Use data Mining Techniques to Predict Users' Engagement on the Social Network Posts in The Period Before, During and After Ramadan

استخدام تقنية استنباط البايانات في التنبوء بتفعالية مستخدمي شبكات التواصل الاجتماعي على المنشورات في الفترة الزمنية قبل , أثناء وبعد شهر رمضان

by

HANEEN MOHAMMAD AL RAWASHDEH

A dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS
(KNOWLEDGE AND DATA MANAGEMENT)

at

The British University in Dubai

October 2017

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

_____

(Haneen Mohammad Al Rawashdeh)

COPYRIGHT AND INFORMATION TO USERS

Abstract

Social media has become ubiquitous and main requirement for individuals lives, its used for socialization sharing information and recently for marketing and advertising. The content of the posts has a significant impact on attracting users and total user's engagement. Companies have adopted social media as an essential tool for realizing potential customer needs and to ensure optimal customer satisfaction. Analyzing the content, type and the best time of posting can effectively impact and benefit a business. Data mining (also referred to as Data Knowledge Discovery) is the practice of examining a dataset to establish the hidden patterns and knowledge in order to represent the results in an understandable format.

In this thesis, the effectiveness of using a data mining classifier helps to predict users' engagement with a social media post before publishing and anticipating the best type of media for the post. Different classification algorithms were applied to the dataset using the Rapidminer tool. The results of the classification considered as a baseline for this research. In contrast to traditional approaches on this topic, this dissertation seeks to analyze the depth of users' engagement with social media posts and discuss the basis for a predictive model that to predict the total engagement of a post before publishing.

At the beginning, a data set was collected from Crowdbabble online tool and the collected dataset presented different periods and different social media networks to examine the user behavior at different times. Before Ramadan presented a typical month, During Ramadan related to the special month for Muslims and after Ramadan referred to the beginning of summer vacation. Facebook, Twitter and Instagram were the main platforms examined as these are popular in the Arab world. This study focused on dataset from the Arab world to investigate Arab users' behaviors and interests. The collected dataset was analyzed from various perspectives to study the post characteristics and the effect on users' engagement. Each platform (Facebook, Instagram, Twitter) has a primary role in users' engagement, Facebook is the top used application tool compared to Instagram and Twitter. The results showed that Post type (Status, Photo, Link, Video) has an important role in attracting users and the analysis confirms that video and photo posts create the maximum levels of engagement. Post fields also had impact on engaging users, whereby the categories of Beauty, Fashion and Celebrity were the most attractive page types.

خلاصة

أصبحت وسائل التواصل الاجتماعي في كل مكان ومن المتطلبات الأساسية لحياة الافراد، حيث أنها تستخدم لنشر المعلومات وتحسين العلاقات الاجتماعية ومؤخرا أصبحت تستخدم كثيرا في مجالات التسويق والإعلان. محتوى المشاركات له تأثير كبير على جذب المستخدمين وإجمالي مشاركة المستخدم. وقد اعتمدت الشركات وسائل الإعلام الاجتماعية كأداة أساسية لتحقيق احتياجات العملاء المحتملة وضمان رضا العملاء الأمثل. يمكن أن يؤدي تحليل المحتوى والنوع وأفضل وقت للنشر إلى التأثير بفعالية على النشاط التجاري والاستفادة القصوى منه. تعد عملية استخراج البيانات (التي يشار إليها أيضا باسم استنباط المعرفة واكتشافها) هي عملية التنقيب في مجموعة بيانات لتحديد الأنماط والمعرفة الخفية لتمثيل النتائج في شكل مفهومي.

في هذه الأطروحة تبين أن استخدام مصنف البيانات يساعد على التنبؤ بنسبة مشاركة المستخدمين وتفاعلهم مع الاعلان المعروض في وسائل الاعلام الاجتماعية قبل نشر وتوقع أفضل نوع من وسائل الإعلام لهذا المنصب. تم تطبيق خوارزميات تصنيف مختلفة على مجموعة البيانات باستخدام أداة برمجية Rapidminer. وتعتبر نتائج التصنيف بمثابة خط أساس لهذا البحث. على النقيض من النهج التقليدية في هذا الموضوع، تسعى هذه الأطروحة لتحليل عمق مشاركة المستخدمين مع وظائف وسائل الاعلام الاجتماعية ومناقشة الأساس لنموذج تنبؤي فعال للتنبؤ بمستوى الانخراط الكلي مع الاعلانات قبل النشر.

في البداية، تم جمع بيانات من عدة وسائل تواصل اجتماعية , وقدمت مجموعة البيانات التي تم جمعها على فترات مختلفة (قبل شهر رمضان , خلال شهر رمضان , وبعد رمضان)

كانت فيسبوك وتويتر وإينستاجرام هي المنابر الرئيسية التي تم فحصها لأنها الأكثر شعبية واستخداما في العالم العربي. ركزت هذه الدراسة على مجموعة البيانات من العالم العربي للتحقيق في سلوكيات ومصالح المستخدمين العرب. تم تحليل مجموعة البيانات التي تم جمعها من وجهات نظر مختلفة لدراسة خصائص الاعلان المبوب وتأثيره على مشاركة المستخدمين. كل منصة (الفيسبوك و إنستاغرام و تويتر) لها دور أساسي في مشاركة المستخدمين، الفيسبوك هو أكثر أداة تستخدم في العالم العربي مقارنة إينستاجرام وتويتر. وأظهرت النتائج أن نوع المشاركة (الحالة، الصورة، الرابط، الفيديو) له دور هام في جذب المستخدمين ويؤكد التحليل أن مشاركات الفيديو والصور تخلق أقصى مستويات المشاركة. كان لنوع ومجال الصفحات أيضا تأثير على إشراك المستخدمين، حيث فئات الجمال والأزياء والمشاهير كانت أنواع الصفحات الأكثر جاذبية.

