Arab American University

Faculty of Graduate Studies

Development of Facebook Data Extraction, Analysis and Visualization Tool

By

Ahmed Qerem

Supervisor

Dr. Khalid Rabayah

This thesis was submitted in partial fulfillment of the requirements for the

Master`s degree in

MSC. in Computer Science

July/ 2018

Development of Facebook Data Extraction, Analysis and Visualization Tool

By
Ahmed Qerem

This thesis was defended successfully on 17/7/2018 and approved by:

Committee members	Signature
1. Supervisor: Dr. Khalid Rabayah	……..
2. Internal Examiner: Dr. Mohammed Maree	……..
3. External Examiner: Dr. Rashid Jayousi	……..

Declaration

This is to declare that the thesis entitled “Development of Facebook Data Extraction, Analysis and Visualization Tool” under the supervision of “Dr. Khalid Rabayah” is my own work and doesn’t contain any unacknowledged work or material previously published or written by other person, except where due reference is made in the text of the document.

Date: 17/7/2018

Name: Ahmed Qerem Signature:

Dedication

I dedicate this project to Dr. Khalid Rabayah, my source of inspiration, knowledge and understanding. He has been the source of my strength throughout this program and on his wings only have I soared.

I also dedicate this work to my parents and my wife, who has encouraged me all the way and whose encouragement has made sure that I give it all it takes to finish that which I have started.

Thank you. My love for you all can never be quantified. God bless you.

Acknowledgement

Firstly, I would like to thank my thesis advisor Dr. Khalid Rabayah. The door to Dr. Rabayah office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to acknowledge Dr. Mohammed Maree and Dr. Rashid Jayousi as the second reader of this thesis, and I am gratefully indebted to them for the very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents and to my wife for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Abstract

Facebook has become an indispensable tool in our everyday lives. It is often used as a platform where people can share their opinions, ideas, special moments, feelings and many other forms of social expressions. It is also used in various fields such as business, education and health care for exchanging and sharing specialized knowledge and experiences.

The frequent use of Facebook by a large number of people on a daily basis results in massive amounts of data. For instance, an amount of 4 petabytes of data were generated by an average of 1.23 billion daily active users in December 2016. Facebook is thus seen as a rich data source which is complex to process and analyze. In this respect, big data and data mining are emerging approaches aimed at the analysis of massive datasets to obtain helpful and useful information that can be used in different areas.

Along the dominance of Facebook as a social network site where such massive datasets are generated, an increasing number of tools have been developed to process and analyze Facebook generated data. These tools can be classified based on their use into a number of categories. First, data extraction tools which are used to fetch data out of Facebook and store it in different formats. A second category of tools refer to tools that allow data analysis. Finally, data visualization tools which are important to generate a readable and useful output.

With the spread use of social network sites, more and more researches started to employ it in many fields. Predicting the future is one of these fields, depending on social network sites content we can predict the real-world outcomes. On the other hand, studying tie strength between friends is another application. Some people use social network sites for commercial purposes, for marketing and studying opponents. Healthcare is another filed addressed by social network sites

where it helps healthcare organizations, clinicians and patients. Social network sites has used as a communication media within the instructors and students to create online classes.

Many of these tools have been studied and investigated in order to develop a single tool that combines extraction, analysis and visualization features and to overcome the existing tools limitations. However, developing such a tool is complex since none of the investigated tools can be used independently.

Our goal is to develop a universal tool to extract the data out of Facebook using Facebook graph API and to analyze the collected data based on different algorithms to get time measures as well as data measures. Finally, visualize the analyzed data in form of charts to be used by analysts and decision makers.

In this thesis, we implement a new web-based tool based on Spring framework that is capable of fetching data out of Facebook using Facebook API and then analyze these data to get the Facebook page key performance indicators such as time related and data related information. Finally, the tool is capable of visualizing the key performance indicators to the user in the form of charts.

The new developed tool allows users to run their own inquires with a few clicks and without any background in datamining or programming. This was achieved via our tool that has its own crawler, analyzer and visualizer. The crawler has been developed on the top of “httpClient” to get the data from Facebook. Currently, it depends on Facebook graph API to fetch the pages-related data since it is the only available API that can be trusted. Earlier studies suggested the design of a new crawler that crawls Facebook DOM. This kind of crawlers faced some issues with Facebook due to access rate limitations. The crawlers will therefore be blocked by Facebook after a period of time.

The tool was successfully used to fetch, analyze and visualize data out of tens of thousands of posts and comments from different kinds of Facebook pages, including university pages, news pages, and commercial pages. The tool successfully and thoroughly generated professional charts out of these pages that can practically help in the decision making process.

Contents

Abstract vi

Chapter 1 Introduction 1

1.1 Research background 1

1.2 Motivation 2

1.3 Research objectives and significance 3

1.4 Thesis structure 4

Chapter 2 Theoretical Background and Literature Review 5

2.1 Social Media 5

2.2 Knowledge gap and research rationale 6

2.3 Conclusion 10

Chapter 3 Analysis and Design of the Facebook Data Extraction Tool 11

3.1 Tool analysis: goals and objectives 11

3.2 Major components 11

3.3 Requirements 12

3.3.1 Non-functional requirements 14

3.4 Software design 14

3.4.1 Structural models 15

3.4.2 Interaction models 17

3.4.3 Context models 21

3.4.4 Complexity measurements 22

3.5 Database design 23

3.6 Software interface 25

3.7 Software testing and validation 28

Chapter 4 Data collection, analysis and visualization 30

4.1 Facebook data collection using developed tool 30

4.2 Facebook Data Extraction and Visualization 31

4.2.1 Educational Pages 33

4.2.2 News Pages 43

4.2.3 Commercial Pages 50

4.3 Tool metrics 58

Chapter 5 Summary and conclusions 60

5.1 Research overview 60

5.2 Contribution of the study 60

5.3 Research limitations **61**
5.4 Future research directions **61**
References **63**
الملخص **66**

Table of Figures

Figure 1: Proposed MVC architecture 13
Figure 2: Crawler class diagram. 16
Figure 3: Analysis class diagram. 16
Figure 4: View class diagram. 17
Figure 5: User use case. 17
Figure 6: System use case. 18
Figure 7: Crawling sequence diagram. 19
Figure 8: Analysis sequence diagram. 19
Figure 9: Word cloud sequence diagram. 20
Figure 10: Data clean up sequence diagram. 20
Figure 11: Context diagram. 21
Figure 12: Activity diagram. 22
Figure 13: Complexity factors. 22
Figure 14: ER diagram. 24
Figure 15: home view 25
Figure 16: Extract view. 26
Figure 17: Crawled items panel. 26
Figure 18: Profile view. 27
Figure 19: Profile view. 27
Figure 20: Profile view. 28
Figure 21: Tool statistics, percentage of posts to comments. 31
Figure 22: Tool statistics, number of crawled posts per month 32
Figure 23: Tool statistics, number of crawled comments per month. 32
Figure 24: Percentage of posts to comments. 33
Figure 25: Comment rate per Posts. 33
Figure 26: Post rate per hour. 34
Figure 27: Comments rate per hour. 34
Figure 28: Post rate per month. 35
Figure 29: Comment rate per month. 35
Figure 30: Arab American University Word Cloud. 36
Figure 31: Percentage of comments to posts. 37
Figure 32: Comments rate per post. 37
Figure 33: Posts rate per hour. 37
Figure 34: Posts rate per month. 38
Figure 35: Comments rate per hour. 38
Figure 36: Comments rate per month. 39
Figure 37: An-Najah National University word cloud. 39
Figure 38: Percentage of comments to posts. 40
Figure 39: Comments rate per post. 40
Figure 40: Post rate per hour. 41
Figure 41: Comment rate per hour. 41
Figure 42: Posts rate per month. 42
Figure 43: Comments rate per month. 42

Figure 44: Birzeit word cloud. 43
Figure 45: Percentage of comments to posts. 44
Figure 46: Comments rate per post. 44
Figure 47: Posts rate per hour. 45
Figure 48: Comment rate per hour. 45
Figure 49: Posts rate per month. 45
Figure 50: Comments rate per month. 46
Figure 51: Ramallah news word cloud. 47
Figure 52: Percentage of comments to posts. 47
Figure 53: Number of comments per post. 48
Figure 54: Post rate per hour. 48
Figure 55: Comment rate per hour. 49
Figure 56: Comments rate per month. 49
Figure 57: Posts rate per month. 49
Figure 58: Shehab word cloud. 50
Figure 59: Percentage of comments per posts. 51
Figure 60: Comments per post 51
Figure 61: Posts rate per hour. 52
Figure 62: Comments rate per hour. 52
Figure 63: Posts rate per month. 52
Figure 64: Comments rate per month. 53
Figure 65: Jawwal word cloud. 53
Figure 66: Percentage of comments to posts. 54
Figure 67: Comments rate per post. 54
Figure 68: Posts rate per hour. 55
Figure 69: Comments rate per hour. 55
Figure 70: Posts rate per month. 56
Figure 71: Comments rate per month. 56
Figure 72: Wataniya word cloud. 57

List Of Abbreviations

DOM: Document object model vii
SNS: Social Network Sites 2
OSN: Online social networks 8
MVC: Model view controller 58
MVC: Model view controller 66

Chapter 1 Introduction

1.1 Research background

Facebook has become an indispensable tool in our everyday lives. It is often used as a platform where people can share their opinions, problems, special moments, feelings and many other forms of social expressions. It is also used in various fields such as business, education and health care for exchanging and sharing specialized knowledge and experiences.

The frequent use of Facebook by large number of people on a daily basis results in massive amounts of data. For instance, an amount of 4 petabytes of data were generated by an average of "*1.23 billion daily active users in December 2016"*[1]. Facebook is thus seen as a rich data source which is complex to process and analyze. In this respect, big data and data mining are emerging approaches aimed at the analysis of massive datasets to obtain helpful and useful information that can be used in different areas.

Along the dominance of Facebook as a social network site where such massive datasets are generated, an increasing number of tools have been developed to process and analyze Facebook generated data. These tools can be classified based on their use into a number of categories. First, data extraction tools which are used to fetch data out of Facebook and store it in different formats. A second category refers to tools that allow for the analysis of data. Finally, data visualization tools which are important to generate a readable and useful output to be used.

Many of these tools have been studied and investigated in order to develop capabilities into a single tool that combines extraction, analytical and visualization features. However, developing such a tool is complex since none of the investigated tools can be used independently. Moreover, certain tools are designed for a special purpose and are only limited to that purpose such as

[1] https://newsroom.fb.com/company-info/ visited in 10/04/17

opponent analysis and police investigation tools. Commercial tools represent an example of special purpose tools which are designed to collect information about competitors' and customers' needs and are therefore limited to data extraction. Other categories such as analysis and visualization tools are partially used since the output information is handled by marketing specialists and their need is limited to visualize information in tabular forms. Further, certain tools are designed to be fully visualization tools to represent the data in an interconnected manner such as friendship tree exporting tools where each leaf represents a user and links represent the friendship relation. These tools can be helpful in studying human behavior and their relations. Analysis part is not fully implemented in the reviewed tools in conjunction with data extraction and visualization. So, basically data is being analyzed by data mining tools before it can be used.

Based on the above discussion, there is an urging need to develop a universal integrated tool to collect and analyze Facebook data following data mining techniques that can be helpful in getting hidden patterns and visualizing results in different formats.

1.2 Motivation

The use of Social Network Sites (SNSs) is pervasive. People create personal social profiles on SNSs which provide a communication medium to share personal moments and experiences related to various aspects of their social lives. They also share their personal interests and attitudes as well as interact with others on SNSs. Studying users' profiles on SNS may provide an understanding of their social properties and trends.

SNSs are not limited to sharing information about social life. People nowadays can form specialized communities in areas like education, commerce, and healthcare. These communities create and co-create content based on mutual interests.

With the spread use of SNSs, more and more researches started to employ it in many fields. Predicting the future is one of these fields, depending on SNSs content we can predict the real-world outcomes [1]. On the other hand, studying tie strength between friends is another application [2]. Some people used SNSs for commercial purpose, marketing [3], studying opponents [4]. Healthcare is another filed touched by SNSs where it helps healthcare organizations, clinicians and patients [5]. SNSs has been used as a communication media within the instructors and students to create online classes [6].

Taking into consideration the importance of SNSs and their effect on our life, we can say that studying SNSs will help us to understand society and provide solutions to make our life better. We use the outcomes of SNSs to support our decisions. Enhance the healthcare services by studying the society health habits from SNSs.

Educational filed could be enhanced by analysis of the student's behavior/habits and focus in finding the reasons behind low scores. Online courses and conversations is another application of SNSs in this field. Finally, we study the SNSs to analyze the commercial opponents and to enhance the quality of the products/services to achieve the maximum benefit.

1.3 Research objectives and significance

This research focuses on exploring data extraction, analysis, and visualization from Facebook. In this context, we review the existing Facebook related tools that help in getting useful information. To do so, we start from data extraction and review and understand existing tools.

Similarly, we review existing analysis methods/tools and determine if there are tools that can do both extraction/analysis. Finally, we explore the visualizations tools. After that, we will compare these tools with each others and find the pros and cons of each tool.

Our research aims to design and implement a new tool having the pros from the explored tools and enhancements for the cons, it will be a universal tool for extracting, analyzing and visualizing Facebook data in order to produce useful information in various relevant areas.

The proposed tool will crawl Facebook pages and store their data in a database after applying data filtration techniques. Then, it will analyze the data and extract the useful information/measures and store them for further use. Finally, visualize the analyzed information in form of charts.

1.4 Thesis structure

The rest of this thesis is organized as follows: Chapter 2 shows the theoretical background and literature reviews and discusses existing approaches and the importance of our research. Chapter 3 provides the analysis and design of the Facebook data extraction tool and discusses the implementation process of the tool considering the database and its major components. Chapter 4 shows how the tool is used in action to extract, analyze and visualize the data. Finally, chapter 5 summarizes the main issues and conclusions from the current research, highlights the limitations of our tool, and presents suggestions for further future work.

Chapter 2 Theoretical Background and Literature Review

2.1 Social Media

Social media is defined as a web-based service that allows users to create their own personal profiles and provide the ability to communicate with other users that share the same connection. These connections can be viewed by users [7]. On the one hand, there are definitions of social media that are application specific referring to platforms such as Facebook and Twitter [8]. On the other hand, social media can be defined considering Web 2.0 and user-generated content as Internet-based applications built on the foundations of Web 2.0 technologies and allow users to create and share content [9].

Referring to the definitions above, we can define social media as web-based applications that allow users to create and share their content taking into consideration the rising technologies such as Web 2.0 and the improvements in the field of big data processing, storage and networking. From application point of view, Facebook and twitter are the dominant platforms so it can be used as a reference for social media.

Talking about social media raises text mining up since it is closely related. It is an automated process to explore a huge amount of data in order to get useful information (patterns) [10]. Other definitions are specific and correspond to the research target area to provide precise description. From information extraction perspective, it can be defined as a process of extracting facts from text, or a process of applying data mining techniques and algorithms to get information from text if we are defining text mining from text data mining perspective. Finally, it is a series of partial steps along with data extraction and data mining processes [11].

Referring to [12], text mining has become involved in several areas such as natural language processing, computational linguistics and information retrieval to help in getting useful

information from unstructured data within a short period of time. Because our research focuses on a specific area (Facebook), we can define it as a process of extracting useful patterns from unstructured text.

Having a dataset is very important in data mining (text mining) to carry out the experiment. In order to get such dataset, web crawling is a good option in case of none of published datasets match the needs, so web crawling is a process of exploiting web tree to move from page to another following in/out links to assist search engines to create their indexes [13]. We can define it as a process of getting structured data that follow a specific format from web [14] and storing it in a dataset in order to make it available for post processing.

Data visualization is one of the fields that should be explained due to its importance in showing up data in readable form. Referring to [15] there are no standards or conventions to be followed in the field of data visualization. There are several communities and each one of them has its own tools/ways based on the way they look at the data. In this respect, data visualization is generally used to visualize the large-scale collections of data [16].

2.2 Knowledge gap and research rationale

Since social media is increasingly becoming prevalent in our daily social lives and interaction, it is reasonable to suggest that it is a rich data source for studying human behavior. To do so, there are a lot of processes that should be followed in order to get useful data from the huge collections of unstructured data. Starting from understanding the goals of the overall process moving to selecting a suitable data set and then preprocessing and cleaning up the data to get it ready to apply different types of data mining algorithms [17].

Social media has been studied for different purposes. Some researchers have studied social media to predict real-word outcomes based on existing data [1], whereas others use social media for marketing purposes such as online tourism [18]. Also it can be used for competitive analysis by tracking what users say about their competitors [4].

Predicting personality from the published user's information on social media is an important part that can be studied to create user personality model which helps in determining user attitudes. This is helpful for instance in predicting romantic relationship success and job satisfaction [19]. Friendship is another factor that can be predicted from social medias by applying similarity measures, so users with the same interests are most likely to be friends [20, 25].

Social media should be analyzed precisely to get a useful information in different areas. So there is a need to have powerful tools to extract, analyze and visualize social media data. Facebook as a rich social media platform has been selected for examination in this research.

Different tools have been studied and developed but unfortunately none of the studied tools has implemented the functions specified above (extraction, analysis and visualization) altogether. A new tool 'Netvizz' was introduced as a general-purpose data extractor for different subsections of Facebook platform [21]. The study provides a new powerful tool to extract the data out of Facebook and operates on Facebook API platform. However, this tool doesn't support data visualization as well as other tools, so adding this feature will enhance the tool to be an end-to-end analytical tool (extraction, analysis and visualization). However, this tool has been proposed since the majority of existing tools are developed for marketing purposes and target Facebook pages. They focus on the owners of these pages rather than the page itself. Researchers are interested in analyzing page content regardless of the owners. However, existing general-purpose tools such as "NameGenWeb"(is no longer available) and "Network Importer" have some

limitations. The former extracts friendship network only and the later extracts the friendship network and data from pages but as a single graph (doesn't support primitive data types).

