3.5 % 11.5 % 58.4 % 24 % 2.6 %	PhD Master BA Medium study Less than

As a result of data collected using information form and after preprocessing, we got a set of characteristics that had been specified as follows: -

1. **Age** (13-52).
2. **Gender** (male, female).
3. **Education:** Including PhD, Master, BA, Medium study, Less than.
4. **The average daily use of the network:** It represents the daily average time spent by the user to browse the various sites, except Facebook.
5. **The average daily use of the Facebook:** Represents the daily time spent browsing Facebook user rate.
6. **The average posts daily in Facebook:** It represents the average number of daily publications for the user within the Facebook site.
7. **Average number of friends:** It represents the average number of friends to the user within the Facebook site.
8. **Interaction:** (active, enactive).
9. **Proposal page:** How much proposal page is helpful to the user?
10. **Facebook games:** It measures user interests of the communion game in Facebook site.
11. **Search for new friends:** The user attention for searching new friends in the verity site.
12. **Share like friend:** It measures user attention to new or unknown publication.

13. **Topics preference:** Some of the main topics that users used in social sites (blog, world news, local news, sciences, technology, business, sport, industry, lifestyle, universities, jobs, music, celebrities community, art, health, fine stuff, game, video, event, entertainment, travel, s hopping, religion, programming, commentators, literature, history, geography, design, family and child and general culture).
14. **Likes topics:** Same topics preference subject.
15. **Posting topics:** Same topics preference subject.
16. **Search:** Same topics preference subject.

#### **4.4.3 Clustering:**

The following stage is to use SOM to cluster the selected attributes. It uses the 16 attributes as one input vector for each user and the maps input matrix of all input vectors to two dimension to cluster them. The first dimension is the input vector itself and the other is the cluster type.

The cluster type is number from 1 to 6 that represent the classified (scientific, policy, religion, sporty, general culture, education). That represent a new attribute for each user based on this new matrix, rough set algorithms is applied.

#### **4.4.4 Reduction:**

In this stage, attributes in clustering data is reduced through applying different rough sets (GA, JA, Holte1R, Manual, SVA Genetic, Johnson

Reducer, Holte1R Reducer, and Manual Reducer) algorithms to select the best reduction based on reduction, rule, cardinality numbers, and support.

**Table 4.2: Set of attributes in the used data set.**

<b>Features</b>	<b>Min</b>	<b>Max</b>	<b>Mean</b>	<b>Std. dev.</b>	<b>Median</b>
Age	13	52	33.9	10.903	35
Gender	0	1	0.377	0.485	0
Education	1	5	3.092	0.760	3
Post average	1	5	1.020	0.209	1
Average Facebook time	1	5	1.092	0.408	1
Average internet time	1	5	1.1926	0.472	1
Friends number	1	5	2.282	1.706	2
Friend like sharing	0	5	3.129	1.128	3
Find new friend	0	5	3.132	1.304	3
Interest play game in Facebook	0	5	3.952	1.038	4
Proposal page	0	5	3.230	1.203	3
Interaction	0	1	0.217	0.334	0
Search	1	33	62.208	72.685	20
Like	1	33	74.901	74.147	41
preference	1	33	62.525	70.971	18
Post	1	33	18.636	34.231	1

Each one of these algorithms has two options: full and object related. Full results are a set with less number of properties that can determine functional dependencies. Object related results are a set of decision rules or

general patterns through minimal attribute subsets that discern on a per object basis.

#### **4.4.5 Classification:**

In this stage, NB classifier is applied on the reduced data of previous step. The classification accuracy measure used in this experiment is computed using the confusion matrix shown in Figure (4.5). This matrix contains information about actual and predicted classifications done by a classification algorithm. The accuracy is the proportion of the total number of predictions that were correct where TP represents True Positive, FP is the false positive, TN is the True Negative, and FN is the False Negative.

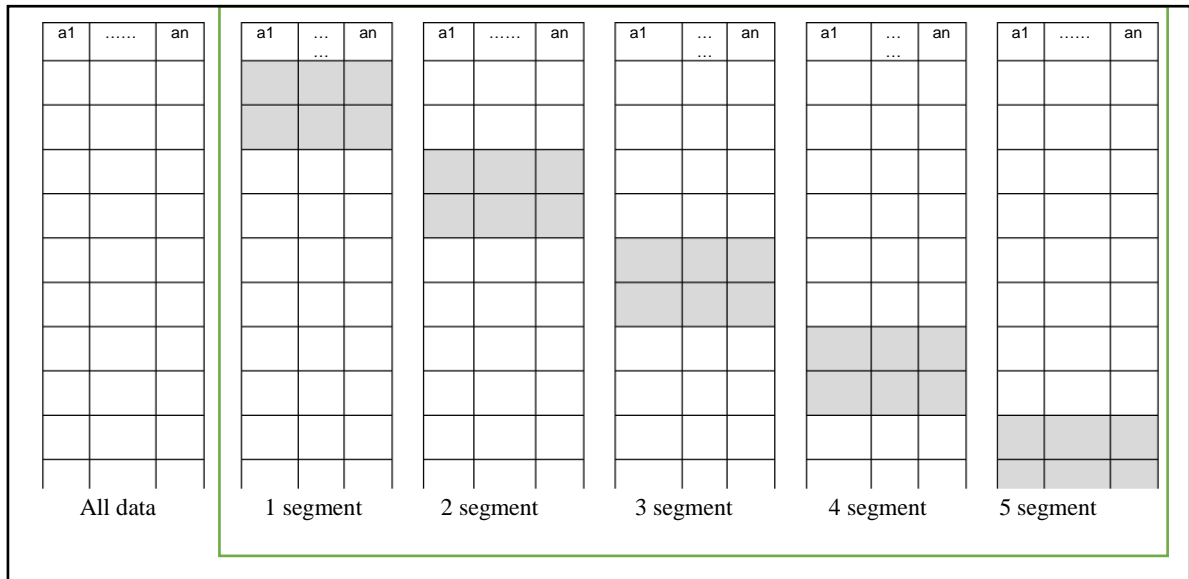
		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

**Figure 4.5: General confusion matrix [72]**

#### **4.4.6 Cross-Validation:**

In this stage, cross-validation is applied for estimating the performance of a classifier to reduction attributes matrix (all objects, reduction attribute) using rules of reduction algorithm and classifier reduction annotation. The

mechanism of cross-validation is used for obtaining utilize observations. Figure (4.6) represents a graphical unit of the K-fold cross validation. Each one of columns indicates to single iteration of testing and training.