Acknowledgement

First, and most of all, I would like to thank Allah for giving me the chance to carry on my study, and I would like to express my appreciation to my supervisor Dr. Sherief Abdullah for his support, encourage and guidance to complete this dissertation.

Also, I would like to express my deepest gratitude to my Mum, her prayers and support were the motives all the time.

The one who had walked alongside me, helped me and I dedicate my success, my beloved husband. This journey would not have to be possible without your tolerance and cooperation.

Finally, my sincere thanks also go to my family members and to all my friends specially the one that I have had the pleasure to work with, during this journey Faten.

Table of Content:

Contents

List of Figures:

List of Tables:

Chapter 1 Introduction:

This chapter introduces the birth and growth of social media and internet users. It shows the importance of social media in advertising and business and. also includes the research motivations and objectives along to the research questions to be answered. Moreover, it describes how this research is structured to target the reader and why it is unique.

1.1 Introduction

Social media has profoundly changed our lives and the interaction between people around the world. Social media platforms like Facebook, Twitter and Instagram have become essential resource for real-time information with a huge amount of users (Yue et al.2014). Social media applications are used for various reasons: socializing, communicating and receiving information. A result, companies have adopted social media as an appliance to raise business values (He, Zha & Li 2013). Social media has gleaned valuable data and information about its users, such that this stream of live information is considered to be a great utility for large corporations and government organizations. Consumers react with positive or negative comments on posts, and these comments may spread on a platform; such knowledge has precious marketing value for companies (Yue et al.2014). Companies use social network to influence customers by advertising and contacting customers to ensure satisfaction. As result, researchers are to unearth relationships between online publication and user engagement (in the form of Likes, Comments, and Shares) of users about a publication (Moro, Rita and Vala ,2016). Predictive systems using data mining algorithms are able to predict users' engagement based on a post's evaluation. This thesis investigates a post's characteristics and the impact on users' engagement in Facebook, Twitter and Instagram applications. In addition, it introduces a predictive model using data mining algorithms to forecast users' engagement with a published post. Data mining provides a helpful technique to extract and predict knowledge from a dataset. The study focuses on three different periods: (1) Before Ramadan (2) During Ramadan (3) After Ramadan. Although the time of Ramadan changes each Gregorian calendar year, Before Ramadan typically represents a month in the period (25^{th} *--25^{th} May in 2017), During Ramadan is the month that stands for special occasion for Muslims and was (26^{th}May-24^{th} June in2017). After Ramadan refers to a general summer vacation (25^{th} June-25^{th} July in2017). Several types of analysis performed in this research demonstrated the effect of post

attributes on user engagement. The analysis includes: Network analysis exploring the network type (Facebook, Twitter, Instagram) on the engagement, Field Analysis that examined the effect of pages' fields or content (Beauty, Cooking, Health, Media, Fashion, Shopping, Social and News) on the users' reactions, Media Analysis to discuss post type (Photo, Video, Status, Link) with users' interaction, Time Analysis to investigate the influence of time (Hours, Workday, Weekend) before, during and after Ramadan on users' responses. Also, Emotions analysis states the positive and negative reactions on posts through different page fields and over different times. The focus of this research specifically is on honing the predictive model to predict the engagement weather (very low, Low, Moderate, High, Very High). Many Data mining algorithms were applied on the dataset to classify the posts (Decision tree, Naïve Bayes, K Nearest Neighbor) and the classifiers evaluated the content using the 10-validation accuracy method.

This research is extends from (**How to improve users' engagement with social media in UAE**) research done in the Data mining course. The goal of this thesis is to utilize the extracted knowledge in supporting the decision maker to validate the post before publishing the post.

1.2 Overview about Social Media.

Social Media is an online form of publishing where users can interact and communicate by creating accounts and share personal information between others. Social media sites exhibit an exponential penetration to users' daily lives and change deeply the communication around the world. (Qualman, Safko & Brake, 2009). Recent research indicates that social media applications like Facebook, Twitter, Instagram and YouTube are used for many reasons; to contact friends, for entertainment as well as the sharing of information. These reasons are why companies adopt social media as a tool to increase the business value and customer loyalty (He, Zha & Li 2013). Social media applications generate a wealth of data, this amount of data has hidden and useful knowledge for different fields such as: business, education and health (Dey, Haque, Khurdiya, & Shroff, 2011).