Referring to [22], a custom tool has been developed to collect and analyze a massive data describing the relation between the Facebook users, the process started on a single seed and stopped after ten days of crawling. A partial sample was collected from Facebook graph containing up to third-level of friendship relation. They used a custom agent to crawl Facebook graph like real user to collect the data from Facebook. The custom agent is a useful app to get friendship relations since Facebook graph API doesn't support fetching public data from user's profiles (Facebook allows this for those with a developer account only). In fact, the visualization part of this tool needs to be enhanced. Moreover, it focuses on friendship relation only.

Another tool has been introduced to represent the large online social networks (OSN) in a meaningful graph since it is not a simple job due to the large number of elements and the need for optimization algorithms [23][26]. Experiment was conducted on Facebook and the resulting undirected graph composed of 547,302 vertices and 836,468 edges. Then, they applied the analysis and visualization techniques. This tool focuses on graph visualization in Facebook friendship relations. To do so, they developed a custom agent using java acts like real user to collect the data from Facebook. The custom agent is a useful app to get friendship relations.

Further, for crime investigation purposes, [24] was proposed where investigators receive archive files from Facebook containing full details about the target user. These files are difficult to filter and analyze, since it is merely data dump files. So an automated application was built based on Facebook graph API to retrieve user's data in a simple way, for who has a permission from Facebook (police). It is a special purpose tool for police investigation. However, the researcher built his own Facebook based app, which makes use of Facebook graph API in an automated

process to fetch the data. We can't utilize this approach in fetching data from public groups/pages since we don't have a permission to get user's data.

In [27] a crawling tool has been designed and implemented to crawl Facebook and get structured data to be used as input for analysis and visualization tools that they rely on as a third-party tool to do the required post processing. They use Gephi as visualization tool since it is free, and easy to use in cross-platforms. Technical challenges have been discussed such as page dimensions, limitation where Facebook doesn't load the entire page but instead loading it asynchronously to avoid network overloading. This adds more complexity and delays to the crawler to deal with asynchronous JavaScript. Dynamic web page and incorrectly formatted characters are other related issues that make the implementation process more challenging. Finally, limited access rate is a major issue where the Facebook profile used by the crawler will be blocked after period of time due to its high access rate. While this study provides a powerful crawling tool to crawl Facebook data and resolve the mentioned challenges above, it ignores the analysis and visualization parts and relies on third-parties instead. So using such tools add limitations to what functionalities are being supported.

Similar to previous solutions, the authors of [28] provide a solution to crawl Facebook without taking care of other post processing steps, so it is simply collecting the data.

A new tool has been introduced in [29] to fetch the data out of Facebook and then visualize it in form of social graph, this tool is a special purpose aimed to prevent organized crimes by focusing on user interactions. Extraction tool with simple word frequency visualizer has introduced in [30].

Around 25 open source tools have been evaluated in [31] to find the optimal tool that fulfill the requirements and needs of the users to collect, store, analyze and visualize OSNs contents. Tools

have categorized into three categories: Crawling tools, storing and analysis and visualization tools, none of evaluated tools have the functions above.

In addition, during our literature search we didn't find existing tool to be used as an integrated tool to manipulate different aspects of Facebook. Some of them focus on crawling and implementation of custom crawlers to fetch out the data and discard other aspects totally or partially. Analysis is the second aspect that wasn't employed enough in the existing tools to make use of crawled data to get useful information to be presented to the end user to help in different ways. Finally, visualization has been implemented in some tools for specific purposes and didn't fulfill the general needs. With that being said, there is an actual need to have a tool that can be used as integrated tool to crawl, analyze and visualize Facebook.

2.3 Conclusion

The rapid growth of social media and increasing generation of massive amounts of user data, increasing number of studies are conducted to develop tools that allow for getting useful information. For the time being, OSNs have become more complex in-terms of data rate, so working with such data needs efficient techniques to make data analysis and processing possible. By reviewing the existing techniques, we have found that none of the explored matches the needs to crawl, analyze and visualize the Facebook, so a new tool should be developed to be a universal integrated tool.

Chapter 3 Analysis and Design of the Facebook Data Extraction Tool

3.1 Tool analysis: goals and objectives

Designing a universal tool to extract, analyze and visualize Facebook data is the main goal of our research. To achieve this goal, the tool design should take the below points into consideration:

1. Data extraction: our tool should have capabilities to fetch data out of Facebook public pages. The fetched data include posts, comments, likes, date-related information such as post date/time, comment date/time. Page general info such as number of followers, type of page (educational, commercial … etc.) should be fetched as well.
2. Analysis: the implemented tool should analyze the collected data to explore the hidden information. This information reflects the page peak time, user interaction rate, user active time, posts to comments relation, comments rate per post, post rate, comment rate, and page word cloud.
3. Visualization: It is the important phase since it will display to results to the users of our tool, so it should display the analyzed data in a clear way to achieve the goals of the tool. Data should be displayed in form of pie charts, scatter charts, bar charts, line charts and word cloud. These charts will support end-user decisions by providing measures regarding to page growth, user's interaction rate and peak time … etc.

3.2 Major components

Our proposed tool has three major components, data extractor, analysis and visualization components. Each one has its own functions.

1. Data extractor component: responsible for fetching data out of Facebook using Facebook graph API to get all page related data and store it in the database. This component applies multiple filters to clean up the data.
2. Analysis component: designed to apply different types of analysis on extracted data, this component takes the output of data extractor component as an input to calculate time related information, data rate information and word cloud.
3. Visualization component (real time component): designed to visualize the output of analysis component in readable form, pie charts, scatter charts, bar charts, line charts and word cloud.

3.3 Requirements

We proposed a model to allow a set of users to get useful data from Facebook. The proposed model shown in figure 1, is composed of multiple user interfaces as presented below:

1. Home interface: displays statistics which represent the number of crawled pages/ groups/ profiles, number of crawled posts/ comments, crawled posts per time and crawled comments per time.
2. Extraction interface: has inputs to data extractor component which are entry points (page, group and profile), Facebook id for the entry point and the limit (max number of nodes to extract).
3. Analytics home interface: displays a dropdown of crawled items to allow user to select which one to view.
4. Profile analytics interface: it’s the main interface which displays all profile related information, including number of crawled comments/ posts, comments rate per post, post/ comment rate per hour, post/ comment rate per month and word cloud.

The proposed model consists of the items below:

1. User should be able to navigate to home page and display tool statistics.
2. User should be able to navigate to extraction view and fill in the fields to extract the data.
3. The tool should get the target data based on the users input and returns response message for the user to determine the status of the extraction process.
4. Tool should apply data filtration processes to clean up the data and get rid of noise.
5. Calculating word occupancy and create word cloud.
6. Storing the information in database.
7. User should be able to navigate to analytics view and select which items to view.
8. Tool should fetch the target information from database and display it to the user.

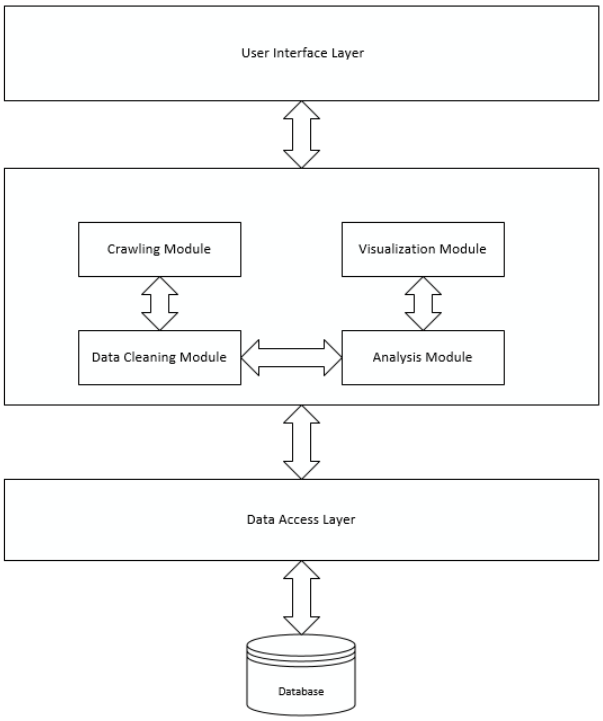

Figure 1: Proposed MVC architecture

3.3.1 Non-functional requirements

Non-functional requirements cover the specific criteria that judge the operation of a system, rather than specific behaviors.

Below is a list of all non-functional requirements:

1. Performance: this includes extraction/analysis time. The tool should utilize resource usage to guarantee an optimal performance. We used a key value pair data structures to enhance the post processing.
2. Scalability: the tool is designed to handle with the growth of Facebook data by using optimal algorithms for fetching/ analysis. Our tool doesn't buffer the data while crawling, it will fetch and insert into database after cleaning up the data.
3. Capacity: It defines the number of Facebook items that the tool can handle. It is limited to the database capacity in our design.
4. Availability: tool availability depends on handling the unexpected behavior to avoid exceptions. We rely on the testing phase to avoid such situations.
5. Recoverability: since our tool store the data in database, recovering information is valid from the existing backups.
6. Maintainability: the tool should be designed to be maintainable by adding software logging to allow us to track failures and find the proper solutions within short time.
7. Security: We used a recent java spring framework version to enhance the security.

3.4 Software design

In this section we introduce different kinds of diagrams that represent the overall system.

3.4.1 Structural models

Class diagram has been split to three sub diagrams, crawler, Analyzer and visualizer class diagrams. Figure 2 shows the crawler class diagram and its relations with other objects, it has interface to “HttpClient” in order to fetch Facebook data, “DBModel” will be used to store the crawled data in database. The remaining classes are used to manage the extracted data and to provide control over this module.

Data analysis is another module designed to manipulate the output of the crawler in order to retrieve the useful information. To accomplish its task, the analysis module has interfaces to “DBMode”, “StopWordsCmd” to get the data and generate the word cloud respectively. The rest of objects are being used as utilities to support this module. Figure 3 shows the analysis module and the relation with other objects.

Finally, we discuss the visualization module as shown in figure 4. It simply fetches the analyzed data from database and displays it to end users in the form of charts. Class diagrams developed referring to [32].

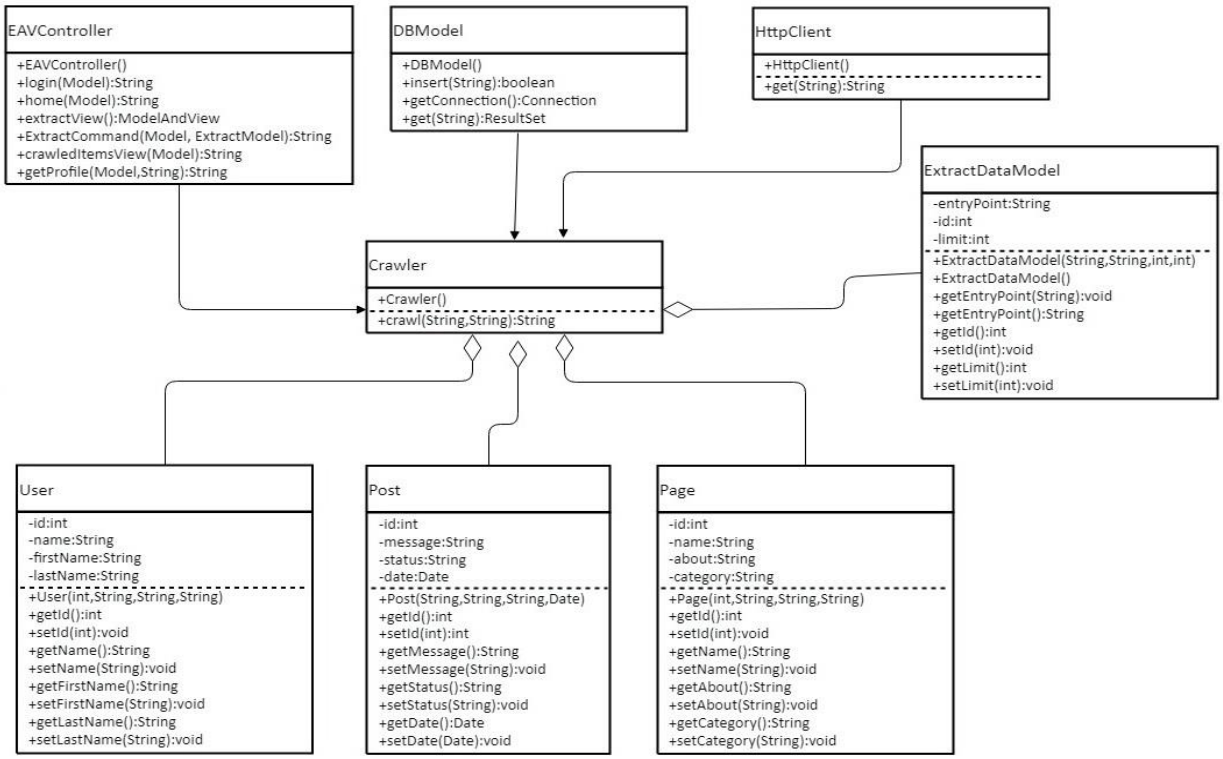

Figure 2: Crawler class diagram.

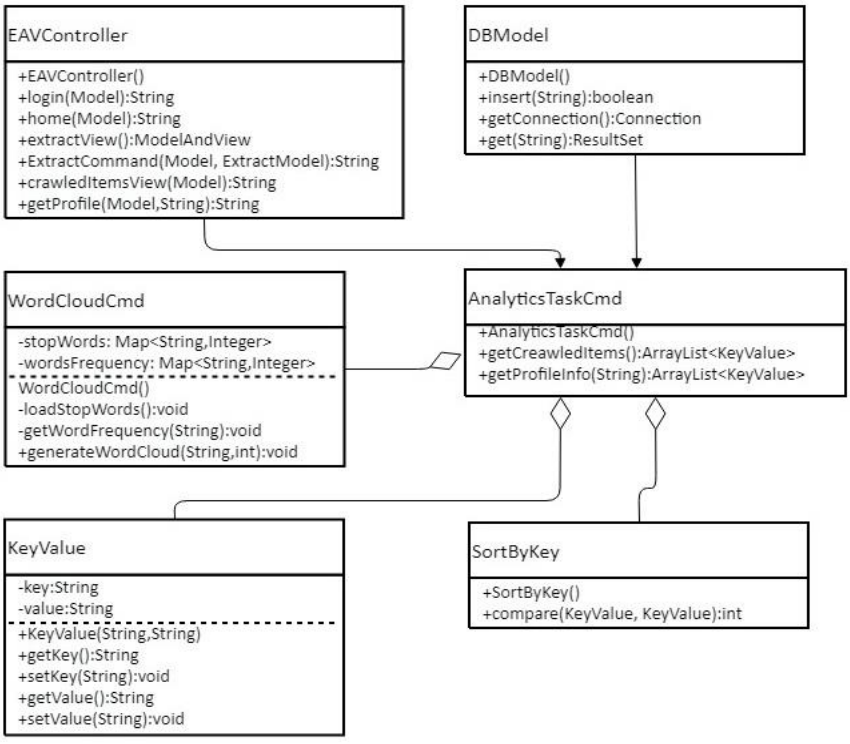

Figure 3: Analysis class diagram.

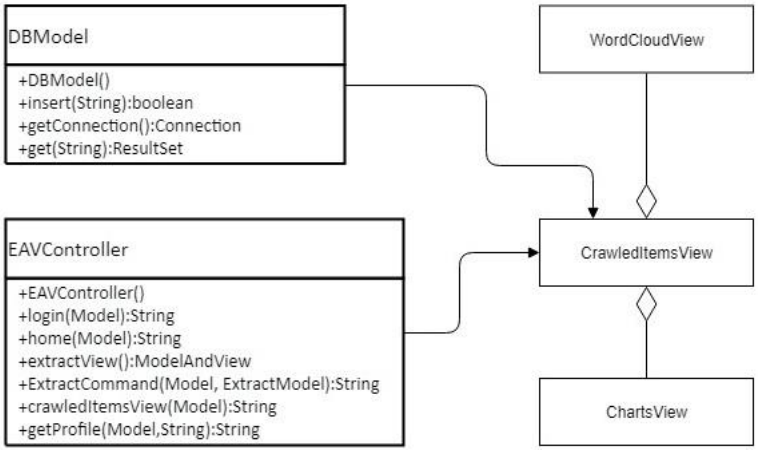

Figure 4: View class diagram.

3.4.2 Interaction models

In this section, we modeled both of use case diagrams and sequence diagrams depending on the information provided by [32].

3.4.2.1 Use case diagrams

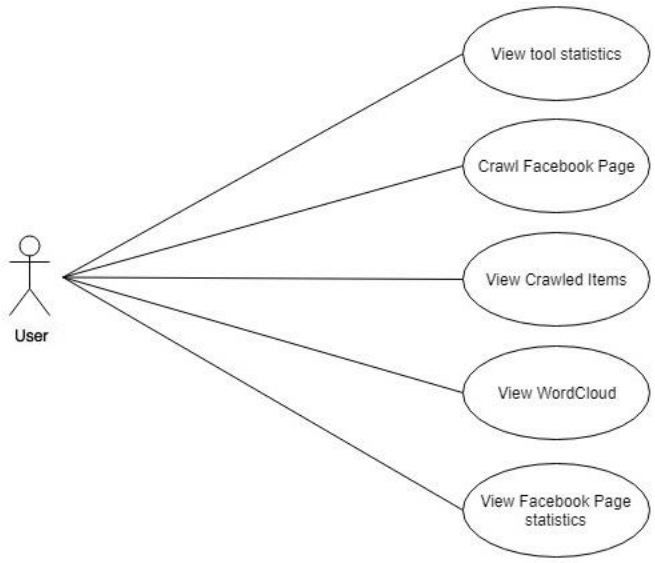

Figure 5: User use case.