**Figure 4.6: Classifier evaluation using 6-Fold Cross-Validation**

In concerning k-fold method, dataset is randomly classified into k subsets. In the first iteration, the examination set is used with the subset, and the remaining four are used to derive the rules. In the second iteration, subset number 2 is used as the examination set and so on. This process is iterates for k-1 times. The result of this stage is computing sensitive, specificity and accuracy for each iteration using Equation (2), Equation (3) and Equation (4) respectively. Sensitivity is the percentage of predicting the number of correctly classified as true positive while specificity is the percentage of predicting the number of true negatives by the classifier.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} * 100\% \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} * 100\% \quad (4)$$

### 4.4.7 Simulation Model:

This stage used the attributes of the resulting from reduction algorithms, usage in building a simulation model to Facebook site. By using the programming language C# and SQL, where used (32) image of different topics. The simulation model include two page, the first is register page of profile user in simulation model and the second get proposal image in simulation model.

### 4.5 Summary:

In this chapter, the proposed structure model is discussed. In the next chapter, we extract the results by processing data by using SOM, and RST as well as the results of building a miniature model to simulate Facebook environment with analyzing this results.



***Chapter Five***

***Experimental***

***Results and***

***Discussion***



## Chapter Five

### Experimental Results and Discussion

#### **5.1. Experimental Results:**

An adaptive system capability for the creation of environments is mainly determined by the collected and correctness of the stored data in every user model. Among the problems that confront user modeling is the collecting data of more than one site for the creation of user models, the noise inside that data, and the inevitability to capture the vague human conduct nature. The methods of completion data and machine learning are capable of handling large data and processing for uncertainty.

Such features enable these methods to be fit for user models' automatic generation which mimic the decision making of human. The data collected, it enables us to study the user attributes (favorites, profile and daily activities) in Facebook site, for getting the lowest number of attributes which are more effective and can give us a clear idea of the interest, and recommendations of the SN users.

#### **5.2 SOM Algorithm:**

The main benefit of using SOM is to add a new feature to data property clustering. It combines similar properties within a single type.

An example of using SOM algorithm is shown Figure (5.1) where the training input is 16 attribute and the output is 6 clusters<sup>1</sup>. Figure (5.2) shows the results of executing SOM algorithm where the number of object belongs in cluster one are (142), cluster two are (139), cluster three are (83), cluster four are (111), cluster five are (117), and cluster six are (89).

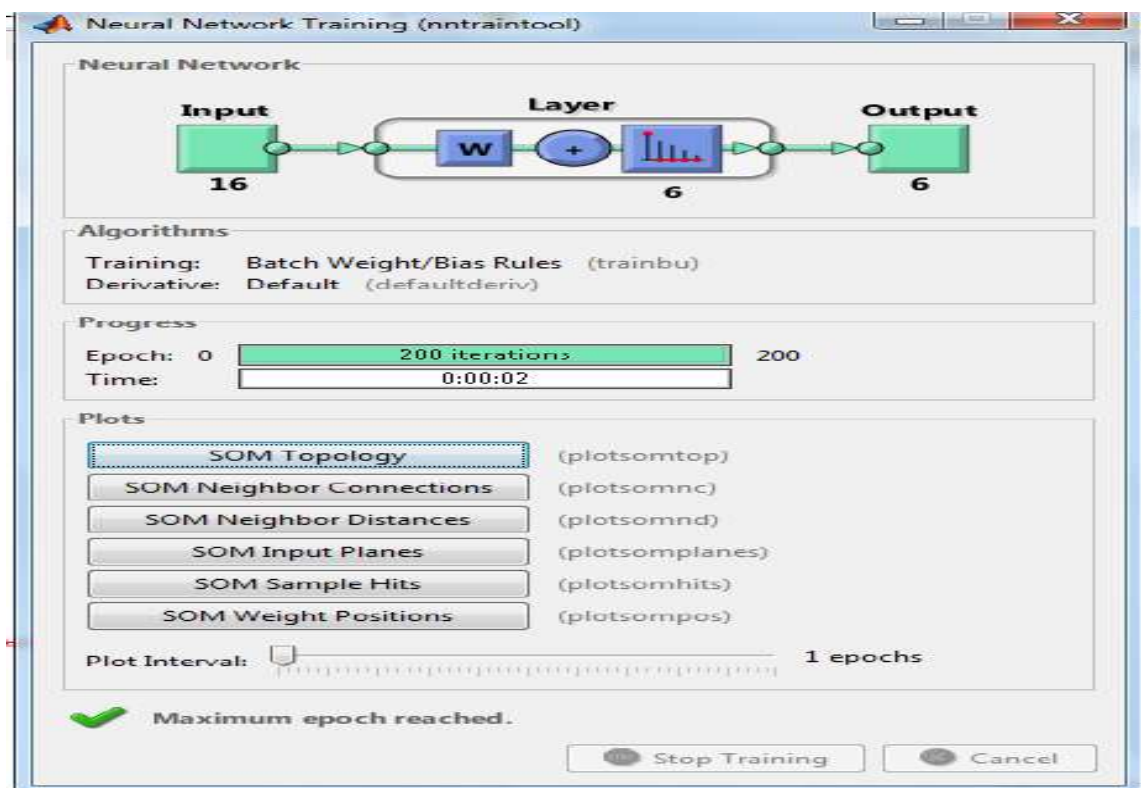


Figure 5.1: Example of using SOM algorithm

<sup>1</sup> According to research conducted by the Forrester Research, where users of social networking sites classifier to six types (Creators, Critic, Collectors, Joiners, Spectators, inactive)

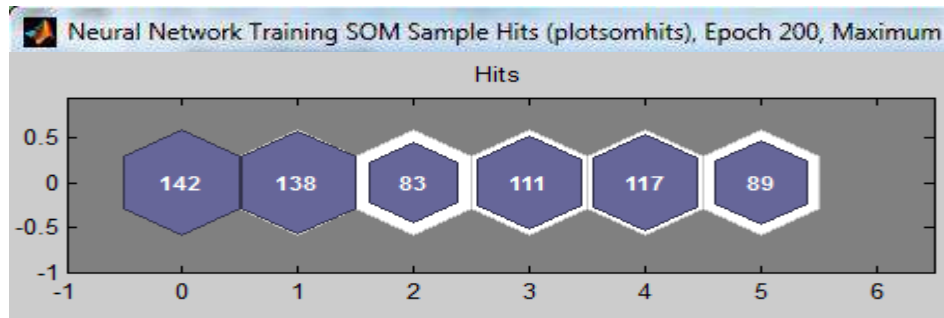


Figure 5.2: Results of executing SOM

### 5.3 Rough Set:

The second stage is to apply RS with reduction algorithm (GA, JA, Holte1R, Manual, SINE-VNTR-*Alus* (SVA) Genetic, Johnson Reducer, Holte1R Reducer, and Manual Reducer), each at a time. This stage has two parts as follows.