The Facebook platform launched on February 2004 and now has more than one billion active users such that more than 50% of users log in to Facebook site or mobile application. Nowadays, business and organizations use Facebook as official channel to communicate, interact and engage with customers ((Hays, Page & Buhalis 2013). Also, Facebook and other social media is used in

education to access content and maintain student engagement with the instructor, so learners are able to define, manipulate and evaluate existing knowledge. Fostering learning is another advantage of social media; by collaborating and discussing content, learners can create new meaning and understanding the material. Furthermore, it enables authentic learning through assisting learners to create video/audio /photo, microblogs that include short sentences and images (Gikas & Grant 2013). Instagram is a medium that is used to create and send photos quickly derived from the two words “Insta” and "Gram”. Instant was a name for the Polaroid camera that shot photos and printed them immediately. Gram term is taken from “Telegram” a tool used to send information quickly. The purpose of Instagram is to shoot and send pictures quickly (Sudrajat et al. 2016).

Twitter launched in 2006 as micro-blogging website where users can post 140 characters as a tweet. Users of Twitter are called followers, because they have to follow an account before browsing it. They can follow persons like celebrities and comedians and share their interests. Organizations consider Twitter as a successful tool to engage with customers and understand their behaviors and help to understand how this behaviour will influence the business (Hays, Page & Buhalis 2013).

1.3 Social Media in Business.

In the last few years, the advent of social media led to a radical change in the ways business and marketing operates. Currently enterprises use social media in Business-to-Business (B2B) transactions. In market and product development there are significant differences between B2B and consumer product sectors. Developing new products is complex and takes time in B2B sectors. In addition, it requires dealing with large organizations as customers where the cooperation with them is more direct but complex. In this context advertising is critical for success. Corporate purchasing managers often have detailed criteria upon which to make decisions, so detailed information about products is essential (Jussila, Kärkkäinen & Aramo-Immonen 2014). Social media allows producers to communicate differences, in turn purchasers take these into consideration. Sales people that engage and communicate with customers in a responsive manner, and leverage on social media can produce customer satisfaction by providing the means to enable a positive experience to meet customer satisfaction (Agnihotri et al. 2016). Social media also has had a significant impact on customer relationship management (CRM) through developing modern

capabilities that foster strong relationships with customers. Organizations invest in Social customer relationship management (SCRM) and over the past few years and spent more than one billion U.S. dollars according to (Sarner et al. 2011). SCRM has many capabilities further enabled by social media like customer-linking and e-marketing capabilities to drive customer satisfaction, loyalty and retention.

1.4 Social Media for Interactive Organizations

The advent of social media generated the opportunity to engage stakeholders with customer in different style better than the available on traditional web sites. Social media offers three elements for the organization -customer's relation: messaging feature with dynamic updating, various media sharing and interactive applications (Saxton & Waters 2014). Twitter and Facebook offer the feature of interactive discrete messages, where customers are able to share comments and link them to create interactive system for the organization. These interactive features are the chief dynamic component the social media sites (Lovejoy & Saxton 2012). Organizations can measure the interaction through observing the number of direct messages and viewing the replies and mentions on the message. Some social networks offer the statistics for interactive messages indicators such as the number of users' clicking and the number of sharing and retweeting for messages (Saxton & Waters 2014).

1.5 Research motivation.

Using the social network as a tool for marketing and advertising in raise. This research gives the opportunity to evaluate the post characteristics (Type, Time, Network application) before publishing it. As long as these characteristics have significant consequences in advertising. This report provides valuable insights about consumer behavior in the dominated social media for marketers and brand managers. In addition, evokes the decision maker to fulfill the desired gratifications. Furthermore, this thesis is unique because it's the only research that analyze dataset for Ramadan.

1.6 Research questions

Through the empirical experiment this research will answer the following questions:

- Research Question 1: Which social network application is the best for advertising?
- Research Question 2: What is the best media type for the post to attract users?
- Research Question 3: When is the best time to publish the post for maximum engagement?
- Research Question 4: How to predict the engagement before publishing the post?

1.7 Research methodology

Since social media posts have a significant effect on marketing and advertising products, this study will contribute in optimizing the best model to enhance posts engagements. The study approach was built on a mixed methodology of both qualitative and quantitative techniques. The qualitative phase is done via an online survey that published on different social networks. It was used to attain an insight into active and broadband Arabic public pages and to disclose the effective page fields had on the social network. Furthermore, the survey received feedback from Arab users in the region to view the impact of post type on Arab pages and assess the effect on business. A quantitative study was used to collect dataset from the filtered public pages using Crowdbabble analytical tool. Crowdbabble was used to export the posts from the desired public pages in the period (before, during and after) Ramadan during the period (25^{th}April-25^{th} July 2017. Furthermore, analyzing the pages and displaying statistics on each page such as: the total page fans, average of page posts per day, engagement per post, engagement rate and the total engagement. The preprocessed dataset was analyzed using MS Excel as well. Different analysis was undertaken using MS Excel starting with ANOVA one way test the impact of social network on total engagement, Network analysis, Field analysis, Time analysis, Media analysis and Facebook reactions analysis. Rapidminer tool for data mining was used to create a prediction model to ascertain the impact of post characteristics (type, time, field) on user's behavior and engagement. In addition, the model was used to expect the engagement on post before publishing

on the page and the reaction on the post before publishing to assess the decision maker (manager) in choosing the best type, time and network. Figure 1 shows the methodology of this research.