Users will have the ability to crawl Facebook pages and view all related info. Including pages statistics, word cloud and tool statists as shown in figure 5.

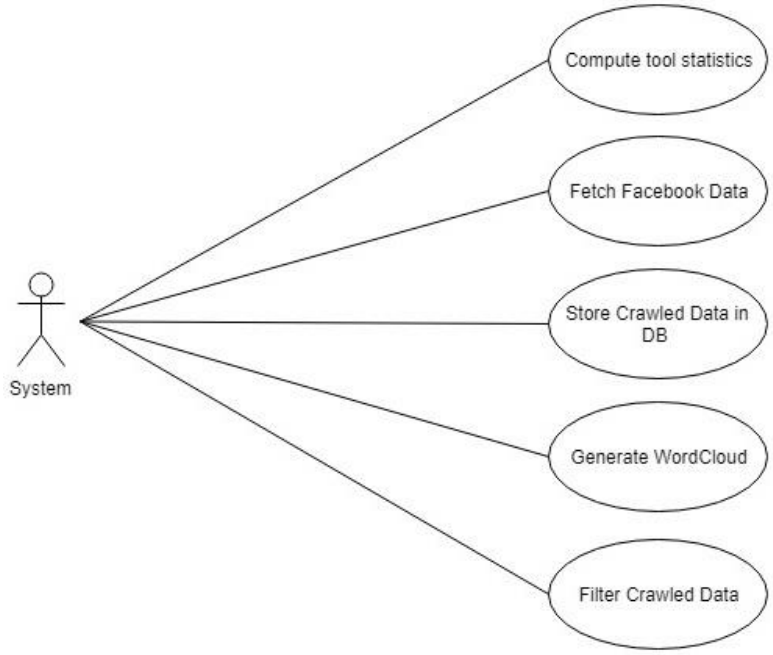

Figure 6: System use case.

Figure 6 above shows the system use cases, system refers to the tool's internal processes. The tool will compute the overall statics, fetch the data from Facebook, store the data in database, generate word cloud and apply the different analysis techniques.

3.4.2.2 Sequence diagrams

Figure 7 shows the interactions between crawler module and other objects during the crawling process where it mainly depends on "HttpClient" object to get the data based on the input from "EAVController".

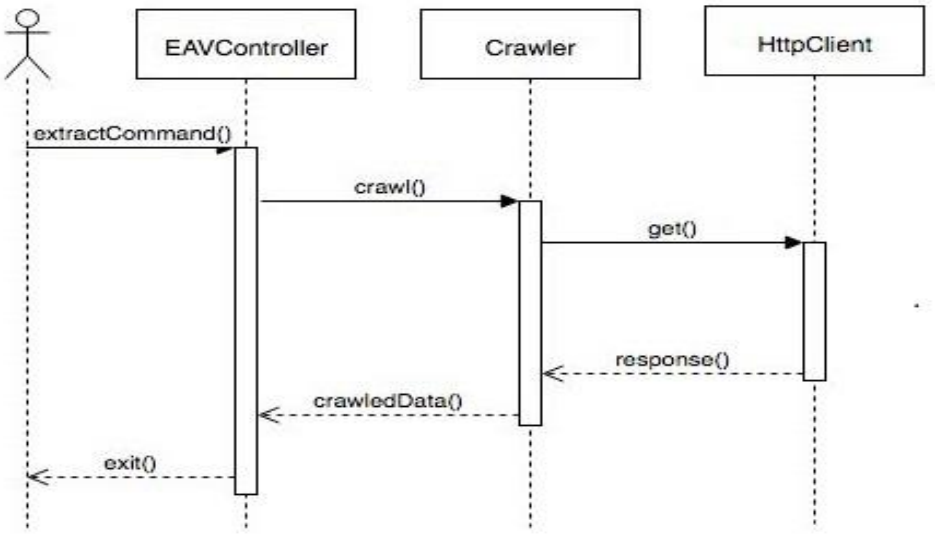

Figure 7: Crawling sequence diagram.

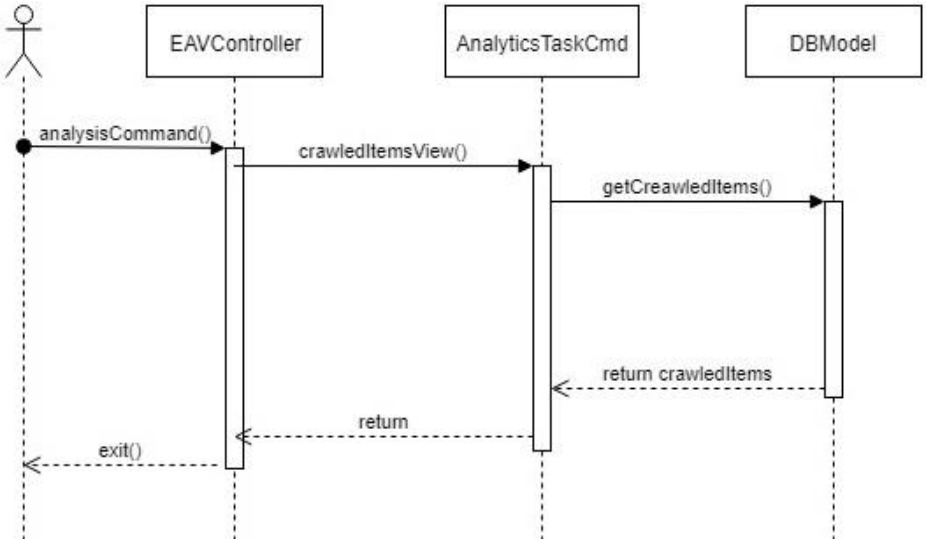

Figure 8: Analysis sequence diagram.

Figure 8 above shows the analysis module in action, where it communicates with the "DBModule" to fetch the data and apply the analysis algorithms.

The word cloud module generates the word cloud for each crawled page, it will fetch the data and output the words with the highest frequency (number of output words is configurable). Figure 9 shows the sequence diagram for word cloud module.

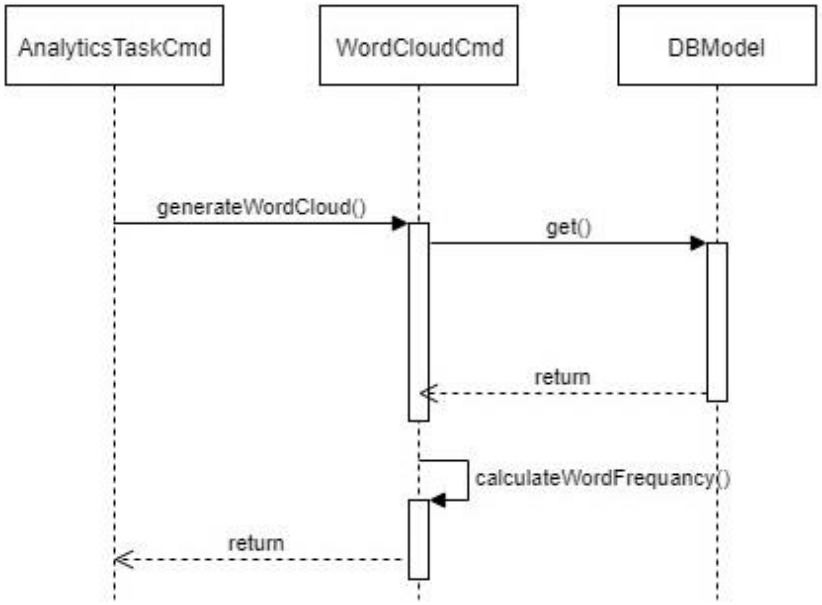

Figure 9: Word cloud sequence diagram.

Data cleanup is an important part of this tool in order to get a useful information. Before analysis process, data should be filtered to remove unnecessary characters and emojis. We designed a module that work simultaneously with crawler module as shown in figure 10.

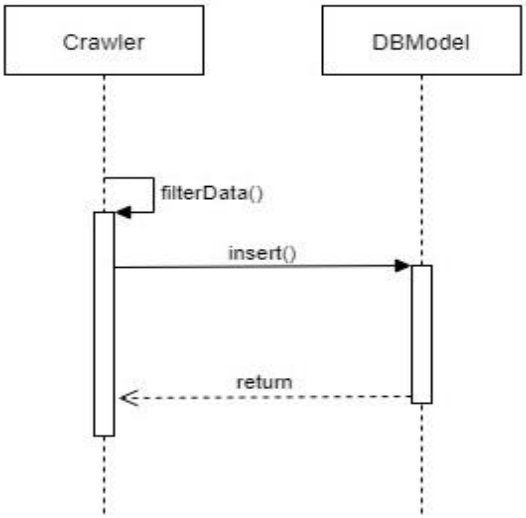

Figure 10: Data clean up sequence diagram.

3.4.3 Context models

Context diagrams are important to show the relationship between our system and the other external systems and to demonstrate the flow of the system. The following two sections will discuss the context and the activity models base on the instructions provided in [32].

3.4.3.1 Context model

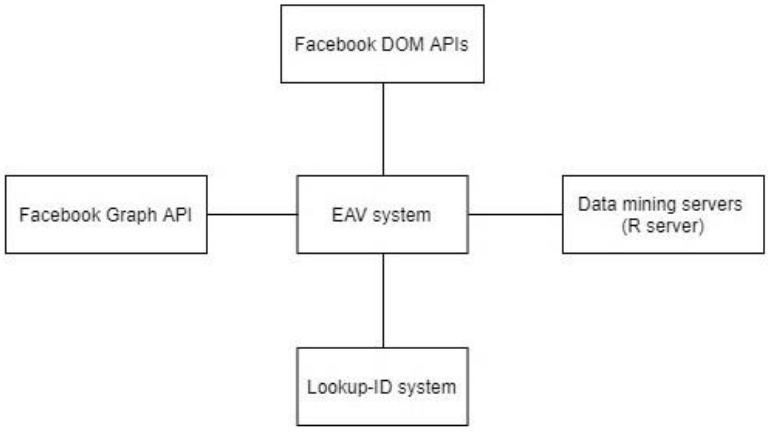

Figure 11: Context diagram.

Our tool has interactions with several systems. Graph API is the most important system to fetch the data where the remaining systems help in analyze the data such as Data mining servers. Lookup-ID system used to get the unique ID for Facebook pages. Some of DOM APIs are used to inspect the DOM of the Facebook pages. Figure 11 shows the context diagram.

3.4.3.2 Activity diagram

Considering a complete cycle of the tool, the process starts from submitting the Facebook page URL. The system will fetch the corresponding ID from lookup-ID system and if it is a valid page URL, the system will extract the data using the Facebook graph API.

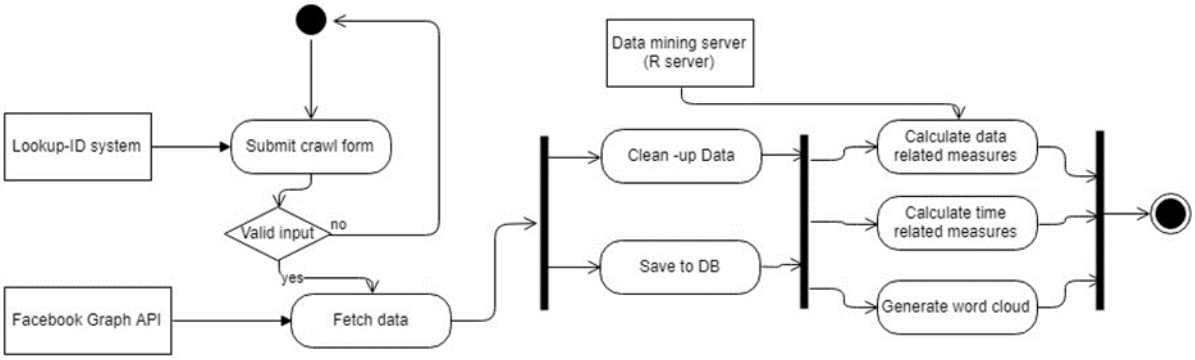

Figure 12: Activity diagram.

The system will clean up the data and save it to database, after that the system will analyze the data to get the data measures and time related measures. Figure 12 shows the activity diagram.

3.4.4 Complexity measurements

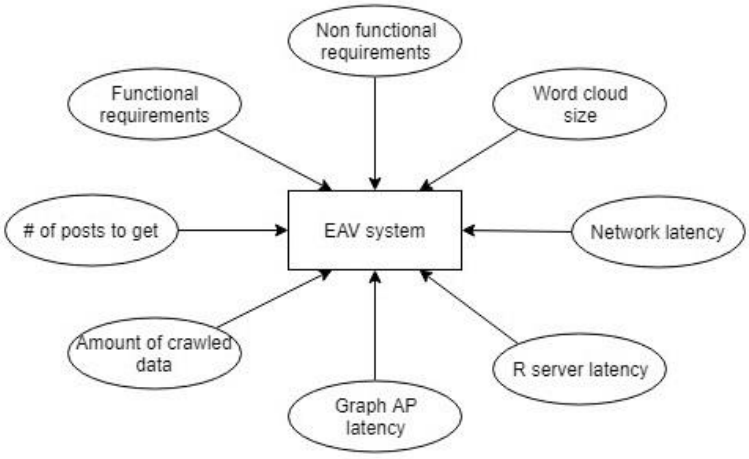

Figure 13: Complexity factors.

Software complexity describes how difficult to understand the software. In software engineering complexity should be measured from development point of view because time equals money. For example, the system that contains large amount of data to process will increase the complexity, figure 13 above shows the different complexity factors of the system.

3.5 Database design

We designed our own database to store the crawled data and to cache the analysis results, to achieve this we proposed a schema with five tables each one represents a unique object. “Post_tbl” will hold the posts related information such as the number of likes, publish date, the content message and the type of the post. This table has a relation with each table in the schema.

“Comment_tbl” represents the comments, so it contains the content message and the creation time, this table has a relation with the “Post_tbl”. Each comment must belong to a single post and the post may have zero or more comments.

“Page_tbl” represents the Facebook page, having the name, likes, followers, id, about and the word cloud. This table has a relation with “Post_tbl”.

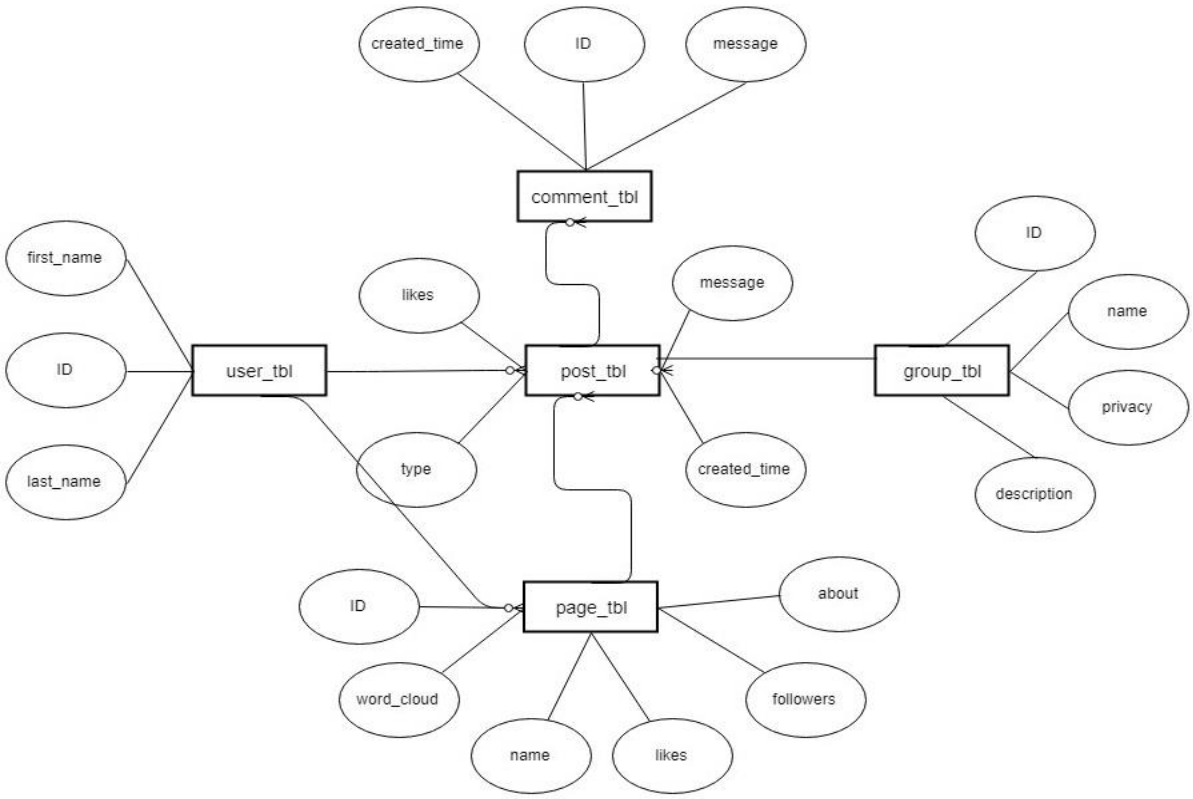

Figure 14: ER diagram.

Each post must have a published in (page, group or profile) and each page may have zero or more posts. “word-cloud” attribute will be used to cache the output of “WordCloudCmd” to improve the performance. “user_tbl” represents Facebook user and store its id, first name and last name. The user may belong to a page, group and can add a post or comment. We used this table to track the posts and the comments with the page since we didn’t design our tool to crawl Facebook user profiles.

“Group_tbl” is the last table in the schema, we used it to store group related info such as id, name, privacy and the description. Our tool has a limitation to access Facebook groups, using Graph API since the groups are private and to get the data, our tool (as Facebook user) should

join the group in order to extract its data. The group may has zero or more posts. Figure 14 above show the database ERD based on the steps mentioned in [33].