#### A- Part one (Reduction and Cross-Validation):

In this part, reduction algorithm is applied, and the object related option is choosed. This option outputs a set of decision rules or general patterns through minimal attribute subsets. Table (5.1) shows the result of executing RS reduction algorithm. From this table, JAs gives a better result because it gives less number of reducts, less number of rules, less number of cardinality, and high number of support. As important result, JA reduces the cardinality number to 4 attributes that represent (age, like, share-like, search new friend).

**Table (5.1): Results of the reduction algorithms**

Reduction Algorithm	No. reduct	No. rule	Cardinality	Support
<b>Genetic algorithm</b>	14306	5654	1,2,3,4,5	100
<b>Johnson algorithm</b>	680	444	1,2,3,4	100
<b>Holte1R</b>	14	934	1	1
<b>Manual</b>	1	678	13	0
<b>SVA Genetic</b>	14357	7464	1,2,3,4,5	100

In Second stage, we applied NB on the reduced attributed and the output confusion matrix in which the results of this stage is shown in Table (5.2).

**Table 5.2: Result of executing NB**

No name								
	Predicted							
	1	2	3	4	5	6		
Actual	1	109	0	0	0	0	1.0	
	2	0	155	0	0	0	1.0	
	3	0	0	94	0	0	1.0	
	4	0	0	2	114	0	0.982759	
	5	0	0	0	0	117	1.0	
	6	0	0	0	0	0	89	1.0
		1.0	1.0	0.979167	1.0	1.0	1.0	0.997059

From Figure (5.3), the accuracy is (99.07%), the sensitivity is (98.27%) and the specificity is (97.91%) of the computed RS on basis of the confusion matrix.

Next stage is applying cross-validation through dividing the data in to 10 sets in which each set includes two parts: one for training and the other for testing. Table (5.3) through Table (5.11) show the percentage of a cross-validation for all dataset using k=10 on different training and testing percentages.

**Table 5.3: Executing C-V with 90% training and 10% testing**

	1	2	3	4	5	6	Undefined	
1	9	0	0	0	0	0	0	100.0%
2	0	14	0	0	0	0	1	93.33333%
3	0	0	9	0	0	0	1	90.0%
4	0	0	0	13	0	0	0	100.0%
5	0	0	0	0	13	0	1	92.85714%
6	0	0	0	0	0	7	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	95.58823%

**Table 5.4: Executing C-V with 80% training and 20% testing**

	1	2	3	4	5	6	Undefined	
1	13	0	0	0	0	0	0	100.0%
2	0	9	0	0	0	0	0	100.0%
3	0	0	11	0	0	0	1	91.66667%
4	0	0	0	13	0	0	0	100.0%
5	0	0	0	0	10	0	0	100.0%
6	0	0	0	0	0	11	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	98.52941%

**Table 5.5: Executing C-V with 70% training and 30% testing**

	1	2	3	4	5	6	Undefined	
1	12	0	0	0	0	0	0	100.0%
2	0	16	0	0	0	0	0	100.0%
3	0	0	6	0	0	0	3	66.66667%
4	0	0	0	10	0	0	0	100.0%
5	0	0	0	0	8	0	0	100.0%
6	0	0	0	0	0	13	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	95.58823%

**Table 5.6: Executing C-V with 60% training and 40% testing**

	1	2	3	4	5	6	Undefined	
1	7	0	0	0	0	0	0	100.0%
2	0	17	0	0	0	0	1	94.44444%
3	0	0	11	0	0	0	3	78.57142%
4	0	0	0	5	0	0	0	100.0%
5	0	0	0	0	13	0	0	100.0%
6	0	0	0	0	0	11	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	94.11764%

**Table 5.7: Executing C-V with 50% training and 50% testing**

	1	2	3	4	5	6	Undefined	
1	14	0	0	0	0	0	0	100.0%
2	0	13	0	0	0	0	0	100.0%
3	0	0	4	0	0	0	1	80.0%
4	0	0	0	17	0	0	1	94.44444%
5	0	0	0	0	11	0	0	100.0%
6	0	0	0	0	0	7	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	97.05882%

**Table 5.8: Executing C-V with 40% training and 60% testing**

	1	2	3	4	5	6	Undefined	
1	10	0	0	0	0	0	0	100.0%
2	0	19	0	0	0	0	0	100.0%
3	0	0	10	0	0	0	3	76.92308%
4	0	0	0	9	0	0	1	90.0%
5	0	0	0	0	8	0	0	100.0%
6	0	0	0	0	0	8	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	94.11764%

**Table 5.9: Executing C-V with 30% training and 70% testing**

	1	2	3	4	5	6	Undefined	
1	15	0	0	0	0	0	0	100.0%
2	0	12	0	0	0	0	0	100.0%
3	0	0	7	0	0	0	0	100.0%
4	0	0	0	12	0	0	1	92.30769%
5	0	0	0	0	17	0	0	100.0%
6	0	0	0	0	0	4	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	98.52941%

**Table 5.10: Executing C-V with 20% training and 80% testing**

	1	2	3	4	5	6	Undefined	
1	8	0	0	0	0	0	0	100.0%
2	0	17	0	0	0	0	0	100.0%
3	0	0	6	0	0	0	2	75.0%
4	0	0	0	10	0	0	0	100.0%
5	0	0	0	0	14	0	0	100.0%
6	0	0	0	0	0	11	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	97.05882%

**Table 5.11: Executing C-V with 10% training and 90% testing**

	1	2	3	4	5	6	Undefined	
1	14	0	0	0	0	0	0	100.0%
2	0	16	0	0	0	0	1	94.11764%
3	0	0	7	0	0	0	2	77.77777%
4	0	0	0	13	0	0	0	100.0%
5	0	0	0	0	5	0	0	100.0%
6	0	0	0	0	0	10	0	100.0%
Undefined	0	0	0	0	0	0	0	Undefined
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	0.0%	95.58823%

**Results:**

The highest accuracy classification is 98.52% was obtained from two partitions (80%-20% and 30%-70%) training-test partitions. An accuracy classification of 95.58% was obtained from the 90%-10% training-test partition, 98.52% from the 80%-20% training-test partition, 95.58% from the 70%-30% training-test partition, 94.11% classification accuracy was obtained from the 60%-40% training-test partition, and 97.05% classification accuracy was obtained from the 50%-50% training-test partition.