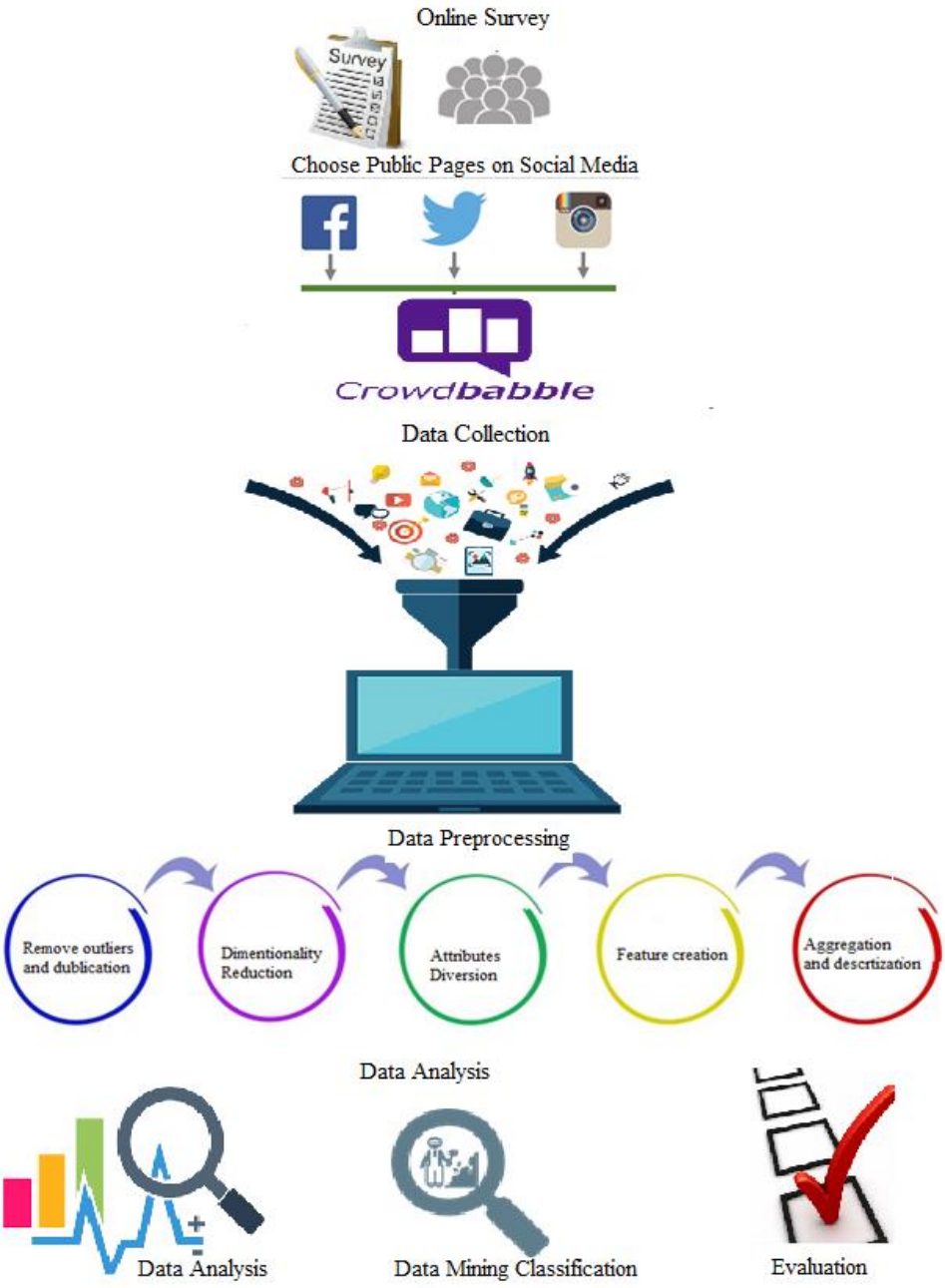


Figure 1 Research Methodology.

1.8 Uniqueness of the study

Most of the completed researches in this field focus on the relationships between published posts on the social network and the influence of these online posts on users' behavior and interaction. (Cvijikj et al., 2011). However, few studies have implemented a predictive model to expect the evaluation and engagement on the post before publishing (Moro, Rita & Vala 2016). Most of predictive models that applied on datasets from the Arab world are used for opinion mining (known as sentiment analysis). The developed system in this research is able to predict the evaluation of the post through predicting the category of total engagement (Very low, Low, Moderate, High, very high) using different classification models. Furthermore, this research analyzes the emojis reactions on Facebook post over the publication of different pages in various fields (Beauty, Cooking, Health, Media, Fashion, Shopping, Social and News). In addition, it explains the effect of day time (hours), duration (before, during and after) Ramadan and (weekday, weekend) on users' reaction reactions. From the other side, this study analyzing data in Ramadan as it is a holy month and the behavior of users will differ in this month.