3.6 Software interface

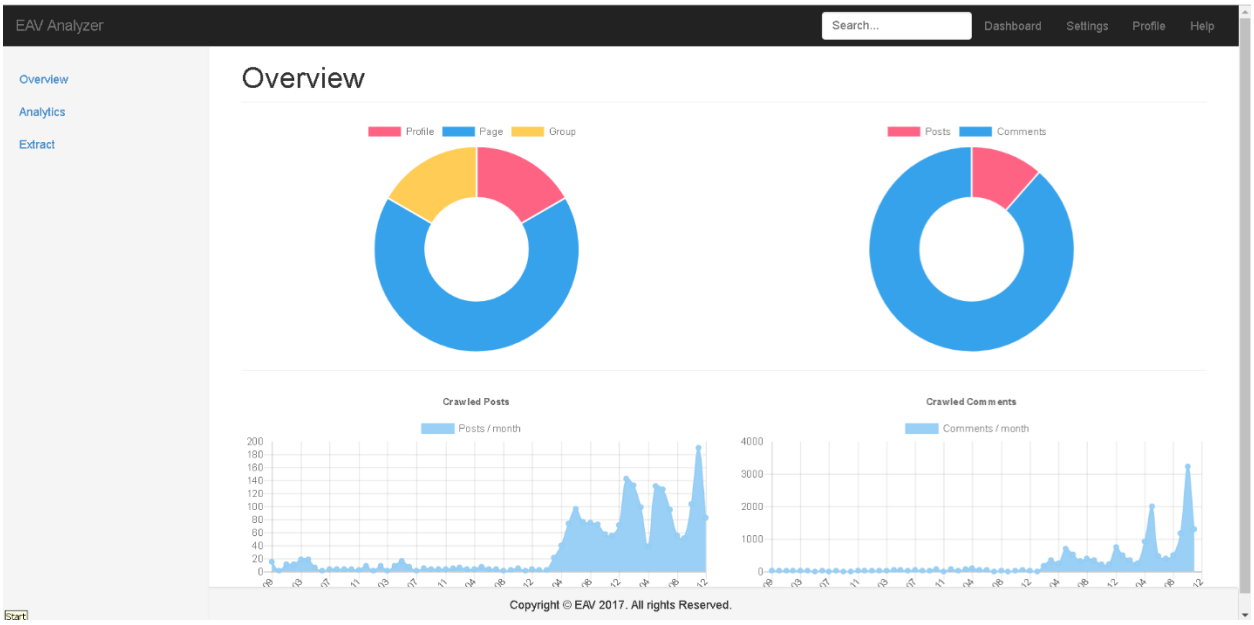

Figure 15: home view

Figure 15 above shows tool home view, where general tool measures are shown to indicate the number of Facebook crawled profiles, pages and groups. Also it shows the number of posts to the number of comments, data rate per time.

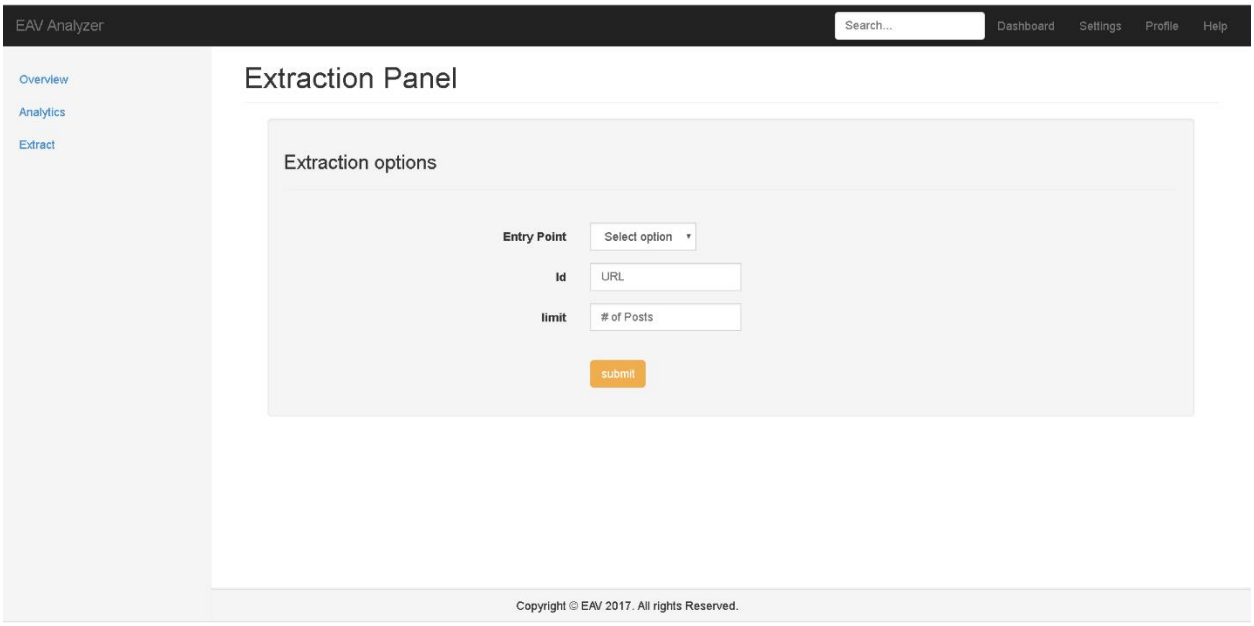

Figure 16: Extract view.

Extraction view takes user input to extract the data, these inputs include entry point (page, group and user profile), Facebook Id of the entry point and the limit (max number of posts to retrieve).

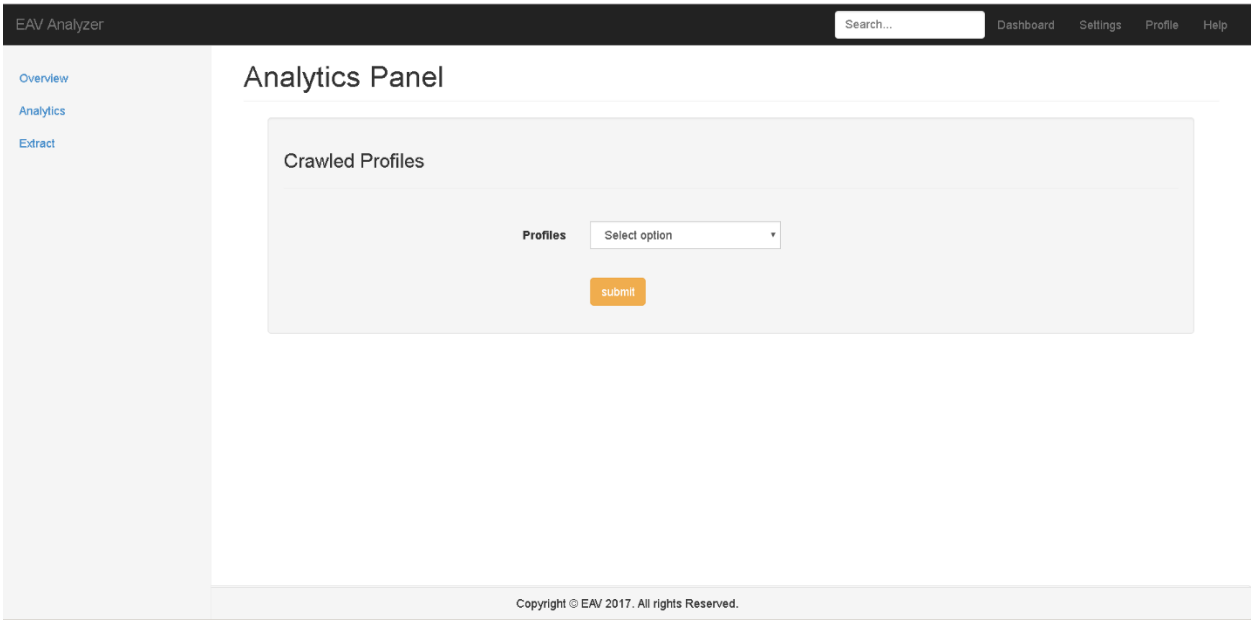

Figure 17: Crawled items panel.

Crawled items panel allows users to select which profile to analyze, so the dropdown in the figure above (figure 17) will display the previously crawled items.

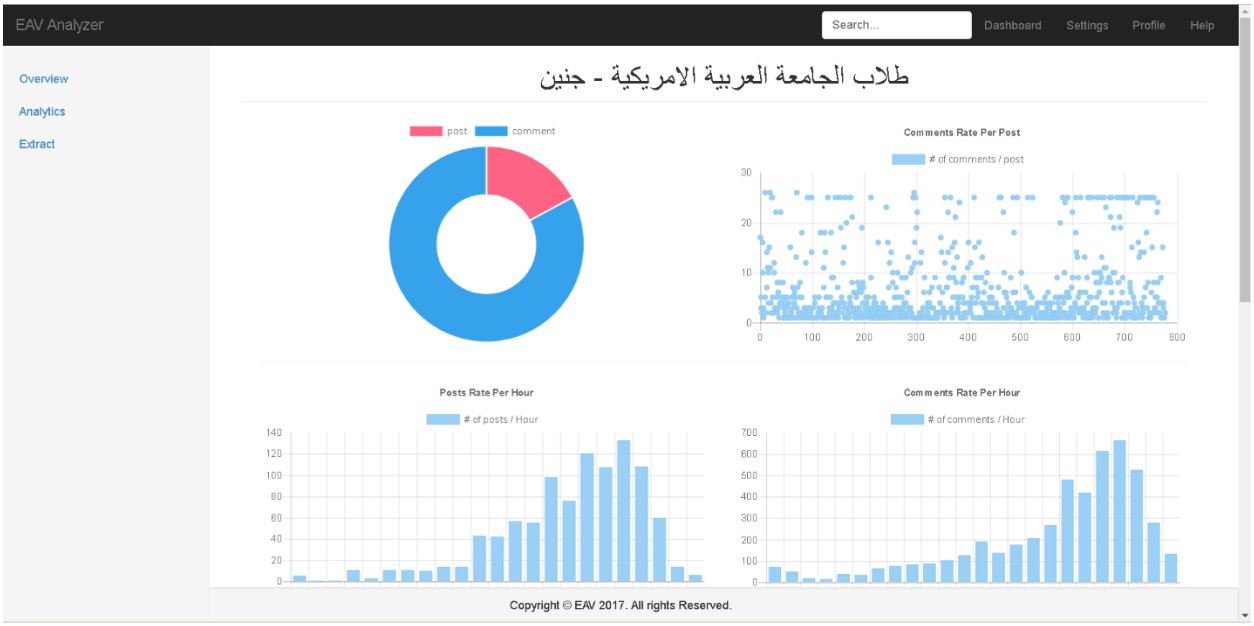

Figure 18: Profile view.

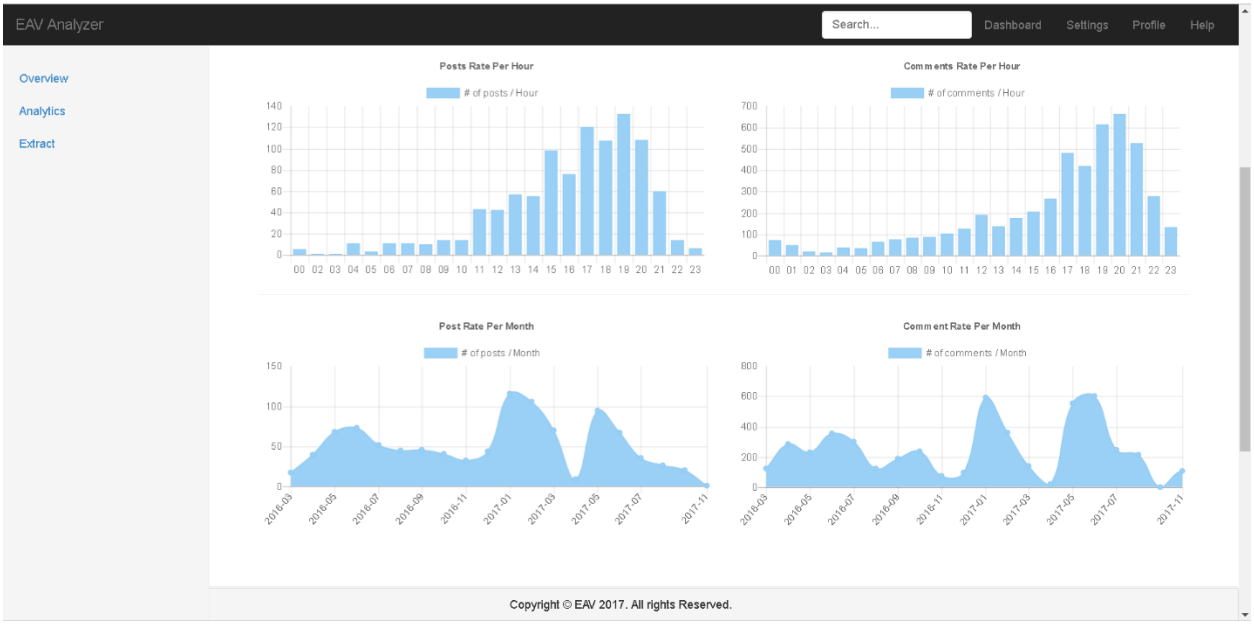

Figure 19: Profile view.

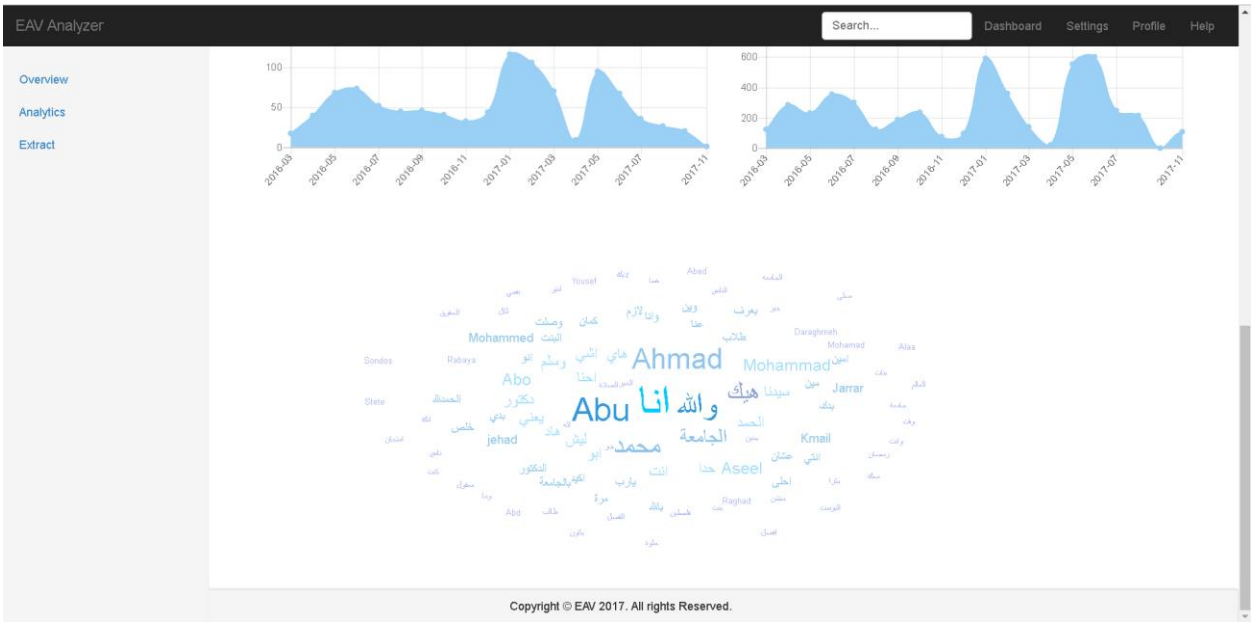

Figure 20: Profile view.

Figures 18 -20 shows different types of profile measures, it shows the number of crawled posts/comments, comments rate per post, comments/posts rate per hour, comments/posts rate per month and word cloud.

3.7 Software testing and validation

Software testing have two main goals, to demonstrate that the output system matches the requirement and to discover potential errors and incorrect behavior.

Software testing can be done on different levels:

- Unit testing: this type of testing carried out by the developers who are developing the software system and in some cases of development processes a pair of developer/tester working together to discover the bugs or defects.
- Release testing: a separate testing team tests the system before releasing it to the end users to ensure requirement matching.
- User testing: users or individuals of users test the system.

We use the first two types of development testing in our project to test the system and requirement matching.

Chapter 4 Data collection, analysis and visualization

4.1 Facebook data collection using developed tool

As mentioned previously, many tools have been proposed to fetch data out of Facebook. Each of these tools focus on specific areas such as fetching friends list, social trends and commercial competitors. In our tool, we focused on fetching data out of Facebook pages to be used as a data source by any analysis techniques.

Fetching Facebook page data using our developed tool is a simple process. The user should provide the page ID and the number of posts to get and the tool will handle everything, starting from page general info to posts and their comments. After this process, the user will get a database containing all page-related information.

In our experiments, we ran the tool over different categories of pages to compare between them:

- Educational pages: we chose three Palestinian university pages on Facebook to run this experiment. The first one was the page of "Arab American University / الجامعة العربية الأمريكية"[2]. Our tool retrieved 2787 post and 16866 comments covering the interval from March 2011 to April 2018. Similar to the previous page, we obtained 4641 post and 43099 comment from "Birzeit University | جامعة بيرزيت"[3] page covering the interval from July 2010 to April 2014.The last page was "جامعة النجاح الوطنية - An-Najah National University"[4]. We got 5141 post and 23102 comment covering the interval from February 2011 to April 2018.

[2] https://www.facebook.com/aauj.pr
[3] https://www.facebook.com/pg/BirzeitUniversity
[4] https://www.facebook.com/ANajahUni

- News pages: a sample of two pages were selected to conduct this experiment, “Ramallah News - رام الله الإخباري”[5] was the first one, where the tool crawled 4590 post and 67204 comments covering the interval from November 2011 to April 2018.