An accuracy classification of 94.11% was obtained from the 40%-60% training-test partition, 98.52% from the 30%-70% training-test partition, 97.05% from the 20%-80% training-test partition, 95.58% classification accuracy was obtained from the 10%-90% training-test partition. Table (5.12) shows the values of classification Sensitivity, Specificity, and Accuracies values of the test data.



**Table 5.12: Sensitivity, Specificity, and Accuracies (%) results for k=10 Different Participations.**

<b>Training-Testing</b>	<b>Accuracy</b>	<b>Sensitivity</b>	<b>Specificity</b>
90-10%	95.58	0.956	0.04
80-20%	98.52	0.985	0.02
70-30%	95.58	0.956	0.04
60-40%	94.11	0.941	0.06
50-50%	97.05	0.971	0.03
40-60%	94.11	0.941	0.06
30-70%	98.52	0.985	0.02
20-80%	97.05	0.971	0.03
10-90%	95.58	0.956	0.04

The values of Accuracy.Mean, Accuracy.Median, Accuracy.Std Dev, Accuracy. Minimum, Accuracy. Maximum, are presented in Table (5.13).

**Table 5.13: Cross Validation Result for all Partition**

Accuracy.Mean	0.963235
Accuracy.Median	0.963235
Accuracy.StdDev	0.015884
Accuracy.Minimum	0.9481176
Accuracy.Maximum	0.985294

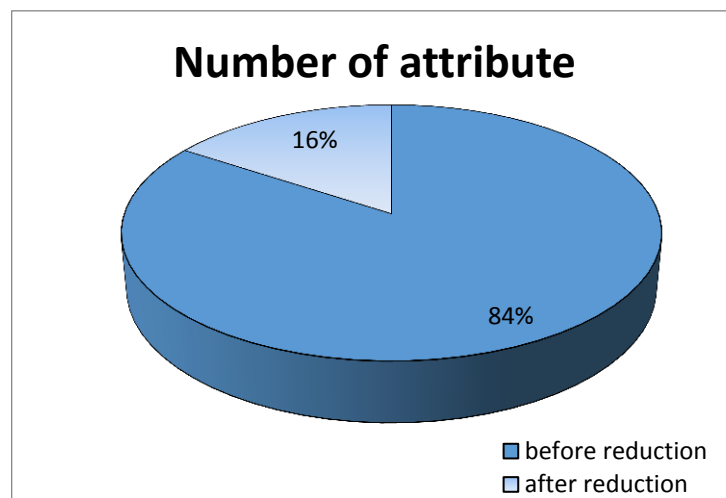
**B- Second part (Reduction and Building Module)**

In this part, reduction algorithm is applied, and the full option (intended output a set contains less number of properties can determine functional dependencies) is chosen. Table (5.14) presents the results of the implementation of these algorithms to get attribute data reduce. From this table, Johnson Reducer gives a better result in which it gives less number of reducts, less number of rules, less number of cardinality, and high number of support.

**Table 5.14: Evaluation measurements of reduction and rules produced by different algorithms**

<b>Reduction Algorithm</b>	<b>No. reduct</b>	<b>No. rule</b>	<b>Cardinality</b>	<b>Support</b>
<b>Genetic algorithm</b>	28	19040	7,8,9,10	100
<b>Johnson algorithm</b>	1	680	7	100
<b>Holte1R</b>	16	942	1	1
<b>Manual</b>	1	678	15	0
<b>SVA Genetic</b>	34	23120	8,9,10,11	100
<b>JohnsonReducer</b>	1	559	3	100
<b>HolteRReducer</b>	16	942	1	1
<b>ManualReducer</b>	1	678	15	0

Figure (5.3) shows the percentage of reduction resulting from the implementation of JA on the dataset features in which it that gives one reduct with three attributes. This means that the percentage of reduction is 84% .



**Figure 5.3: The percentage of the attributes before and after reduction**

#### **5.4 Model Results:**

Features that resulted from the reduction process had been used to build a simulation model. This model consists of two page. First, the user is recorder profile information (name, E-mail address ,and age. Second page, a picture showing a group different topics and the user selects whether the pictures is (like, dislike). Result execute getting opinions to (70) users.

## **5.5 Discussion:**

To test the proposed model, 70 participants are used. They distributed as: Age from 13 to 68, 0.14% participants outside Facebook site, and (0.86%) from Facebook site.