1.9 Dissertation structure (Design)

The rest of the research is organized as follows: Chapter 2 includes a literature review about data mining and data mining algorithms, it also provides background about social media in Arab world and data mining in social media for Arab world. Chapter 3 describes the used dataset collection and processing. Chapter 4 contains the dataset analysis from different perspectives and analyze the effect of each perspective on the average engagement.

Chapter 5 includes the classification process for the dataset and the performance evaluation, also the use of different classification algorithms and the evaluation for each technique.

The experiments results discussion and researches question answers are in Chapter 6 and finally Chapter 7 have the conclusion and the future work.

Chapter 2 Literature review

This chapter introduces background about data mining data mining algorithms. In addition, this literature describes how data mining techniques are used to predict knowledge from social media, data mining in social media for Arab pages. The discussion for previous studies in using data mining in social media for different purposes is mentioned as well.

2.1 Overview of Data Mining.

Data mining is an emerging technology that attempts to extract meaningful, novel and useful patterns from large numerical or textual datasets. Data mining has two models (a predictive model and a descriptive model). These models can help to support decision makers in various domains. Data mining has different algorithms and tools that have the ability to work with large amount of data and discover the hidden patterns. Data mining involves four phases to accomplish an outcome: understand data, preprocess data, apply the algorithm to establish the model and finally evaluate this model (Moro, Rita & Vala 2016). To glean useful information, researchers used data mining technology and apply it in different fields and different case studies. Educational data mining is important because it helps the learning process by predicting students' performance and understands how students learn (Romero, Sebastian 2013). A study by Abeer and Ibrahim (2013) was undertaken to improve students' performance and reducing failure rates. The model had forecasted student performance by comparing mid-term marks, homework accomplishment, and laboratory test grades. Text mining has a role in data mining because it is used to predict model from textual attributes. In business, sentiment interpretation can predict the financial market. Maks and Vossen (2012) present sentiment analysis as a model for customer reactions as either positive or negative. By using a text classifier, they were able to detect customer emotions. This method helps in business operations by anticipating the engagement of a new product for market based on customer feedback or a review. Data mining is used widely to unearth significant hidden information in many fields in order to produce meaningful and useful knowledge.

2.2 Data mining on social media.

Nowadays social media is considered a valuable repository of data. The ease and versatility of using social media in various fields polarizes the researchers to use social media data in their studies. Due to the capability of data mining to effectively handle the three dimensions of a social network's data (namely size, noise and dynamism), data mining methods are used to mine important patterns from data. (Mariam et al. 2014). There are two main identified analyses for Data mining in social media (Linkage-based and Structural Analysis) and (Dynamic Analysis and Static Analysis).

In linkage based analysis the focus is for link social media behavior with relevant node, links, and communities with the network. Dynamic Analysis has interaction between entities data in dynamic social networks like Facebook and YouTube is difficult to carry out, where data generated in high speed and capacity. However, in static analysis for social network the analysis done in batch mode due to the gradually changes in networks over time. (Olowe et al., 2017).

A relevant study to this thesis was undertaken by Moro Rita and Vala (2016) to predict social media impact on brand building. The study was based upon Facebook posts only, whereby 790 posts were gathered from a brand page. They noted that the posts had different characteristics such as the type of content (link, photo, video) and the lifetimes of the post (such as the category, page total likes, type, month, hour, weekday, paid). Data was processed, and the outliers removed. 751 posts were analyzed.

Support vector machine algorithm was applied to predict the performance metric on the social brand. They resulted the followings: the post type was considered as the most relevant attribute that increased the performance; the month of posting (seasonality) produced great impact. This study can be improved by analyzing more datasets and include text mining in the performed model.

Another study done by Sudrajat et al. (2016) sought to obtain and classify personality types and find a companion based on personality types. The success depended on the classification of documents that were used. Classifying personalities was determined by the largest VMap from each category. The Naïve Bayes Method was used in this experiment for text mining, as well as a training set from dataset about personality assessment. The testing set was collected from Instagram photos that had captions. The classification results for personalities on Instagram

showed that 21.5% of users were rated high for conscientiousness, while 59% of Instagram users were evaluated high on Openness to New Experiences as a personality trait.

Measuring the effectiveness of messages and reviews is an important part in evaluation marketing communications; many researchers turn to measure the emotional response from social media. Williams and Mahmoud (2017) presented a study that aimed to detect, classify, interpret emotions from user's tweets. Their experiment had 1000 sampled from a broad system. Automated sentiment analysis was applied on the dataset with supervised approaches (Naive Bayes and Support Vector Machines) on labeled tweets to make sentiment predictions. The results showed that Naïve Bayes and SVM text classifiers were more accurate in detecting the emotions expressed than the general sentiment techniques.

Furthermore, understudying students' behavior using data mining in social networks was helpful in understating students' problems and issues in their learning. An analytical study conducted by Chen, Vorvoreanu, and Madhavan (2014) they selected random sample of 25,000 tweets and found that engineering students are suffering from heavy load studies, lack of social engagements and sleep deprivation. These results were implemented into multi-label classification to classify tweets that reflected students' problems, algorithm had used them to train 35,000 tweets to produce model. This model shows how informal social media can provide insights into students' experiences.