  The second page was, “Shehab News Agency” [6] with a total number of 4667 post and 54076 comments covering the interval from March 2011 to April 2018.

- Commercial pages: this experiment was conducted on two commercial pages, “Jawwal”[7] where the tool crawled 4486 post and 84587 comment covering the interval from November 2011 to April 2018.

  The second page was “Wataniya Mobile Palestine”[8] where the total number 3441 post and 50207 comment covering the period from March 2008 to April 2018.

4.2 Facebook Data Extraction and Visualization

Referring to the tool statistics during this experiment, a total number of 30072 posts and 343121comments have been crawled, so the number of comments around eleven times greater than the number of posts as shown in figure 21. This number indicates that the crawled pages have an average number of eleven comments per post which reflects the user interaction rate. So if we have a page with a comment rate over than eleven, then we can judge that the page has a good user interaction rate relative to the crawled pages.

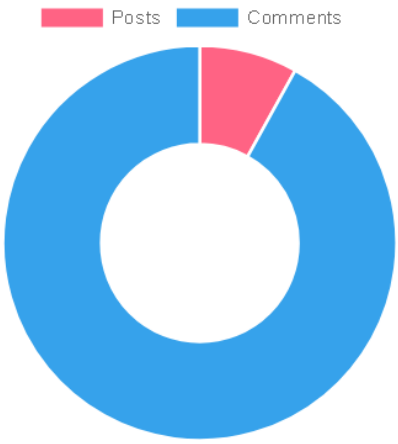

Figure 21: Tool statistics, percentage of posts to comments.

[5] https://www.facebook.com/RamallahNewsOfficialPage
[6] https://www.facebook.com/ShehabAgency.MainPage
[7] https://www.facebook.com/Jawwal.059
[8] https://www.facebook.com/WataniyaMobilePalestine

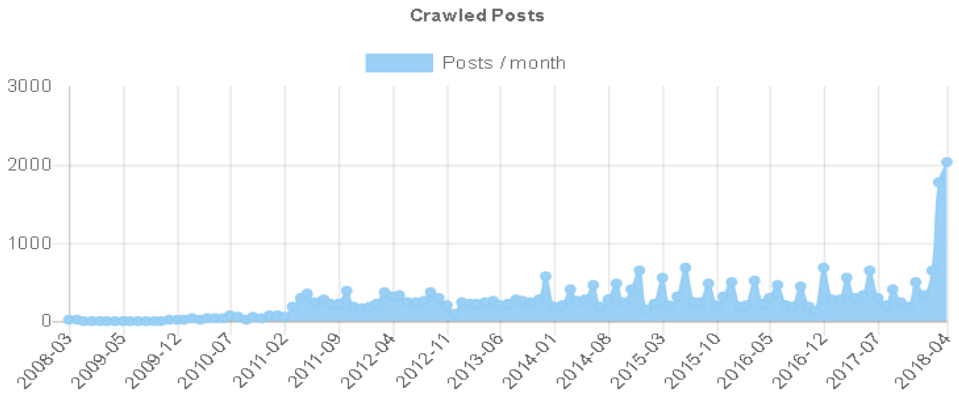

Figure 22: Tool statistics, number of crawled posts per month

Crawled data from all Facebook pages cover the period from March 2008 to April 2018. Figure 22 shows the number of crawled posts per month, where the number of posts keep growing from 22 posts in March 2008 to reach around 2027 posts in April 2018.

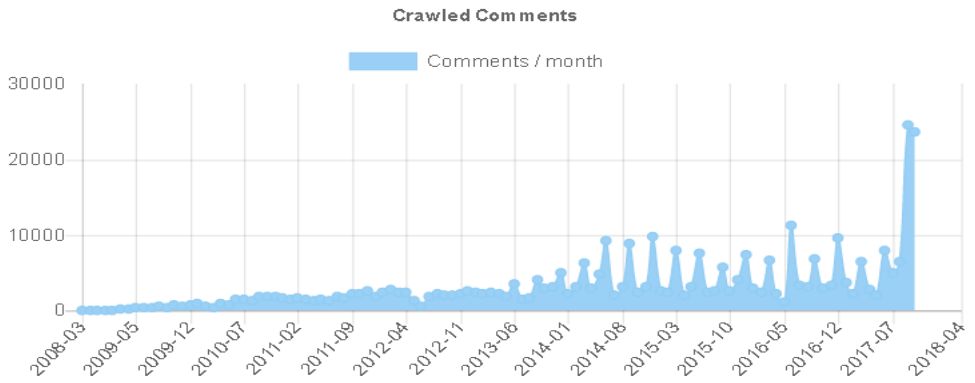

Figure 23: Tool statistics, number of crawled comments per month.

Figure 23, shows that the number of comments grew up from ten comments in March 2017 to 23682 comment in October 2017. In the following sections, we'll discuss each group of crawled pages separately.

4.2.1 Educational Pages

Three pages have been crawled to conduct this part of our experiment. Starting from the page of the Arab American University, we found that the page has around an average number of six comments per post referring to figure 24 and figure 25.

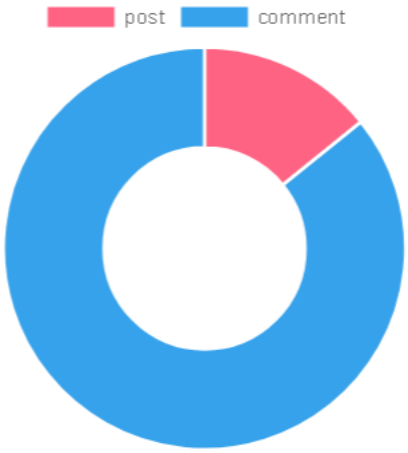

Figure 24: Percentage of posts to comments.

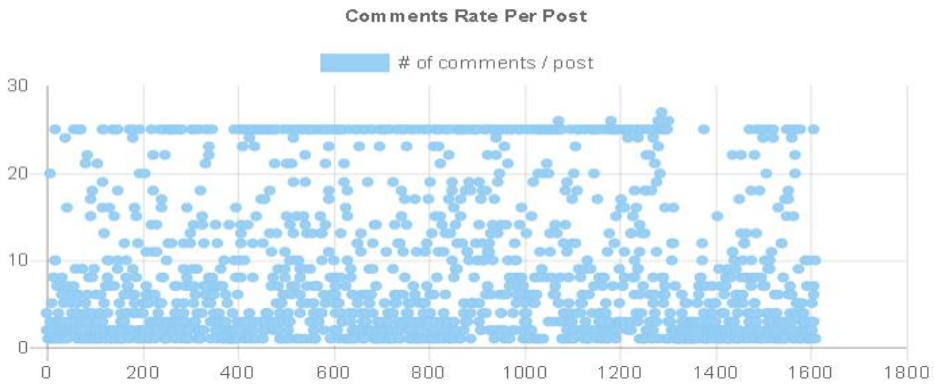

Figure 25: Comment rate per Posts.

The majority of posts have a comments number less than 5. Some of them have around 25 comments and the rest of posts have a number of comments between them, so based on that the users interact with a few number of pots.

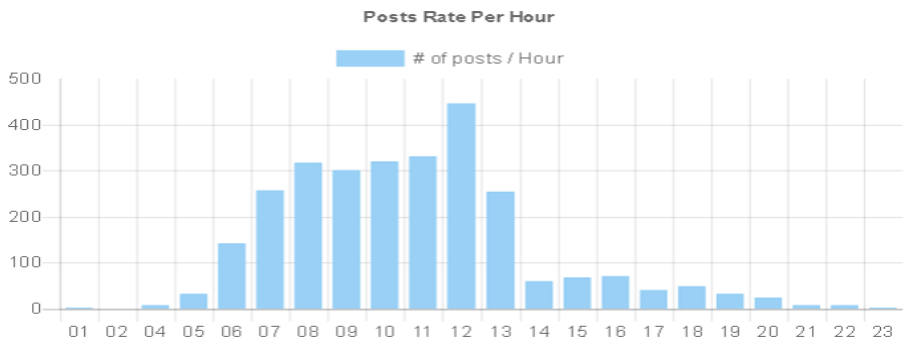

Figure 26: Post rate per hour.

Figure 26 shows post rate per hour, where the majority of posts have been posted from 7 AM to 1 PM (12 PM is the peak posting time).

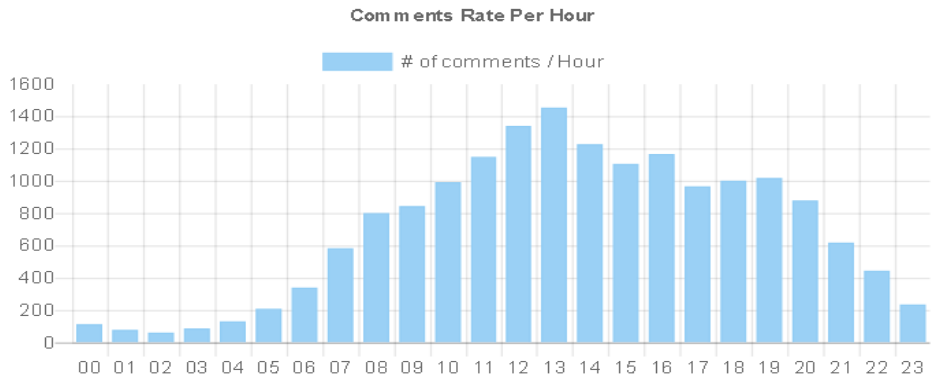

Figure 27: Comments rate per hour.

Figure 27 shows the comments rate per hour. We noticed that the comments rate increases from 2 AM until 1 PM and decreases after that. Commenting rate reached the peak at 1 PM during students' break time where all students don't have any lectures. So it is reasonable to have a peak time at this point of time.

The second half of a day has commenting rate greater than the first half, since the students had finished their lectures and returned back home. There is a noticeable threshold between 4 PM and 5 PM and this change in commenting rate can be explained relative to the time when a

regular university day ends and students are back home. So most of them have to use transportation and sometimes wait to get a taxi since all students finish at the same time.

Comparing the posting rate with commenting rate, we didn't find a direct relation between them. Most of posts have been posted in the first half of the day and comments increased at the second half.

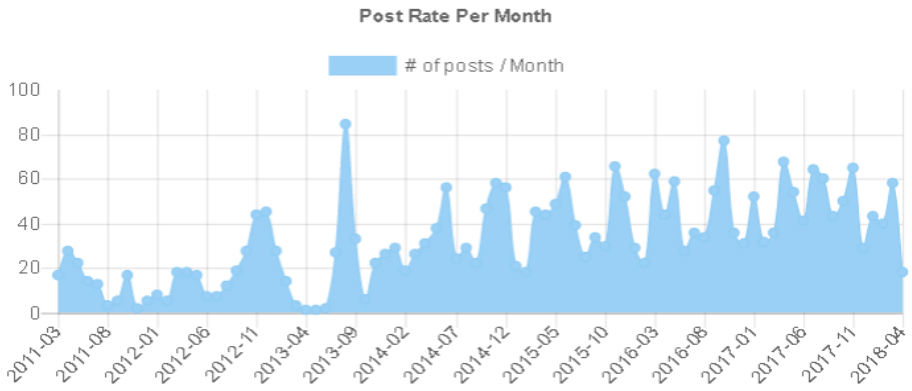

Figure 28: Post rate per month.

Figure 28 shows the posts rate per month. It is noticeable that the posting rate has improved starting from April 2014 until now. We have a peak value in August 2013. Returning back to that date we found several conjugate occasions (annual graduation party, open new master programs and Ramadan) and the increase in posting rate was to cover these occasions. Regarding the commenting rate per month, users' interest got increased by the time.

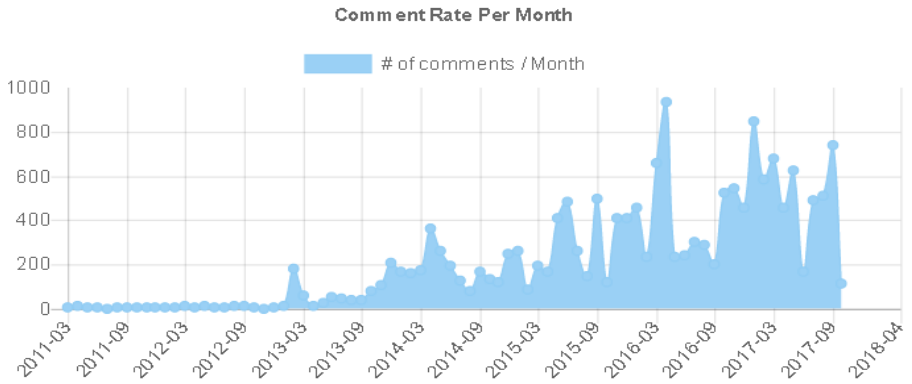

Figure 29: Comment rate per month.

Figure 29 shows the relation between comments and months. In first three years, users didn't show interest in this page but after that the interest increased significantly.

Figure 30 shows the word cloud for the Arab American University page, words such as ("الجامعة", "العربية", "الفلسطينية", "الامريكية", and others) are frequently used by users of this page. This provides an overview regarding the content of this page.

Figure 30: Arab American University Word Cloud.

The second page was An-Najah National University. Our tool crawled the posts and comments starting from February 2011 until April 2018. Referring to page corresponding statistics which are shown in figure 31 and figure 32, we found that the average number of comments per post is around four.

Comparing with the page of the Arab American University, the comments rate is lower and this indicates that the user is less interested in this page.

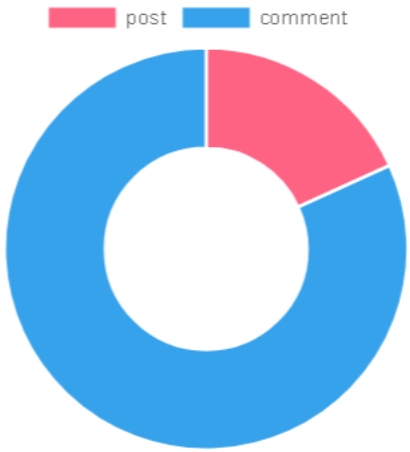

Figure 31: Percentage of comments to posts.

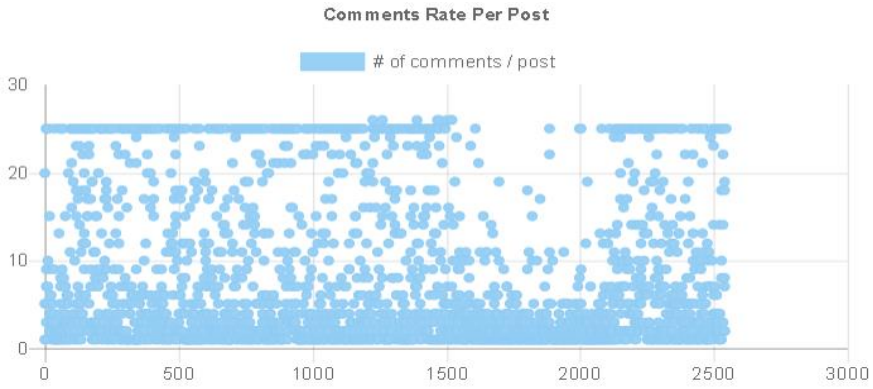

Figure 32: Comments rate per post.

Figure 32 shows that the majority of comments between one and five. Certain posts have around 25 comments.

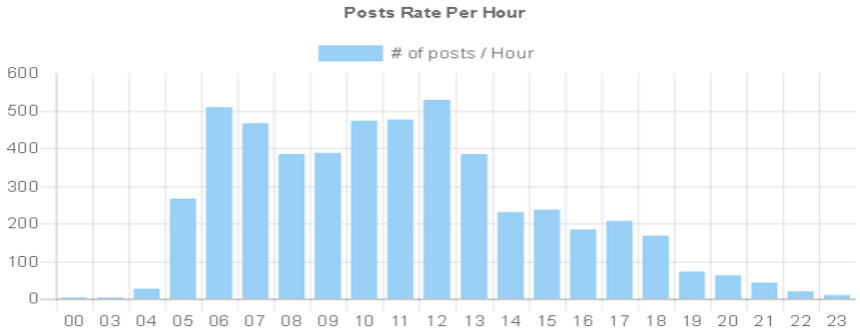

Figure 33: Posts rate per hour.

Posting to the page increased at 5 AM and kept in the range from 400 to 500 until 1 PM. At that point, posting rate started decreasing to reach the lowest level at 12 AM. Figure 33 shows the relation between posts and the hour.

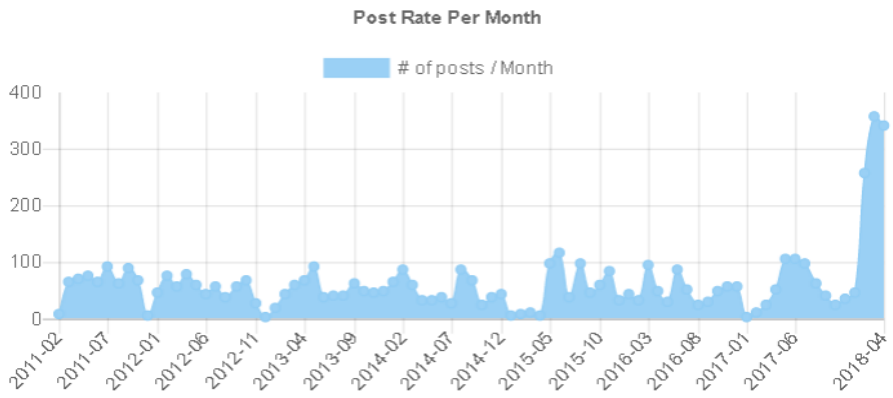

Figure 34: Posts rate per month.