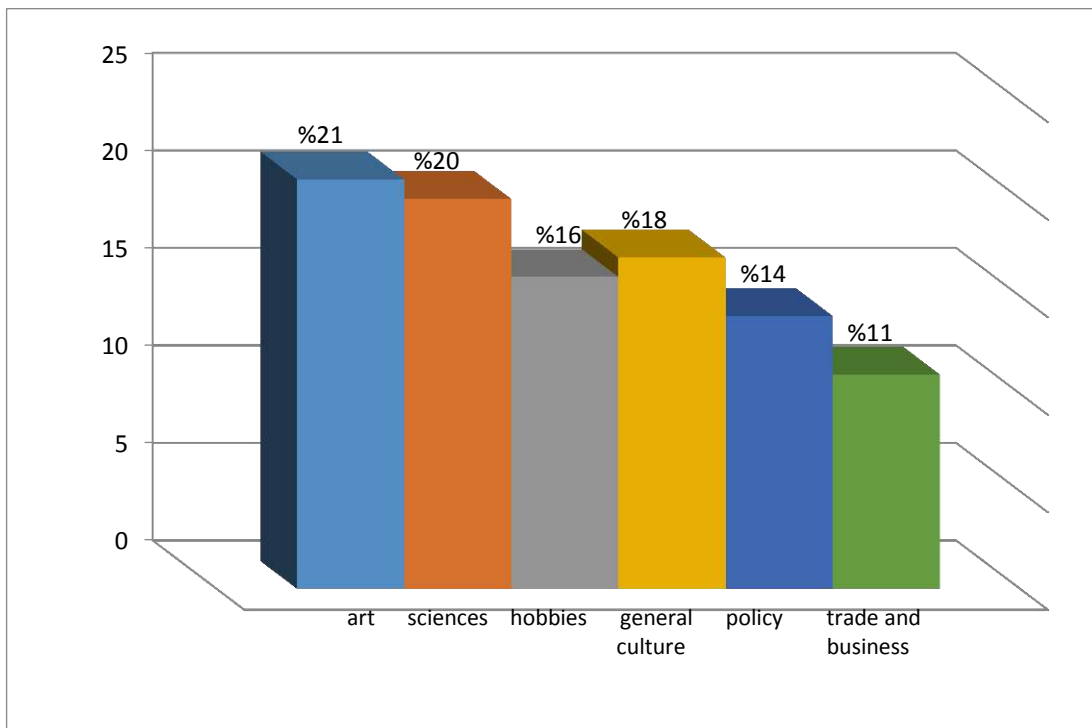
Figure (5.4) shows participants descriptions in Facebook site where:

1. (21%) of the participants with an interest in the art field such as (drawing, video, celebrity news, design, life style , and music)
2. (20%) of the participants had scientific interests represented (scientific news, education, technology, software, geology, health, family and child),
3. (16%) of the participants were concerns hobbies, interests represented (sport, commentators, games, hobbies, travel),
4. (18%) of the general culture and interests of the participants was represented by (blog, a variety of information, religion, literature, history, general culture),
5. (14%) of the participants with an interest in the policy field (local news, world news, politics, events),
6. (11%) trade and business, industries, jobs, shopping.

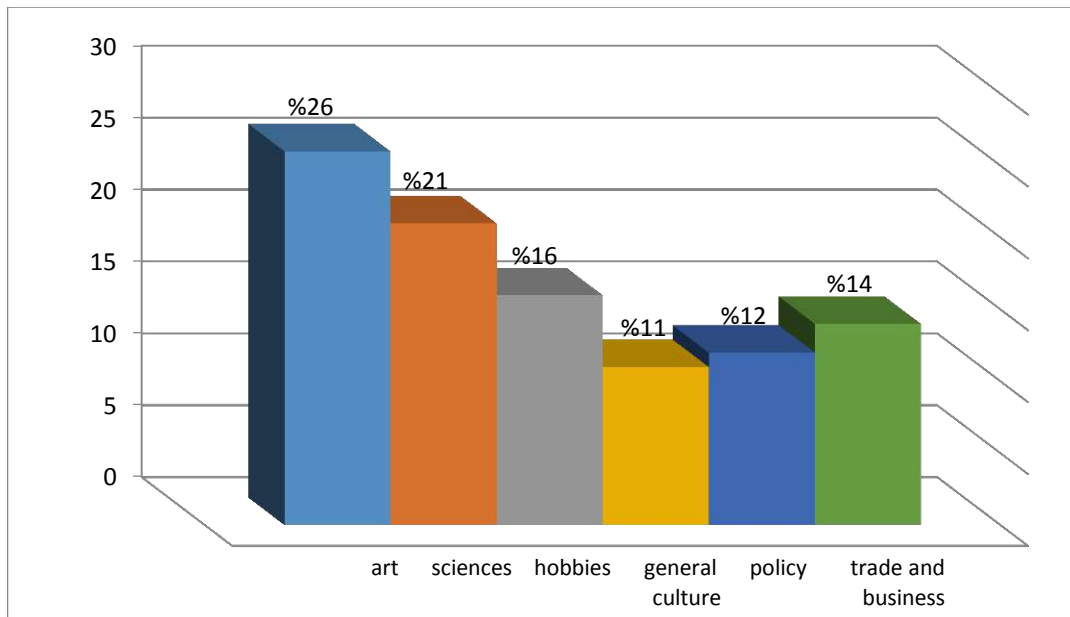
Similarly Figure (5.5) describes the participants outside Facebook site in which:

1. (26%) of the participants interest in art field.

2. (21%) participants interest concerns of scientific interests.
3. (16%) of hobbies interest.
4. (11%) from participants interested with general culture.
5. (12%) of the participants their policy interests.
6. (14%) of the participants concerns of trade and business.



**Figure 5.4: Description of the particular distribution in testing model**



**Figure 5.5: Description of the particular distribution in testing model (users outside Facebook site)**

Finally Table (5.15) shows the confusion matrix in which the overall accuracy of model is (0.94%).

**Table 5.15: The output confusion matrix**

		Predicted			
		0	1		
Actual	0	32	1	0.969697	
	1	3	34	0.918919	
		0.914286	0.971429	0.942857	



***Chapter Six***  
***Conclusion and***  
***Future Work***

## Chapter Six

### Conclusion and Future Work

#### **6.1 Conclusion:**

User information in social networking sites provides a great benefit to predict the concerns of the new users and the permanent user preferences supported in the Facebook site. Also, this information provides a great interest to systems developers and marketers in social sites. The interests of the user of each period may vary greatly since it depends on the psychological state or interest emergency (searching for a particular topic for some reasons). We try to get the largest possible amount of topics of interest to the user to get the most for the construction of modeling user support.

This thesis proposed a Facebook user modeling using direct features derived from Facebook users depending on unsupervised learning method and rough set theory. The cold start problem is one of the steps that we are trying to discuss and suggested solutions for it, by depending of user information in another site or in user history. The adoption of information on social networking sites have very effective in solving this problem.

One of the major challenges that we faced in the preparation of the study is to collect user information because of Lack of credibility in the user information, and what characterized by Arab societies of the



determinants because of privacy. So been preparing information form to collect information from users of social networks sites, where we get (680) participants.

The proposed model for Facebook users depends on the activity in social networking sites in addition to the daily activity at the same sites.