2.3 Social media in Arab world.

The proliferation of internet use and the rapid spread of mobile phones with social applications have led the users in Arab world to respond quickly and embrace the modern technology. In turn, this has fostered users to utilize all the offered features from the social network to connect, communicate and share information. The most consistent social network user group in the Arab world is the youth. Business and media quickly ride the new wave and connect the targeted customers to offer services and advertise using the virtual presence on the social network. Government soon started to leverage the new online territories to support communication between the population and its institutions.

2.3.1 Arab social media users

Arab social media users can be classified into five categories based on their usage and needs to: Explorers, Achievers, Escapists, Pragmatists and Social Butterflies. Explorers are continuously seeking novelty; they use social media as a tool for knowledge discovery. Achievers are ambitious and highly active users looking to improve their knowledge. Achievers use social media to receive and share news on the spot, so they can interact immediately to the content and with other users. Escapists are communication enablers, they are social and contact others without feeling shy or inhibited. Pragmatists are often using the social network when necessary and to gain useful knowledge. Social Butterflies are highly engaged individuals who use social media to interact with large groups and open to making new friendships. There are many reasons for people to use a social network: watching videos, sharing photos, gaining information and listening to music. Chatting is popular activity and the smart mobile phone is main mode of access for the majority of users that generally access platforms in the evening.

2.3.2 Impact on Arab Society

Social media has three aspects to the effects on Arab society and business: Communication, Knowledge and Entertainment. For communication, social media reduces the distance and makes the world a small village by reducing the cross-cultural and geographical barriers and helping users to become closer to each other. In terms of knowledge, social media opens channels for searching, learning and inspiration. It also allows users to interact and receive up to date information. Social media opens the opportunities to develop professional careers and business opportunities.

2.3.3 Studies for Arabic social media

The table1 below summarizes and reviews some studies about data mining in Arab social media it was noted the most of reviewed studies focus on text mining and sentiment analysis.

Authors	Field	Study purposes	Sample	DM Algorithm	Findings
Mohamed M. Mostafaa, Ahmed A. El-Masry(2013)	Profiling e-government services' users.	to observe the effect of various demographic, cognitive and psychographic factors on Egyptian citizens' use of e-government services.	Data were collected drop-off, pick-up methods. Total of 1500 questionnaires were distributed. All questionnaire items, originally published in English, were translated into Arabic using the back-translation technique.	Classification and regression trees CART SVM	trust in e-government systems had the most important role in rule induction. Neural network models can learn input–output relationships to make perfect forecasting for data. The Ease of use for e-government systems increased the use of the government Web site.
Wa'el Hadi 2015	Sentiment Analysis	To classify Arabic Twitter corpus and define the tweets to (Positive, Negative, Neutral).	Arabic Twitter corpus that collected from Twitter ARCHIVISIT	Decision tree NB KNN SVM	SVM model that used to classify the corpus outperformed the KNN NB and Decision tree.
Agarwal , Sureka , and Goya 2015	Intelligence and security application for online radicalization and civil unrest	mining free-form textual content present in social media to describe the online radicalization and civil unrest	Collected dataset from Twitter and Youtube API	Clustering, Logistic Regression. KNN, Naive Bayes, SVM, Rule Based Classifier, Decision Tree.	The Data mining tools were able to predict upcoming events related to civil unrest. Classification algorithms were able to identify extremism and hate promoting content
Salloum et al 2017	Text mining for social in social media	looking for different text mining methods to identify key themes in data.	Survey study No dataset	Different classifiers for sentiment analysis	Arabic text is overlooked in researchers' studies, they open the door to increase the number of studies in social media

Schroeder et al. 2017	Mining Twitter Data	Explore how social media is able to frame grievances	Al Jazeera Arabic and Al Arabiya tweets and comments	Clustering	Social media Twitter specifically played a role in the emergence of the Egyptian Arab Spring revolution.

Table 1 Related work for Arabic social media

Chapter 3 Building the dataset

This chapter describes the collected dataset with the original features (or attributes), and shows the features of each element of data and the possible values of each attribute.

3.1 Data Collection methods and tools

The original dataset was extracted from an online tool generator (Crowdbabble). Crowdbabble trial account is active for 7 days only and the account allows analyzing and extracting 30 public pages. An online survey using Survey Monkey was conducted to select the active and familiar pages on the three social networks. This survey is published through Facebook and provides access to users in different Arab regions (Levant region, GCC, Egypt). From the survey, I filtered 90 public pages for analysis, taking into consideration to ensure consistent 30 pages on each social network, that the page fans must exceed 2M, as well as the number of posts in one month, average/post, engagement/post, and the engagement rate. Pages were from different fields (Beauty, Cooking, Health, Media, Fashion, Shopping, Social and News).data was collected from the period 25th April 2017 until 25th July 2017 to represent the intended period (before, during and after) Ramadan. Al Eid posts were excluded from the dataset because online behavior unusual as it is considered a special occasion.