Figure 34 shows the posts rate per month starting from February 2011 to April 2018, posting rate has a value up to 100 post per month except the first few months in April 2018 where the posting rate was improved to reach around 350 posts in April 2018.

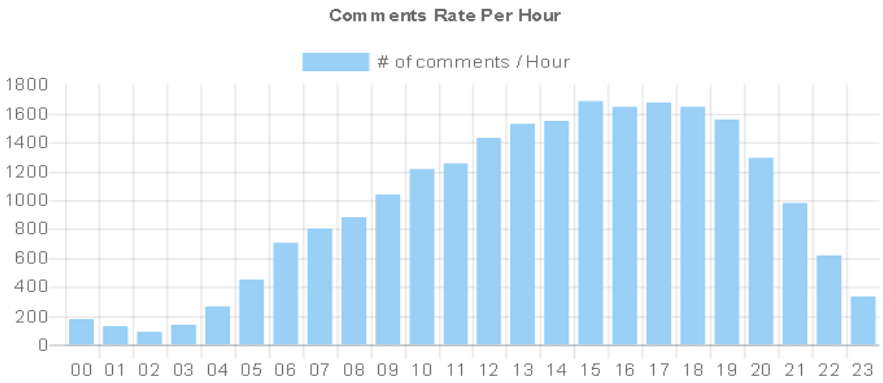

Figure 35: Comments rate per hour.

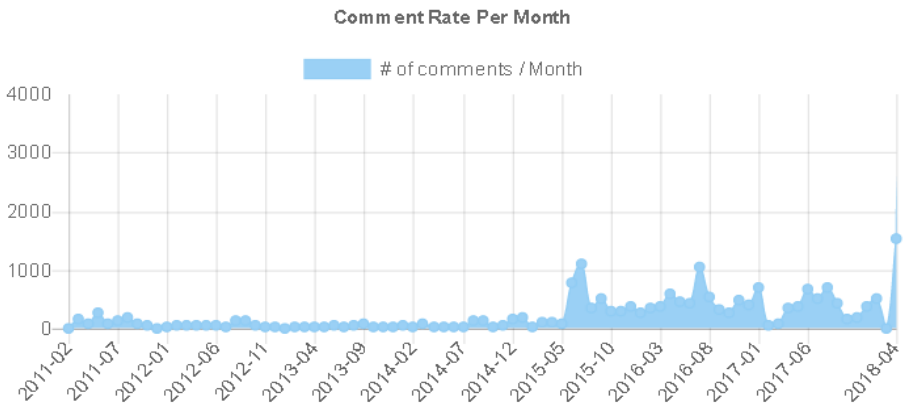

Figure 36: Comments rate per month.

Commenting rate increased linearly starting from 2 AM reaching the peak time at 5 PM and returning to decrease after that. Figure 35 shows the relation between number of comments and the month. This indicates that the students/users interact with the page after returning to home.

Users didn't show a high level of interest until May 2015 since the comment rate is too low. After that date the users became more interested and interact with the page. Figure 36 shows the comment rate per month.

Figure 37 shows the word cloud for An-Najah National University page, here some of frequently used words ("النجاح", "جامعة", "علاقات", "عامة", "دكتور", "الاحترام", "الوطنية"). These words provide an indication regarding the content of this page where the users are interested in the university, lectures and projects.

Figure 37: An-Najah National University word cloud.

The final educational page we use to run our experiment was the page of Birzeit University from July 2010 to April 2018. Referring to figure 38 and figure 39, Birzeit page has around nine comments per post.

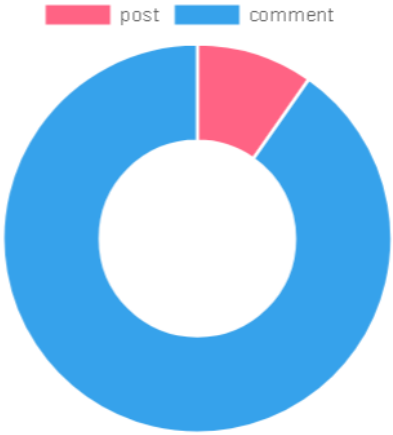

Figure 38: Percentage of comments to posts.

Most of the posts either have a high commenting rate (around 25) or a few comments which indicates that the users/students are highly interested in this page but for some reason, certain posts didn't have high user interaction rate.

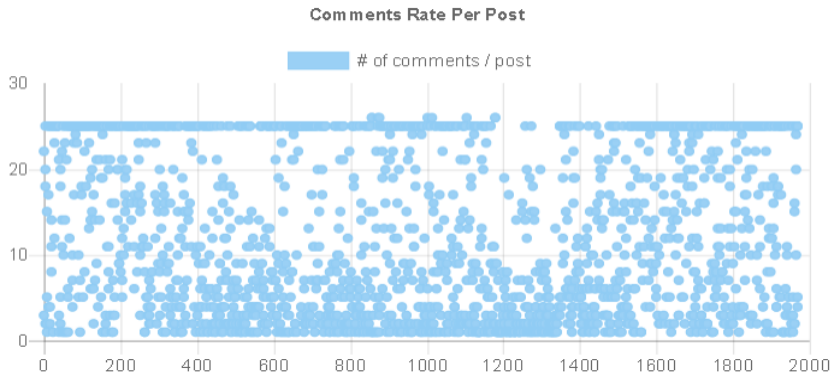

Figure 39: Comments rate per post.

Birzeit page has the highest interaction rate relative to other pages. Figure 40 shows post rate per hour where the number of posts started to increase from 267 posts at 6 AM to reach the maximum of 469 posts at 12 PM, after that posts rate are around 200. Unlike other pages, this

page has a high posting rate after the university closes its doors at the end of each day, since the users are allowed to post on this page.

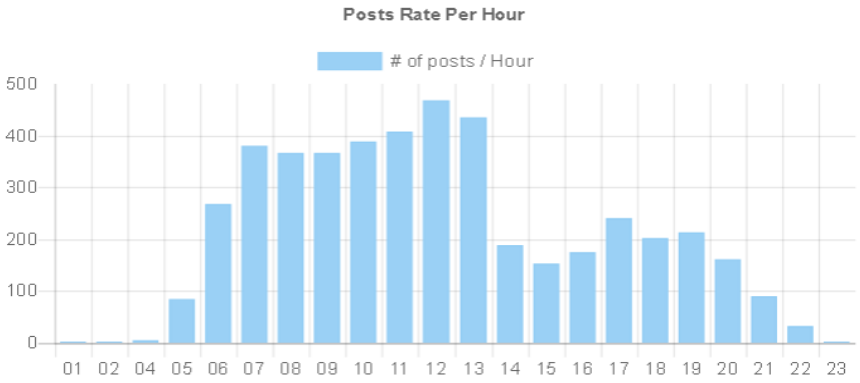

Figure 40: Post rate per hour.

Commenting rate doesn't increase linearly to follow the posting rate. While the posting rate increases from 7 AM to reach the peak at 12 PM, the commenting rate increased until 12 PM and stayed high despite the posting rate going down.

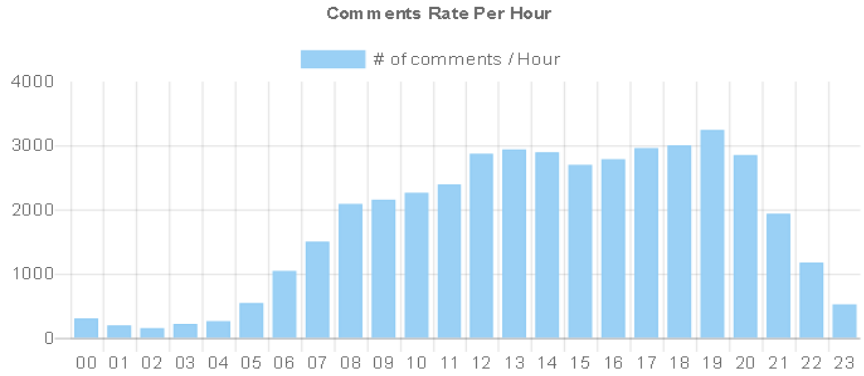

Figure 41: Comment rate per hour.

Most of users are active between 12 PM and 8 PM but there is a noticeable threshold after 2 PM since the students have to finish the remaining lectures before getting back home. Figure 41 shows the relation between the number of comments and the specific time during the day.

Figure 42 illustrates the relation between number of posts and month where the posting on this page started on July 2010 with around 20 post reaching the maximum of 154 post in March 2018.

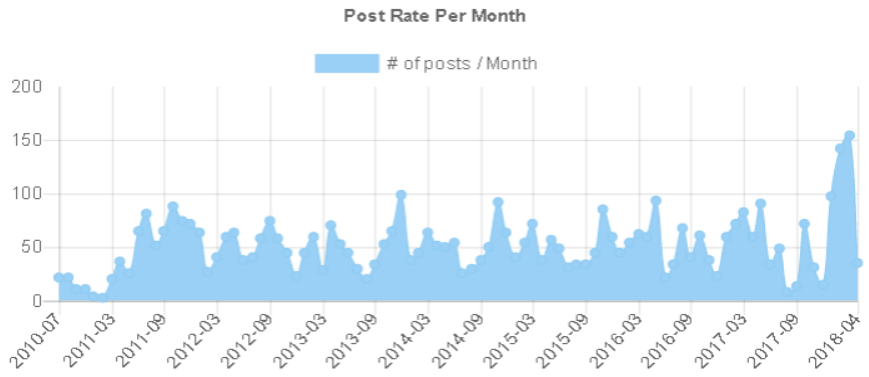

Figure 42: Posts rate per month.

During the months, the page kept a posting rate around 50 except the last year, whereas commenting rate kept growing from a few comments to around 1500 comments per month.

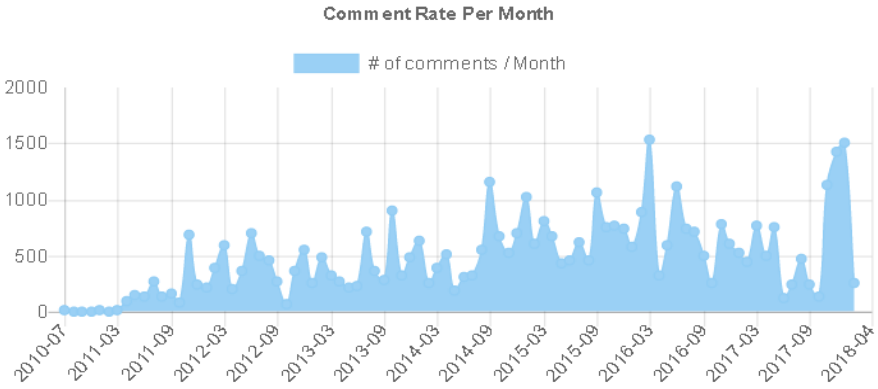

Figure 43: Comments rate per month.

Returning to figure 43, we noticed a peak commenting rate on March 2016. This can be explained if we review the content of the posts on that date which focuses on student kidnapping, master events and some TV programs.

Figure 44 shows the word cloud for this page and it is noticeable that the users are taking abut (“الجامعة”, “بيرزيت”, “الطلبة”, “الفلسطينية”, “العمل”, “الاحتلال”, “ الجامعات”, “دائرة”, “العلوم”).

Figure 44: Birzeit word cloud.

We discussed three educational Facebook pages as part from our experiment. Some of them have a higher interaction rate than others where Birzeit page with around nine comments per post have the highest rate, reviewing other metrics such as page growth, all pages have had a significant growth in the last few years.

Finally, each group of users have their own characteristics that identify their behavior, so the students free time, travel time, study time can be easily determined from the corresponding charts.

Our tool will help in studying student's behavior and find relations between grades and students online time, publish the important posts when at peak time, how the students spend free time … etc.

4.2.2 News Pages

In this section we present results from studying two news pages. The first one was Ramallah news page with an average number of 15 comments per post. The page has a good user interaction rate as shown in figure 45 and figure 46.

The number of comments is normally distributed between zero and 25 as shown in figure 46, so the posts have a level of acceptance among the users.

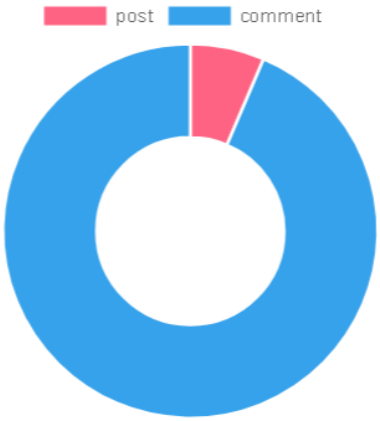

Figure 45: Percentage of comments to posts.

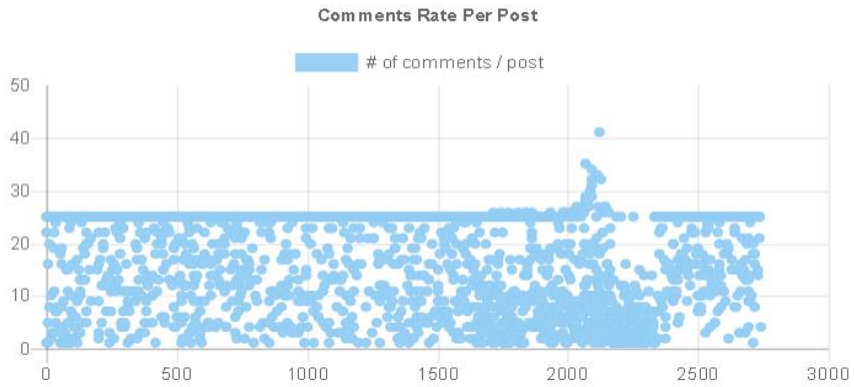

Figure 46: Comments rate per post.

Posting rate started to increase from around 80 posts at 5 AM reaching around 450 posts at 8 PM. The posts are distributed during the day to cover all news in a timely manner. Figure 47 shows the relation between number of posts and the hour.

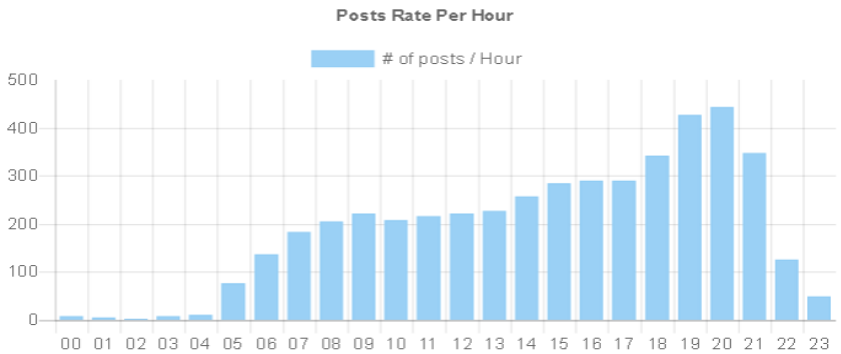

Figure 47: Posts rate per hour.

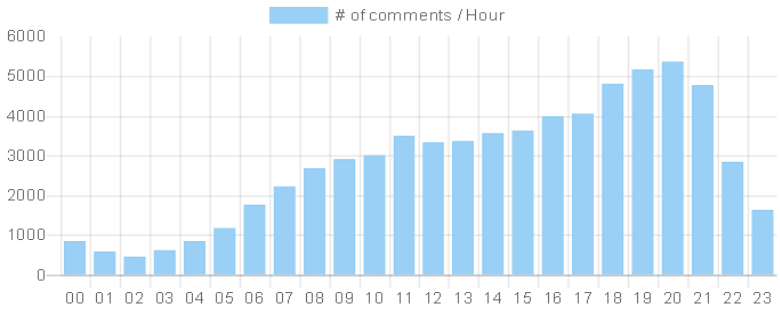

Figure 48: Comment rate per hour.

Similar to the posting rate, commenting rate increased to cope with posts. So the audience of this page didn't have restrictions on accessing Facebook during the day unlike educational pages, where the users/students had to enroll their lectures. The audience of news pages may be employees, unemployed or even retired. Figure 48 shows the relation between the number of comments and the time.

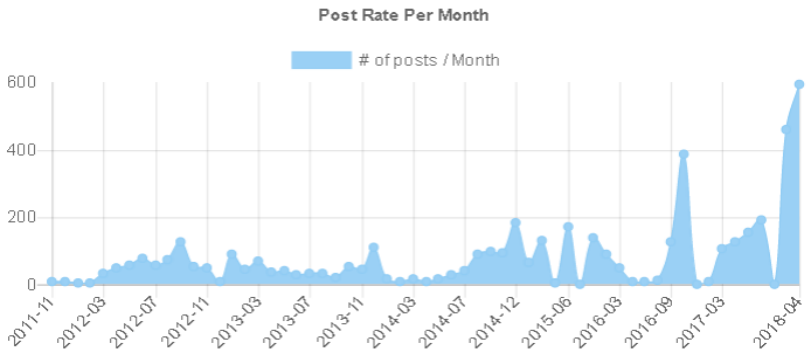

Figure 49: Posts rate per month.

Commenting rate reached a peak value between 6 PM and 9 PM. This can be understood as a time when all employees end their business day and return back home. This is helpful to

understand employees' behavior during the day. Commenting rate increased from 8 AM when most of employees start their work day. Before that some of them had to use transportation and this led to traffic jam at the beginning of day. Similar to that, another traffic jam occurred at the end of day.

The page has grown in a non-linear relation, starting from few posts at the end of 2011 reaching around 600 in April 2018. There is a noticeable edge in December 2016. In this period most of posts are focused on Christmas. Figure 49 shows the relation between the number of posts and the month.