In this thesis, the use of information form to collect data from social network users, then used un supervision learning that helps to group similar characteristics, and extract features of the social network user, relied on the apply RST, which would reduce the number of attribute and reduce complexity as well as the possibility of dealing with a complete data, were able to reduce the number of features by 84%, and also gave us the accuracy rate of 94.28%.

## **6.2 Future Work:**

In future work, we plan to evaluate our work within the framework of the proposed model applied on Ready dataset. For the purpose of obtaining more data comprehensiveness of users of communication networks, social should be taken the attributes from the proposed model as essential attributes to building future systems.

Also using the advantage of existing services in Semantic Web from the structural building Favorites could improve results.



# ***References***

## References

- [1] A. A. Alassiri, M. B. Muda, R. B. Ghazaliet , " Usage of Social Networking Sites and Technological Impact on the Interaction-Enabling Feature", *International Journal of Humanities and Social Science* , 4(4), PP.47-48, Special Issue – February, 2014.
- [2] <http://www.invo.org.uk/wp-content/uploads/2014/11/9982-Social-Media-Guide-WEB>. [Accessed at 9/2/2016].
- [3] C. M. Cheung, P. Chiu, and M. K. Lee. "Online social networks: Why do students use Facebook?." *Computers in Human Behavior*, PP.1337-1343, 2011.
- [4] T. Jagatic, N. Johnson, M. Jakobsson and F. Menczer "Social phishing", *Communications of the ACM*, 50(10), PP.94–100, 2007.
- [5] A. Stadd, "Bookmark It: A Comprehensive Look At 2014 Global Internet And Social Media Stats [DECK]", 2014, Available at: <http://www.adweek.com/socialtimes/2014-global-internet-social-media-stats/495357>.
- [6] T. Zhou, "Understanding online community user participation: a social influence perspective", *Internet Research*, 21(1), PP.67-81, 2011.
- [7] N. B. Ellison and d. boyd , "Sociality through Social Network Sites", In Dutton, W. H. (Ed.), *The Oxford Handbook of Internet Studies*. Oxford: Oxford University Press, pp. 151-172, 2013.
- [8] T. Rodrigues , F. Benevenuto , M. y. Cha, K. P. Gummadi , and V. Almeida," On word-of-mouth based discovery of the web",

In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, PP. 381-396, 2011.

- [9] N. Rob, and C. Near. "Jesus is my friend: Religiosity as a mediating factor in Internet social networking use", *AEJMC Midwinter Conference, Reno, NV*. 2007.
- [10] M. E. Zaglia , "Brand communities embedded in social networks", *Journal of business research*, 66(2), PP. 216-223, 2013.
- [11] S. Doyle, "The role of social networks in marketing", *Journal of Database Marketing & Customer Strategy Management*, 15(1), pp. 60–64,2007.
- [12] S. M. Mark, J. Galaskiewicz, "Networks of inter organizational relations", *Sociological Methods & Research*, 22(1), pp. 46–70, 1993.
- [13] J. Raacke, J. B. Raacke, "MySpace and Facebook: Applying the uses and gratifications theory to exploring friend-networking sites", *Cyber psychology & Behavior*, 11(2), pp. 169–174, 2008.
- [14] A. Rapoport, J. H. William, "A study of a large sociogram", *Behavioral Science*, 6(4), pp. 279–291,1961.
- [15] K. d. Valck, G. H. Bruggen, and B. Wierenga, "Virtual communities: A marketing perspective", *Decision Support Systems*, 47(3), pp. 185–203, 2009.
- [16] M. G. White , "What type of Social Network Exist", available at: <http://socialnetwork.lovetoknow.com>.

- [17] T. Aichner, F. Jacob "Measuring the degree of corporate social media use", *international journal of market research*, 57(2), PP. 257–275, 2015.
- [18] D. Kirkpatrick, "*The Facebook effect: the real inside story of Mark Zuckerberg and the world's fastest-growing company*", 2011.
- [19] B. J. Raymond, M. N. Erich, P. O. David, D. Saha, Z. Shae, and C. Waters, "A study of internet instant messaging and chat protocols", *IEEE Network*, 20(4), PP. 16-21, 2006.
- [20] J. Payne, "Choosing the Right Social Media Channels for Your Business" , available at: <http://www.business2community.com/social-media/choosing-right-social-media-channels-business-0860210#k8Bopvqlj9FCFFUI.97>.
- [21] H. Ajmera, "Social Media 2014 Statistics – an interactive Infographic you've been waiting for!", available at: <http://blog.digitalinsights.in/social-media-users>, 2014.
- [22] T. Steiner, "A meteoroid on steroids: ranking media items stemming from multiple social networks", *In: Proceedings of the 22nd International Conference on World Wide Web ACM*, PP. 31-34, 2013.
- [23] M.S. Hufschmidt, U. Malinowski, and T. Kuhme, "*Adaptive user interfaces: Principles and practice*", Elsevier Science Inc., 1993.
- [24] F. Martinez, Y.C. Sherry, and L. Xiaohui, "Survey of data mining approaches to user modelling for adaptive hypermedia. Systems, Man, and Cybernetics, Part C: Applications and Reviews" , *IEEE Transactions on Computational Social Systems*, 36(6), PP.734-749, 2006.

- [25] R. Girardi and C. Faria, "An ontology-based technique for the specification of domain and user models in multi-agent domain engineering", *CLEI electronic journal*, 7(1):7, 2004.
- [26] M. Potey, and P. K. Sinha, "Review and analysis of machine learning and soft computing approaches for user modeling", *International Journal of Web & Semantic Technology IJWest*, 6(1), P. 40, January 2015.
- [27] F. Carmagnola, F. Cena, O. Cortassa, C. Gena, and I. Torre, "[Towards a tag-based user model: how can user model benefit from tags? In the Proceedings of UM](#)", *International Conference on User Modeling*, pp. 445-449, [2007](#).
- [28] N. B. Ellison, and D. M. Boyd, "Social Network Sites: Definition, History, and Scholarship", *Journal of Computer-Mediated Communication*, 13(1), PP.210–230, October, 2007.
- [29] R. V. Kulkarni, A. Förster, and G. K. Venayagamoorthy, "Computational intelligence in wireless sensor networks: a survey", *Communications Surveys & Tutorials IEEE*, 13(1), PP.68-96, 2011.
- [30] M. Mittal, and K. Kumar, "Quality of Services Provisioning in Wireless Sensor Networks using Artificial Neural Network", *International Journal of Computer Applications*, 117(5), PP.28-40, 2015.
- [31] A. Rosel, "Exploration of Methods for Optimizing Portfolio Tracking with Partial Replication", Spring 2015. [http://michael.hahsler.net/SMU/EMIS8331/tutorials/NN\\_slides.pdf](http://michael.hahsler.net/SMU/EMIS8331/tutorials/NN_slides.pdf). [Accessed at:21/8/2015].