3.2 Data Description

Crowdbabble is an analytical tool for social networks such as: Facebook, Twitter, Instagram and LinkedIn. It extracts posts from public pages with specifying post as link, type of post (photo, video, status link), time of posting in GMT timing, date, number of page followers, the user's reaction on this post including total of likes, comments, shares. On the other hand, it shows the Facebook reaction (happy, wow, love, sad, angry) with the total of each reaction. The collected dataset contains 76,735 posts from broadcast public pages in the Arab region. Crowdbabble returned the data for each page separately in an Excel spreadsheet file. The raw dataset contains the following attributes:

1. Post link.
2. Post content keywords
3. Post Media Type.
4. Date of post.
5. Reactions (for Facebook only)/Likes.
6. Comments/Reply.
7. Shares/Retweets.
8. Total of Engagement.
9. Used filter (for Instagram only).

In table-2 below the collected datasets features and the possible values before prepossessing.

Attribute	**Data Type**	**Description**	**Possible values**
Post link	Polynomial	The URL for the post in the page.	URL link "http://"
Post content	Polynomial	Key works mentioned in the post, special character # or @	Text
Date(GMT+0)	Date_time	The date and time for publishing the post (GMT+0).	Text that contains the date and time.
Post Media type	Polynomial	The media type of the post	Status, Image, Video, Link
Filter	Polynomial	Applied filers on the photo before posting.	Normal, Inkwell, Clarendon, Valencia
Reactions/Likes	Integer	Number of users liked the post	Integer number (0, ∞)
Comments	Integer	Number replies on the post by the users (followers)	Integer number (0, ∞)
Shares\Retweets	Integer	Number of users who shared or retweet the posts - 'share' in Facebook \ - 'retweet' in Twitter	Integer number (0, ∞)
Total Engagement	Integer	Summation for the reactions, comments, shares numbers.	Integer number (0, ∞)

Table-2: Raw dataset description and features.

3.3 Data Preparation and Preprocessing

Before starting the experiment, some preprocessing tasks should be done on the collected dataset, since it is crucial to prepare the raw data for proper data mining algorithms and analysis. Data is prepared using the following steps:

- **Sample Gathering**:

  For each individual public page, the data file returned separately, so 90 distinct files must be combined. Using Excel, I combined the dataset into one file.

- **Feature Creation:**

  The collected dataset did not consider the social network attribute (Facebook, Twitter, Instagram) because Crowdbabble returns an individual file for each page. Likewise, the field type or category feature for the page is not considered. So two attributes were added to the dataset:

  - **Network application attribute**: contains the belonged social network for each post. (Instagram, Facebook, Twitter).
  - **Field type attribute**: this feature shows the page type for the post based on the page (Beauty, Cooking, Fans, Fashion, Health, Media, News, Shopping, Social).

- **Dimensionality reduction:**

  Some attributes considered as irrelevant and omitted from the dataset, because they had information about the post. such as: post id, post link and post content.

- **Attributes Diversion**

  The Date attribute had the time and date for post creation, it was in (GMT+0) format. Time is crucial in the analysis.

  - Date attribute was transformed to (GMT+2) for time as long as most of pages from (Levant region, GCC, Egypt).
  - The new date attribute converted to day attribute in order to show weekday/workday attributes.

- **Aggregation:**

  Combine all attributes of like, share, reaction, comment, tweet into one attribute (total_ engagement).

- **Discretization:**

- The total_engagement on the post attribute was categorized into five categories: very low, low, moderate, high, very high.

- Day of the post also categorized into (weekday), (weekend).

The dataset after applying preprocessing techniques is shown in table-3.

Attribute	**Description**	**Possible values**
Network_Application	The type of social media network that the post belongs for.	(Facebook, Twitter or Instagram)
WeekDay	The day which the post was posted on	(Saturday, Sunday, Monday, Tuesday, Wednesday, Thursday or Friday)
Day_Type	The Categorical attribute of the week day	contains date and time.
Time (Hour)	The time when the message was posted	Time representing hours of the day (from 00 to 23)
Media_type	The media type of the post	Status, Image, Video, Link
Field_type	The media type of the post	(Beauty, Cooking, Fans, Fashion, Health, Media, News, Shopping, Social)
Total_engagement	The total number of engagements on a post (Likes, comments, shares, or retweets)	number ≥ 0
Total_engagement_Group	The nominal value of the attribute *Total_engagement*	(Very low) $\rightarrow x \leq 500$ (Low) $\rightarrow 501 \leq x < 1500$ (Moderate) $\rightarrow 1501 \leq x < 2500$ (High) $\rightarrow 2501 \leq x < 5000$ (Very high) $\rightarrow x \geq 5000$

Table-3: Dataset after preprocessing.

Chapter 4: Posts Analysis

This chapter covers several types of conducted analysis and shows the results for better interpretation of the data. This section explains the analyzed dataset starting with a ANOVA test, based on network type, page field, type of media and time of post. On the other hand, it shows the analysis for Facebook reactions (like, happy, angry and sad).

4.1 Network Analysis:

The Network analysis is focusing on the average of user's engagement before, during and after Ramadan using the posts from public pages for each social network (Facebook, Twitter and Instagram). Furthermore, it describes the number of followers for the social media across the three networks. As Figure-2 shows the number of followers and the percentage of users' growth in 2017. The chart demonstrates that Facebook has the maximum number of users followed by Twitter and Instagram in the region. The percentage of users' growth for Facebook has a maximum rate 39%, followed by Twitter with 1.9% then Instagram whereby the rate is 1.8%. The chart confirms that Facebook is the preferred social channel in the Arab world while Instagram has the lowest preference compared to Facebook and Twitter.