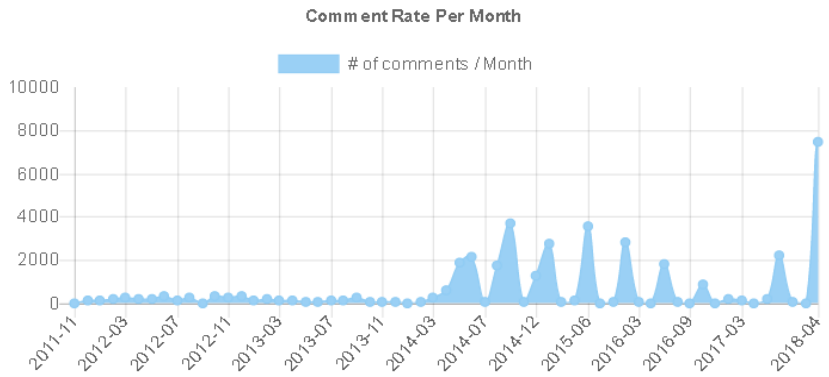

Figure 50: Comments rate per month.

Apparently, the users are interested in a few number of posts. These posts are talking about Palestinian cause in general. Figure 50 illustrates the relation between the number of comments and the month.

Figure 51 shows the word cloud for Ramallah news page where users are taking about Palestine, Jerusalem, Arabs and Jews.

Figure 51: Ramallah news word cloud.

The second crawled page was Shehab news page with around eleven comments per post as shown in figure 52 and figure 53. The user interaction was less than the previous page due to the content of posts so the users are less interested in this type of posts.

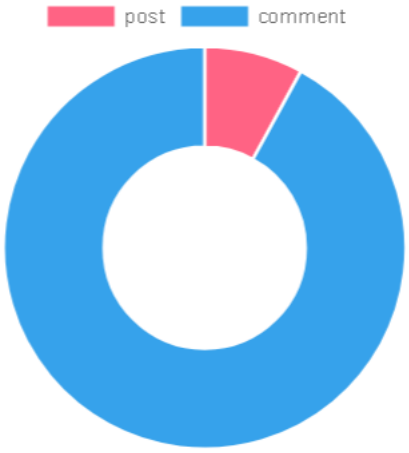

Figure 52: Percentage of comments to posts.

Most of posts had a number of comments between zero and 25 except for a few comments over than 25 comments per post. Figure 53 shows the relation between the number of comments per post.

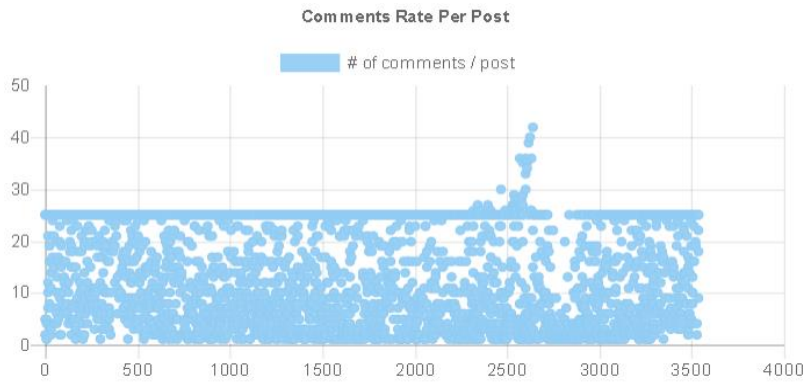

Figure 53: Number of comments per post.

Referring to figure 54, the number of posts increased from around 100 posts at 5 AM to reach the peak value of 420 at 1 PM and started to decrease after that. Most of posts are between 12 PM to 3 PM unlike the previous page where the posts increased from the beginning to the end of the day.

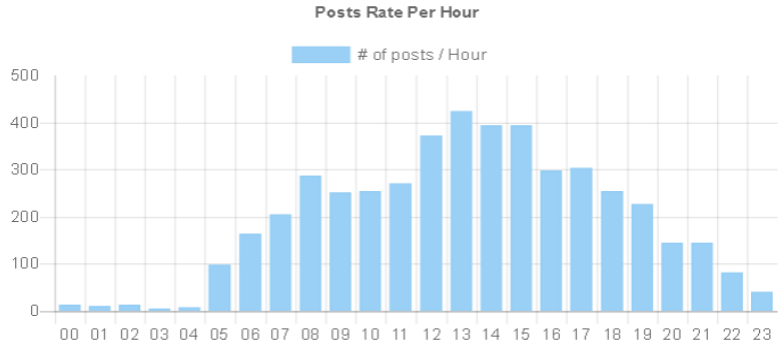

Figure 54: Post rate per hour.

Based on figure 55, commenting rate is proportional to posting rate. If the posting rate increases then the commenting rate will increase and vice versa. Similar to the previous page, most of the users didn't have access limitations to social media.

The page had an average number of 180 post per month. Despite the change in posting rate, commenting rate grew up and sometimes went down. Because the users were interested in some of these posts similar to previous page.

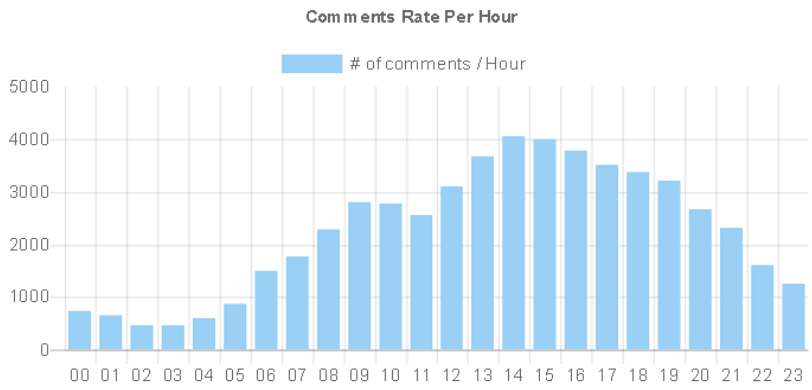

Figure 55: Comment rate per hour.

Figures 56 and figure 57 show the commenting and posting rate respectively.

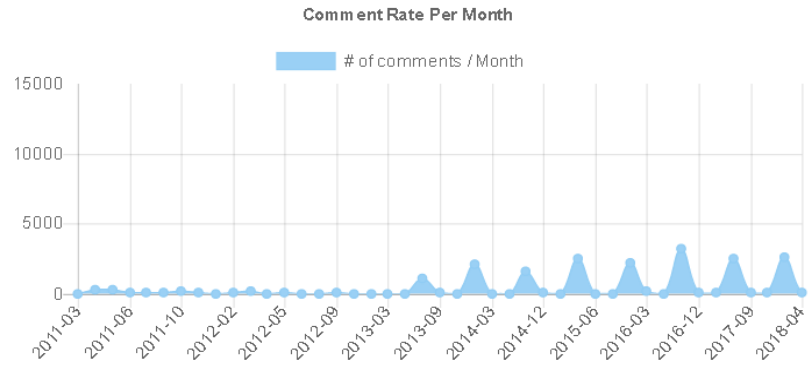

Figure 56: Comments rate per month.

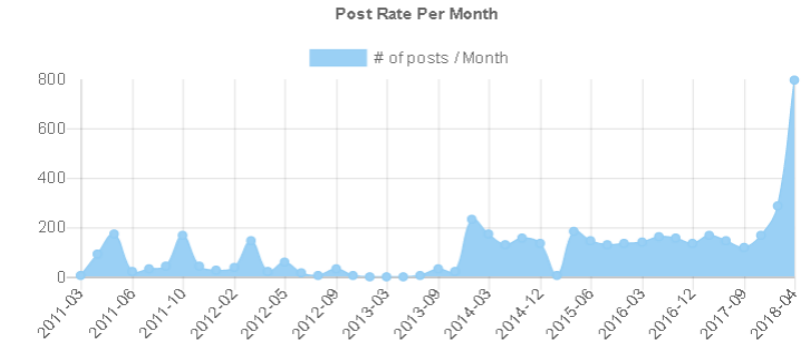

Figure 57: Posts rate per month.

The users were less interested in this page from March 2011 to April 2013. Then users interaction got better but still have ceil points.

Figure 58, shows the word cloud for this page. The word cloud identify the content of this page, where users talked about (“فلسطين”, “غزة”, “الاحتلال”, “المسلمين”, “الشعب”, “العرب”, “مصر”).

Figure 58: Shehab word cloud.

4.2.3 Commercial Pages

Two different Facebook pages have been crawled to cover this section. Starting with Jawwal’s page, it has an average of 19 comments per post as shown in figure 59 and figure 60, so the users are interested in this page and interact with the posts. Most of the posts have around 25 comments per post and the rest of the posts have a number of comments up to 25.

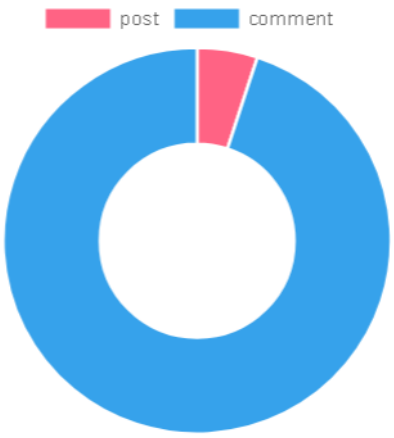

Figure 59: Percentage of comments per posts.

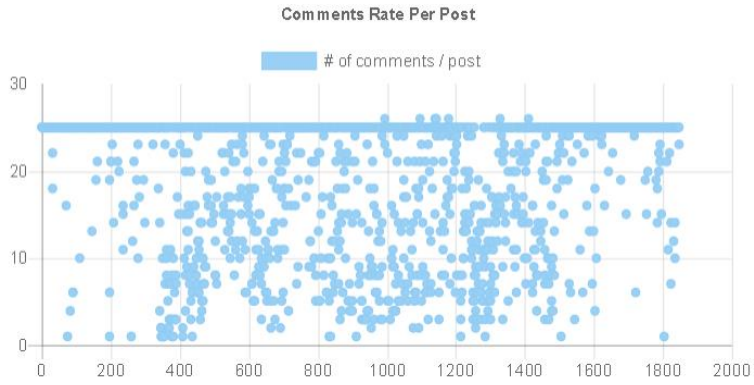

Figure 60: Comments per post.

Having a post with less than 19 comments means that the post is below the average number of comments and this maybe a result of uninteresting post content. So the page owners should review this kind of posts to enhance the quality of future posts to get the users interested in the page.

Referring to figure 61, we noticed that the posting rate had been studied carefully to select a period of time where most of users have free in-home time. These intervals are (5 – 7 AM) when people wake up and stay at home, (11 AM – 1 PM) break time for employees/students and (5 – 7 PM) when people get back to home.

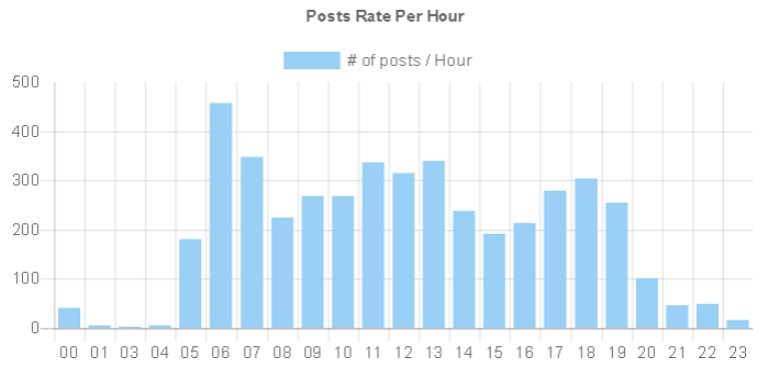

Figure 61: Posts rate per hour.

Choosing the suitable time to publish a post will improve users' interaction and this appears in figure 62, where the user interaction went up and down similar to the posting timing.

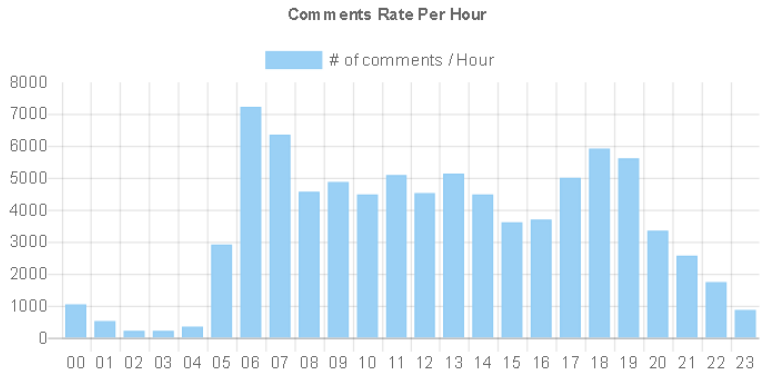

Figure 62: Comments rate per hour.

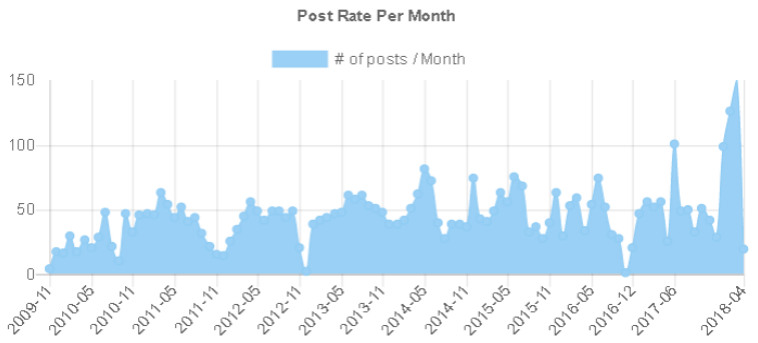

Figure 63: Posts rate per month.

The page have an average number of 50 post per month until the beginning of this year. Going back to the posts. We found a lot of posts talking about 3G release and corresponding offers. Figure 63 and 64 shows the growth of the page in terms of posts and comments respectively.

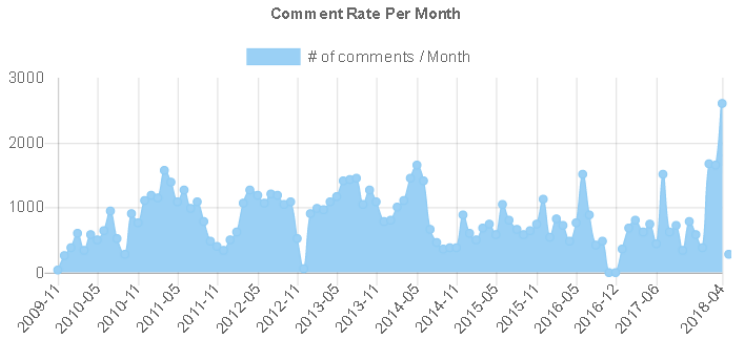

Figure 64: Comments rate per month.

Figure 65 shows Jawwal word cloud where jawwal is the dominant word in the cloud. Here are some of words that are frequently used (“جوال”, “فلسطين”, “شركة”, “غزة”, “مبروك”, “الشعب”).

Figure 65: Jawwal word cloud.

The second page was the Wataniya Mobile page with average of around 15 comments per post. Jawwal page has a better user interaction, figures 66 and 67 shows the percentage of comments to posts and the number of comments per post respectively.

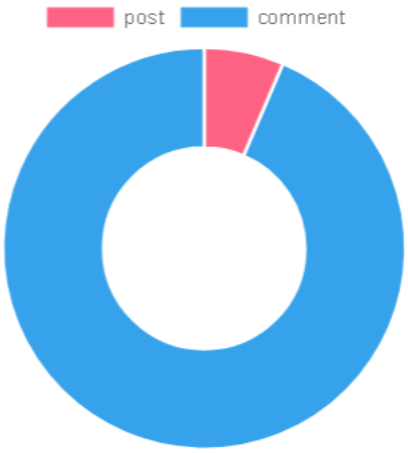

Figure 66: Percentage of comments to posts.

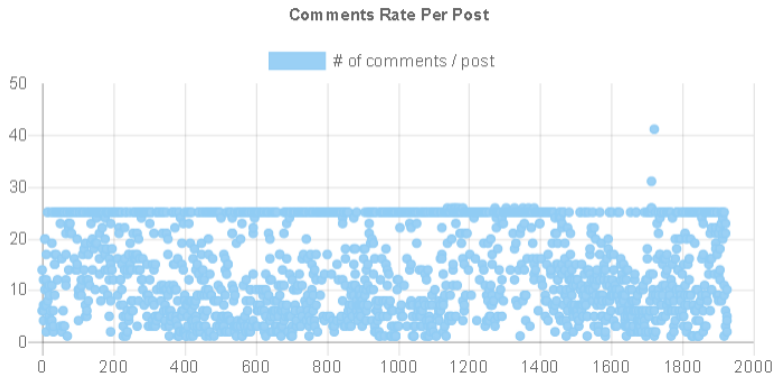

Figure 67: Comments rate per post.

Some of posts have 25 comments and the majority were distributed between zero to 25 as shown in figure 67. Unlike Jawwal page where most of comments concentrated on 25, this page had less user interaction than pervious page.

Posting rate started high at 7 AM and decreased slowly (except the interval from 12 – 1 PM) reaching the minimum level at 1 AM. So the highest posting rate occurred when employee/ students were travelling to their jobs/ universities and this might lead to a low user interaction rate. So choosing the right time to post is the key point to get a high user interaction rate. Figure 68 illustrates the posting rate per hour.

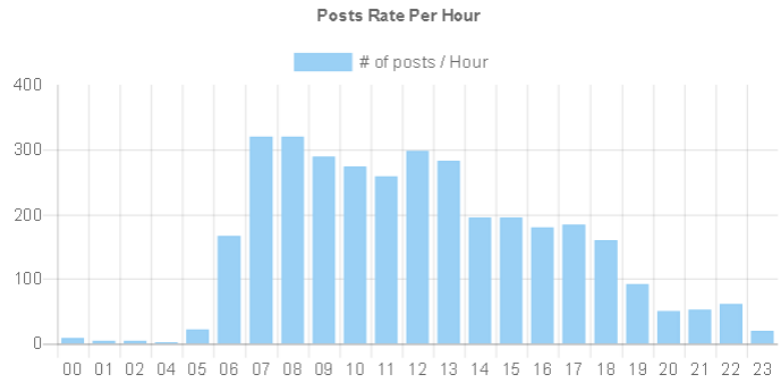

Figure 68: Posts rate per hour.