- [32] C. Shang and Q. Shen, "Aiding neural network based image classification with fuzzy-rough feature selection", *International Conference on IEEE*, PP. 976 – 982, 2008.
- [33] A. Growe, "Comparing Algorithms and Clustering Data: Components of the Data Mining Process", PhD Thesis. Grand Valley State University, 1999.
- [34] X. Hang, X. Zhang, and Y. Du, "A comparison of neural network, rough sets and support vector machine on remote sensing image classification" *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, 7, 2008.
- [35] J. Kamruzzaman, "Artificial neural networks in finance and manufacturing", *Thesis in University of New South Wales, Australia*, IGI Global, 2006.
- [36] E. H. S. Alhrer, "Diagnoses of Erythemato-Squamous Diseases Using Rough-Neuro Model", Thesis in Computer Science Department Faculty of Computers & Information Mansoura University, 2014.
- [37] R. J. Sassi, L.A. Silva, and E. D.Hernandez, "Neural Networks and Rough Sets: A Comparative Study On Data Classification", *Neural Networks and Rough Sets: A comparative study on data classification*. In: IC-AI., PP. 175-180, 2006.
- [38] M. E. yahia, and R. Mahmud. "hybrid expert system of rough set and neural network" , *Malaysian journal of computer science* 12 (1), PP.1-8, 1999.
- [39] T. Kohonen, "Self-Organizing Maps", *Springer Series in Information Sciences*, 30, Springer, Heidelberg, 1st ed., 1995; 2nd., 1997

- [40] J. Yang, R. L. Cheu, X. Guo, and A. Romo, " Analysis of vehicle-following heterogeneity using self-organizing feature maps", *Computational Intelligence and Neuroscience*, 27, 2014.
- [41] Gan, C. Ma, and J. Wu, "*Data clustering: theory, algorithms, and applications*", 20, Siam, 2007.
- [42] N. Yorek, U. Ilker , and H. Aydin. "Using Self-Organizing Neural Network Map Combined with Ward's Clustering Algorithm for Visualization of Students' Cognitive Structural Models about Aliveness Concept", *Computational Intelligence and Neuroscience 2016*, 2015.
- [43] K. Pang, "Self-organizing Maps", [https:// www. google. com.eg /advantage and disadvantage of self-organization map](https://www.google.com.eg/advantage%20and%20disadvantage%20of%20self-organization%20map), 2006.
- [44] V. A. Patole, V. K. Pachghare, and P.Kulkarni, "Self-Organizing Maps to build intrusion detection systems", *Journal of Computer Applications*, 1(7), 2010.
- [45] T. Slimani. "Application of rough set theory in data mining" *arXiv*, PP. 1311.4121, 2013.
- [46] A. E. Hassanien, A. Abraham, and F. Herrera, "Ultrasound biomicroscopy glaucoma images analysis based on rough set and pulse coupled neural network", *In Foundations of Computational Intelligence* ,2, pp.275-293, 2009.
- [47] A.B. M. Salem, M. Roushdy, and S. A. Mahmoud, "Mining patient data based on rough set theory to determine thrombosis disease", *Academic Manuscript Central*, ICGST-AIML, 5(1), PP. 27-31, 2005.



- [48] A. R. Hedar, J. Wang, and M. Fukushima, "Tabu search for attribute reduction in rough set theory", *Soft Computing*, 12(9), PP. 909-918, 2008.
- [49] A. E. Hassanien, A. Abraham, P.F. James, and G. Schaefer, "An overview of rough-hybrid approaches in image processing", *IEEE World Congress on Computational Intelligence, IEEE International Conference on. IEEE*, 2008.
- [50] S. K. Pal, and P. Mitra, "Case generation using rough sets with fuzzy representation", *Knowledge and Data Engineering, IEEE Transactions on*, 16(3), PP. 293-300, 2004.
- [51] A. E. Hassanien, A. Abraham, J.F. Peters, and J. Kacprzyk, "Rough sets in medical imaging: foundations and trends", *Computational Intelligence in Medical Imaging: Techniques and Applications*, PP. 47-87, 2008.
- [52] J. Bazan, H. S. Nguyen, and M. Szczuka, "A view on rough set concept approximations", *Fundamenta Informaticae*, 59(2-3), PP. 107-118, 2004.
- [53] A. Saxena , L. K. Gavel, and M. M. Shrivastava, "Rough Sets for Feature Selection and Classification: An Overview with Applications", *International Journal of Recent Technology and Engineering (IJRTE)*,3(5), November, 2014.
- [54] X. Wang, J. Yang, R. Jensen, X. Liu, "Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma", *Computer Methods and Programs in Biomedicine*, 83, PP.147-156, 2006.
- [55] L. Xiaohan, "Attribute Selection Methods in Rough Set Theory", PhD Thesis, San José State University, 2014.

- [56] K. Al-Aidarous, A. Abu Bakar, and Z. Othman, "Improving Naïve Bayes Classification with Rough Set Analysis", *International Journal of Advancements in Computing Technology*, 5(13), P.49, 2013.
- [57] C. Lian, H. Liu, and Z. Wan, "An attribute reduction algorithm based on rough set theory and an improved genetic algorithm", *Journal of Software*, 9(9), PP. 2276-2282, 2014.
- [58] K. A. Aidarous , A. A. Bakar, and Z. Othman, "Data classification using rough sets and naïve bayes", *In: International Conference on Rough Sets and Knowledge Technology*. Springer Berlin Heidelberg, PP.134-142, 2010.
- [59] M. N. Rahman, M. L.Yuzarimi, and F. Mohamed, "Applying Rough Set Theory in Multimedia Data Classification", *International Journal on New Computer Architectures and Their Applications IJNCAA*, 1(3), PP.683-693. 2011.
- [60] Y.Kay, Ö.F.Ertuğrul and R.Tekin, "A Rough Set Approach for Modeling E-mail Usage Habits", *Computer Science*, 1(4), PP.259-264, 2014.
- [61] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity", PP.1202.0332, (2012).
- [62] S. Sulaiman , S.M. Shamsuddin, and A. Abraham, "Rough neuro-PSO web caching and XML prefetching for accessing Facebook from mobile environment", *Nature & Biologically Inspired Computing, 2009 (NaBIC 2009)*, World Congress on. IEEE, 2009.
- [63] S. K. Ray and S. Singh, "Rough set based social networking framework to retrieve user-centric information", *In Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*. Springer Berlin Heidelberg, PP.184-191, 2009.

- [64] A. Mitra, S. Rani and S. S. Paul, "Clustering Analysis in Social Network using Covering Based Rough Set", *IEEE 3rd International Advance Computing Conference (IACC)*, 2013, PP.476-481, 2013.
- [65] M. Bekkali, I. Sahmoudi and A. Lachkar, " Enriching Arabic Tweets Representation based on Web Search Engine and the Rough Set Theory", *In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, PP. 1573-1574, 2015.
- [66] W. Zuo and W. Zhe, "Research in Social Network Based on Rough Set Clustering Algorithm", *International Journal of Advancements in Computing Technology*, 4(15), 2012.
- [67] R. A. AboKhachfeh and I. Elkabani, "Using rough sets in homophily based link prediction in online social networks", *In Computer Applications and Information Systems WCCAIS, World Congress on*, PP.1-6, January 2014.
- [68] T. Plumbaum, S. Wu, E. W. D. Luca, and S. Albayrak." User modelling for the social semantic web". In *SPIM*, PP. 78-89, 2011.
- [69] B. H. Lim, D. Lu, T. Chen, and M.Y. Kan , "# mytweet via Instagram: Exploring User Behaviour across Multiple Social Networks", *In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ACM*, PP.113-120, 2015.
- [70] F. T. O'Donovan , C. Fournelle, S. Gaffigan, O. Brdiczka, J. Shen, J. Liu, and K.E. Moore, "Characterizing user behavior and information propagation on a social multimedia network", *in: multimedia and expo workshops (ICMEW), 2013 IEEE International Conference on. IEEE*, PP.1-6, 2013.

- [71] E. F. Martinez, S. Y. Chen, and X. Liu, "Survey of Data Mining Approaches to user modeling for adaptive hypermedia", *IEEE Transactions on Systems*, 36(6), PP.737-749, November 2006.
- [72] H. K. Mahdi, H. K. Mohamed and S. S. Attia, "Data Mining for Decision Making in Multi-Agent Systems", In Tech, 2011.



# ***Appendix***

الملحق

# الملخص العربي

### المخلص:

تمثل الشبكات الاجتماعية حيز اهتمام واسع للباحثين في مجالات البحث العلمي، وذلك بسبب ظهور وانتشار عدد كبير جدا من مواقع الشبكات الاجتماعية، وكذلك الاهتمام الواسع للأشخاص في هذه المواقع، ايضا لما تحتويه هذه المواقع من كمية كبيرة ومتنوعة من البيانات والتي تتناقل في الموقع الواحد اوبين المواقع المختلفة (مثل الصور والرسائل، والمعلومات الشخصية، والأخبار العالمية والمحلية، والبحوث العلمي، وغيرها من المعلومات).

الهدف الرئيسي من هذه الرسالة هو بناء نموذج للمستخدم في المواقع الاجتماعية، اعتمادا على نشاط المستخدم في المواقع المختلفة، واستخدام السمات الأكثر ارتباطا، لبناء النموذج للمستخدم الموجودة في الموقع، فضلا عن التنبؤ باهتمامات المستخدم الجديد للموقع، وهي نقطة مهمة في هذا المجال البحثي.

واحدة من أهم التحديات التي واجهتنا في إعداد الدراسة هو جمع معلومات المستخدم ولأسباب عده منها عدم وجود مصداقية في بعض معلومات التسجيل للمستخدم، كذلك ما تتميز به المجتمعات العربية من المحددات منها الخصوصية والامنية. مما جعل الباحثة تعمل على إعداد استمارة معلومات لجمع المعلومات من مستخدمي مواقع التواصل الاجتماعي قمنا بتوزيع استمارة المعلومات على العائلة والاصدقاء وطلبة كلية الحاسبات والمعلومات قسم علوم الحاسب في جامعة المنصور كذلك بنشر الرابط في عدد من المواقع الاجتماعية، حيث حصلنا على (٦٨٠) مشاركا. المراحل الاساسية للنموذج المقترح هي ( جمع البيانات، المعالجة الاولية للبيانات، عمل تكتلات للبيانات، واجراء تقليص للخصائص).

اظهرت النموذج المقترح باستخدام خوارزميات التقليص لعدد السمات نسبة ٨٤% وكانت النسبة الناتجة للدقة ٩٤,٢٨%.



بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

قَالَ رَبِّ اشْرَحْ لِي صَدْرِي<sup>٢٥</sup>  
وَيَسِّرْ لِي أَمْرِي<sup>٢٦</sup>  
وَاحْلِلْ عُقْدَةَ مِنِّ لِسَانِي<sup>٢٧</sup>

صَدَقَ اللَّهُ الْعَظِيمُ

سورة طه ((الآيات ٢٥-٢٧))

جامعة المنصورة  
كلية الحاسبات والمعلومات  
قسم علوم الحاسب



## نمذجة مستخدمي الشبكات الاجتماعية

اسم الباحثة

**وسن عبدالله عبداللطيف الأوسي**

قسم علوم الحاسب- كلية الحاسبات والمعلومات - جامعة المنصورة - مصر  
كلية العلوم- جامعة القادسية - وزارة التعلم العالي والبحث العلمي- العراق

رسالة

مقدمة الى قسم علوم الحاسب- كلية الحاسبات والمعلومات- جامعة المنصورة  
كجزء من متطلبات الحصول على درجة الماجستير في علوم الحاسب

تحت إشراف

**أ.م.د. سمير الدسوقي الموجي**      **د. شاهنده صلاح الدين سرحان**

مدرس في قسم علوم الحاسب

كلية الحاسبات والمعلومات

جامعة المنصورة

رئيس قسم علوم الحاسب

كلية الحاسبات والمعلومات

جامعة المنصورة

جامعة المنصورة – جمهورية مصر العربية

2016