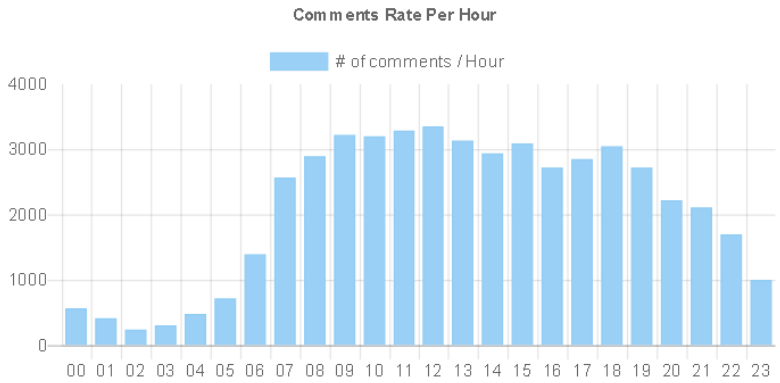

Figure 69: Comments rate per hour.

Although the posting rate has its peak time at 7 AM, commenting rate doesn't have its peak time at the same time, since people are usually busy. Figure 69 shows the relation between comments number and hour. The number of comments started to increase at 4 AM reaching the peak at 12 PM and started to decrease after that.

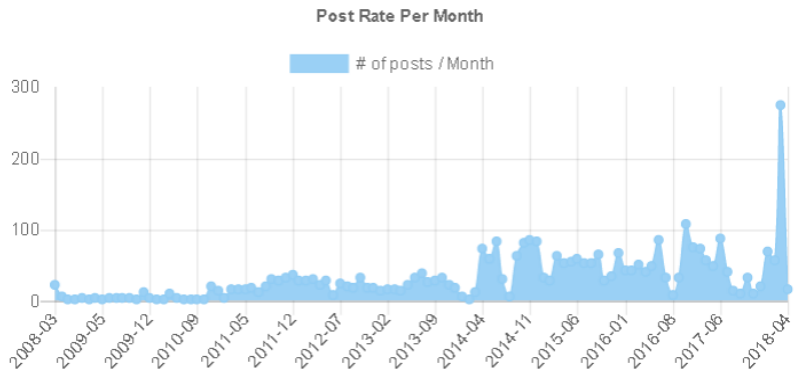

Figure 70: Posts rate per month.

Page grew from a few posts to around 50 posts per month. We have a threshold value at March 2018 due to 3G release like Jawwal page. Figure 70 shows the growth of the page from March 2008 to April 2018.

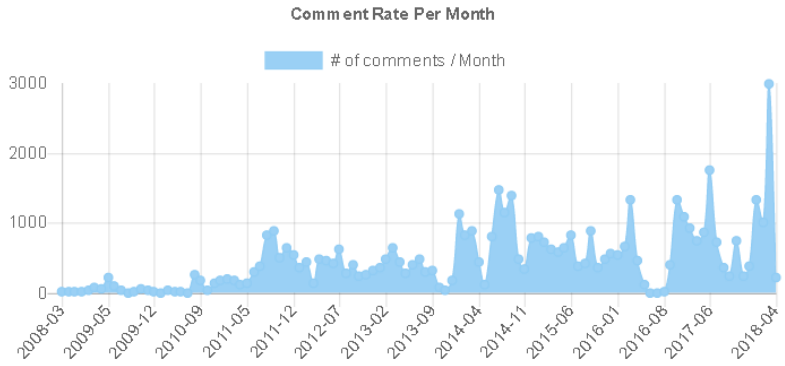

Figure 71: Comments rate per month.

Commenting growth increased similar to posting growth and the users got more interested in this page. Figure 71 shows the relation between the number of comments related to the month.

Figure 72 shows the word cloud for Wataniya page. Here are a few words were that frequently used (“الوطنية”, “موبايل”, “جوال”, “غزة”, “مبروك”,”شركة”). These words reflect the content of this page and the hot subjects.

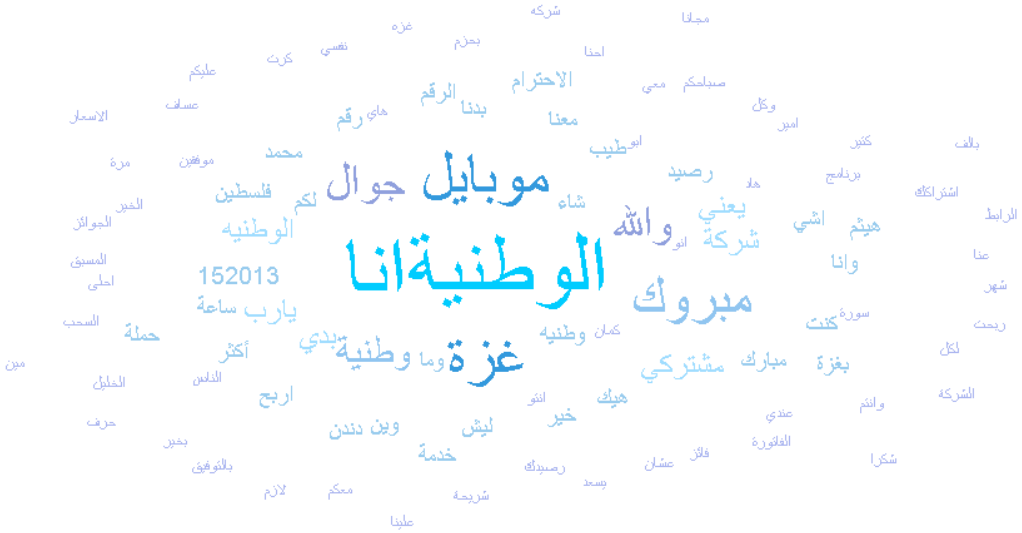

Figure 72: Wataniya word cloud.

4.3 Tool metrics

Our tool had processed 30,072 posts and 343,121comments during the experiment. This was stored in the database to allow us to retrieve the useful information through the analysis modules. Dealing with big data sets requires software performance tuning with strong hardware capabilities. To achieve this we followed the programming standards/ best practices as below:

- Avoid data buffering: so that the data flows from the “httpClient” and is then filtered directly to the database within the crawling process since the buffering at this level will generate a large amount of data and end with memory exceptions (system failure).
- Close database/httpClient connection after completing each process: such kinds of connections have a limited number of instances. If we don’t close each one, the application will fail due to connections reaching the limit.
- Use efficient data structures to analyze the data: we used the hashing in order to generate the word cloud. The code will count the words using the hashing (key: word, value: count). So accessing the words will be direct and will not add performance overhead in contrast with iterative techniques.
- Using MVC architecture: code was split into three modules, models to handle manage the database connection, view as a representation module and controllers to handle the business processes.
- Using robust frameworks: Our application was built using Java Spring, bootstrap 4 and deployed on Apache tomcat server.

The experiment was run in a virtual machine with the specifications below:

- Operating System: Windows 7 Ultimate 64-bit.

- Processor: Intel(R) Core(TM) i7-6600U CPU @ 2.60GHz, ~2.8GHz.
- Memory: 1024MB RAM.
- Network: Intel(R) PRO/1000 MT Network Connection.

With all specifications above, our tool completed the experiment with 26 minutes using 20 M ADSL link.

Chapter 5 Summary and conclusions

5.1 Research overview

In this research, we reviewed the existing Facebook applications/tools that crawl, analyze or visualize Facebook data. Some of them work as Facebook crawlers whereas others as analyzers/visualizers of the data.

We didn't find a universal tool to extract, analyze and visualize Facebook data so the users interested in studying Facebook should use multiple tools to get the data and then apply the analysis tools. Some of them need specialists to analyze the data correctly, for example, applying datamining techniques using R language requires a background in programming in general. There is a need to avoid this kind of complexity to allow any person to conduct his experiment and to get the useful information without having a knowledge background in programming or datamining algorithms.

5.2 Contribution of the study

A new tool has been developed to allow any person to run his own experiment with a few clicks and without any background in datamining or programming. To achieve this we developed a tool that has its own crawler, analyzer and visualizer.

The crawler has been developed on the top of "httpClient" to get the data from Facebook. Currently, it depends on Facebook graph API to fetch the pages-related data since it is the only available API that we can trust to fetch the data. Earlier studies suggested the design of a new crawler that crawls Facebook DOM. This kind of crawlers faced some issues with Facebook due to access rate limitations. The crawlers will therefore be blocked by Facebook after a period of time. Another issue was the dynamic HTML structure, AJAX usage.

Analyzer employs embedded analysis techniques to analyze the collected data with user awareness. After that, the analyzed data will be displayed to the user as a charts/graphs.

5.3 Research limitations

We faced some issues while developing my tool. Facebook Graph API limitations was the main issue. It didn't provide all the required functions, except for the page posts and comments that are available. Additional information such as post likes count/users are not available.

Hardware was another issue, where we were dealing with a huge data. Crawling and storing Facebook data required a large storage capacity. Analysis required high processing speed so it took a long time. To work better, a high performance computers (servers) should be used.

Internet connectivity and bandwidth affect the crawling speed. So having a better internet connection will accelerate the crawler module by reducing the Facebook request time through the httpClient.

5.4 Future research directions

Our research focused on developing a new tool to leverage many tasks from crawling to visualization through the analysis. Since designing such tool requires a long time, we focused on some features and left others for future. So applying data mining techniques such as sentiment analysis is one of the major features that should be implemented in future to add more capabilities to the proposed tool. After that the tool will show the user satisfaction as one of the measurements along with the existing.

Getting user-related information such as post/comment/like owner is an important feature. To do so, a new crawler should be implemented to crawl Facebook HTML DOM directly without a

need to use Facebook Graph API internally. Designing such crawler should consider Facebook access rate to avoid detection and blocking from Facebook.

References

1. Asur, Sitaram, and Bernardo A. Huberman. "Predicting the future with social media." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE (2010): 492-499.

2. Eric Gilbert, Karrie Karahalios “Predicting tie strength with social media” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2009): 211-220.

3. Saravanakumar, M., and T. SuganthaLakshmi. "Social media marketing." Life Science Journal 9.4 (2012): 4444-4451.

4. He, Wu, et al. "A novel social media competitive analytics framework with sentiment benchmarks." Information & Management 52.7 (2015): 801-812.

5. Courtney, K. "The use of social media in healthcare: organizational, clinical, and patient perspectives." Enabling health and healthcare through ICT: available, tailored and closer 183 (2013): 244-249.

6. Tess, Paul A. "The role of social media in higher education classes (real and virtual)–A literature review." Computers in Human Behavior 29.5 (2013): A60-A68.

7. Ellison, Nicole B. "Social network sites: Definition, history, and scholarship." Journal of Computer-Mediated Communication 13.1 (2007): 210-230.

8. Obar, Jonathan A., and Steven S. Wildman. "Social media definition and the governance challenge: An introduction to the special issue." (2015): 745-750.

9. Kaplan, Andreas M., and Michael Haenlein. "Users of the world, unite! The challenges and opportunities of Social Media." Business horizons 53.1 (2010): 59-68.

10. Cohen, K. Bretonnel, and Lawrence Hunter. "Getting started in text mining." PLoS computational biology 4.1 (2008): e20.

11. Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß. "A brief survey of text mining." Ldv Forum. Vol. 20. No. 1. (2005): 2-5

12. Salloum, Said A., et al. "A survey of text mining in social media: facebook and twitter perspectives." Adv. Sci. Technol. Eng. Syst. J 2.1 (2017): 127-133.

13. Pant, Gautam, Padmini Srinivasan, and Filippo Menczer. "Crawling the web." Web Dynamics. Springer Berlin Heidelberg, 2004. 153-177.

14. Arasu, Arvind, and Hector Garcia-Molina. "Extracting structured data from web pages." Proceedings of the 2003 ACM SIGMOD international conference on Management of data. ACM, (2003): 337-348.

15. Clement Levallois, "Where is data visualization going?" Emlyon Business School (2015): 1-10.

16. Friendly, Michael, and Daniel J. Denis. "Milestones in the history of thematic cartography, statistical graphics, and data visualization." URL http://www. datavis. ca/milestones 32 (2009): visited in December 2017.

17. Narang, Sushil Kumar, Sushil Kumar, and Vishal Verma. "Knowledge Discovery From Massive Data Streams." Web Semantics for Textual and Visual Information Retrieval. IGI Global, 2017. 109-143.

18. Xiang, Zheng, and Ulrike Gretzel. "Role of social media in online travel information search." Tourism management 31.2 (2010): 179-188.

19. Golbeck, Jennifer, Cristina Robles, and Karen Turner. "Predicting personality with social media." CHI'11 extended abstracts on human factors in computing systems. ACM, (2011): 253-262.

20. Aiello, Luca Maria, et al. "Friendship prediction and homophily in social media." ACM Transactions on the Web (TWEB) 6.2 (2012): Article no. 9.

21. Rieder, Bernhard. "Studying Facebook via data extraction: the Netvizz application." Proceedings of the 5th annual ACM web science conference. ACM, (2013): 346-355.

22. Catanese, Salvatore A., et al. "Crawling facebook for social network analysis purposes." Proceedings of the international conference on web intelligence, mining and semantics. ACM, (2011): Article No. 52.

23. Catanese, Salvatore, et al. "Analyzing the facebook friendship graph." arXiv preprint arXiv:1011.5168 (2010): 1-6.

24. Flor, Nick V. "Technology Corner Automated Data Extraction Using Facebook." Journal of digital forensics, security and law 7.2 (2012): Article no. 9.

25. Markovikj, Dejan, et al. "Mining facebook data for predictive personality modeling." Proceedings of the 7th international AAAI conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA. (2013): 1-4.

26. Catanese, Salvatore, et al. "Extraction and analysis of facebook friendship relations." Computational social networks. Springer London, 2012. 291-324.

27. Wong, Chi-In, et al. "Design of a crawler for online social networks analysis." WSEAS Transactions on Communications 3 (2014): 264-274.

28. Kijas, Wojciech. "Facebook crawler as software agent for business intelligence system." Studia Informatica 35.4 (2014): 89-110.

29. Aliprandi, Carlo, et al. "Caper: Crawling and analysing facebook for intelligence purposes." Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. IEEE, (2014): 665-669.

30. Mfenyana, S. I., et al. "Development of a Facebook crawler for opinion trend monitoring and analysis purposes: case study of government service delivery in Dwesa." International Journal of Computer Applications 79.17 (2013): Article no. 17.

31. Al-Qurishi, Muhammad, et al. "Selecting the best open source tools for collecting and visualzing social media content." Web Applications and Networking (WSWAN), 2015 2nd World Symposium on. IEEE, (2015): 1-6.

32. Sommerville, Ian. “Software engineering.” New York: Addison-Wesley (2010): 118-146.

33. Silberschatz, Abraham, Henry F. Korth, and Shashank Sudarshan. Database system concepts. Vol. 4. New York: McGraw-Hill, (1997): 259-269.

الملخص

ان انتشار وسائل التواصل الاجتماعي وغزوها لحياتنا اليومية جعل منها حيزا للتعبير عن الرأي، حيث اصبحت هذه الوسائل وعلى راسها الفيس بوك مكانا لابداء الاراء ومشاركة الاحداث والمشاعر. لم يقتصر دور الفيس بوك على الجوانب الشخصية بل تعداها ليصل الى النواحي الاقتصادية والصحية والتعليمية.

بالنظر الى الاهمية التي يتبوؤها الفيس بوك في حياتنا فانه من المهم دارسة المحتوى الذى ينشر عليه والاستفادة منه لتحسين مختلف القطاعات، فان دراسة الصفحات التعليمية سوف يساعد في تحسين نوعية التعليم من خلال دراسة المجتمعات الطلابية وبالتالي تحسين اداء الطلاب، ومن ناحية اخرى سيساهم في ايجاد المعيقات التي تؤثر على الطلاب ومحاولة ايجاد الحلول المناسبة لها. ما ينطبق على الصفحات التعليمية ينطبق على غيرها من الصفحات التى تعنى بالصحة، التجارة، الاخبار ...الخ

خلال هذه الدراسة قمنا بمطالعة المراجع والادبيات لايجات الادوات التي تساعد على استخراج البيانات من الفيس بوك وتحليلها وعرضها، فقد قمنا بمراجعة العديد من هذه الادوات فكان بعضها يُعنى باستخراج البيانات فقط والبعض الاخر يحلل البيانات او يعرضها. هناك بعض الادوات التي تقوم بأكثر من وظيفة (استخراج، تحليل, عرض) ولكن لا توجد اداة مستقلة تقوم بالوظائف الثلاثة المذكورة ومن هنا نبعت فكرة هذا البحث لتطوير الاداة الناسبة لتلبية الحاجات الثلاثة (استخراج، تحليل, عرض).

تم تطوير اداة جديدة لاستخراج البيانات من الفيس بوك وتحليلها وعرضها, وقد تضمنت هذا الاداة العديد من الخوارزميات التي تقوم بفلترة البيانات قبل تخزينها و تحليل البيانات بالنسبة للزمن و مقدار انتاج البيانات وكذل تكرار الكلمات (سحابة الكلمات). أما الجزء الاخير فهو خاص بعرض هذه البيانات على شكل رسوم بيانية لتستخدم من المختصين. ولتجربة الاداة الجديدة، قمنا بتشغيلها على عدد من الصفحات التعليمية والاخبارية والتجارية.

وفي النهاية، خلصت الدراسة الى انه يوجد افتقار بالادوات التي تُعنى بدراسة الفيس بوك وعليه تم تطوير الاداة التى تساهم في دراسته في ضوء المعطيات الحالية.