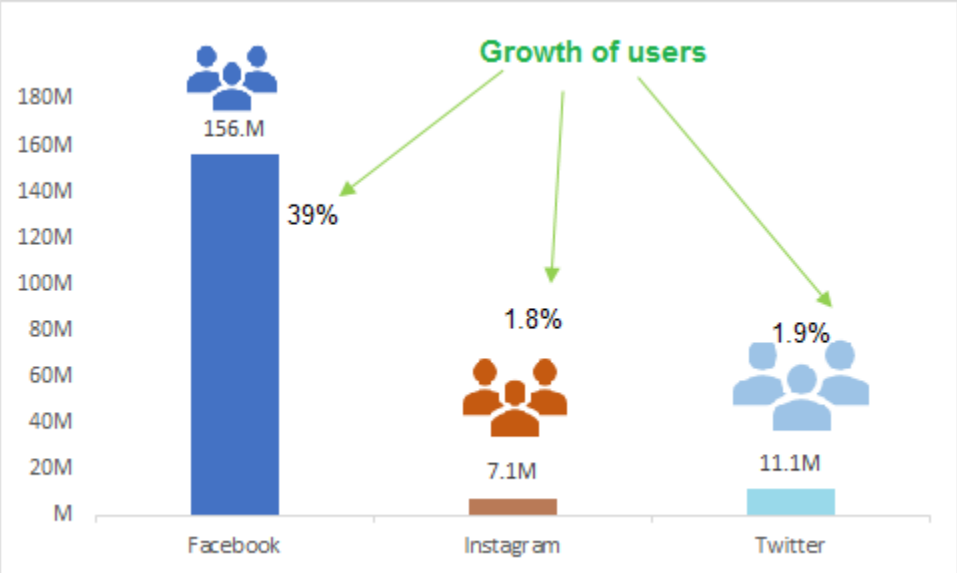


Figure 2: The number of followers for the social network channels in 2017.

The other phase of Network analysis focuses on the average of engagement on each post comparing to the count of posts on each network (Facebook, Instagram and Twitter) within the mentioned period (before, during and after Ramadan). In Figure 3 it's clear that page owners on Facebook were more active and posted around 28K posts before Ramadan, 12.5K during Ramadan with the posts after Ramadan jumping to 21K posts again (in the summer). On Twitter, users tweeted around 3K posts; almost the same during each period. Although Instagram's posts don't exceed 3.3K the posts were widely spread comparing to Facebook and Twitter. Figure 3 shows that the average of users' engagement exceeds the 34K before Ramadan on Instagram. This indicates that Instagram is a better choice to post then get more broadcast range in the region. When comparing Twitter with Facebook, we found that the engagement rate on Facebook posts were higher than Twitter before and after Ramadan. Despite of the variation in post numbers on Facebook during Ramadan compared to tweets, the reaction during Ramadan on Twitter was larger as Figure 4 displays.

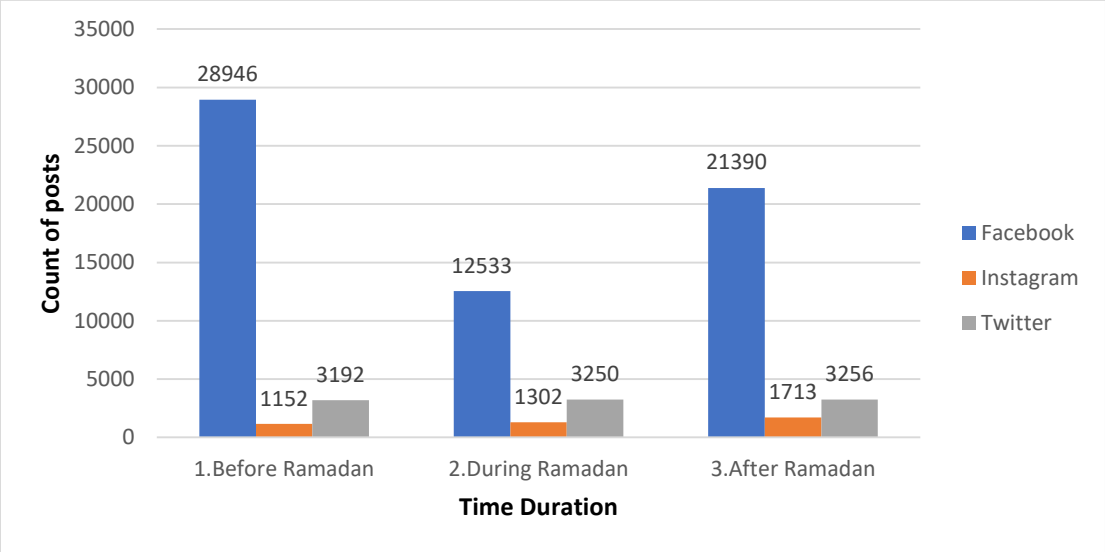


Figure-3: The count number of posts on each social network during the three months.

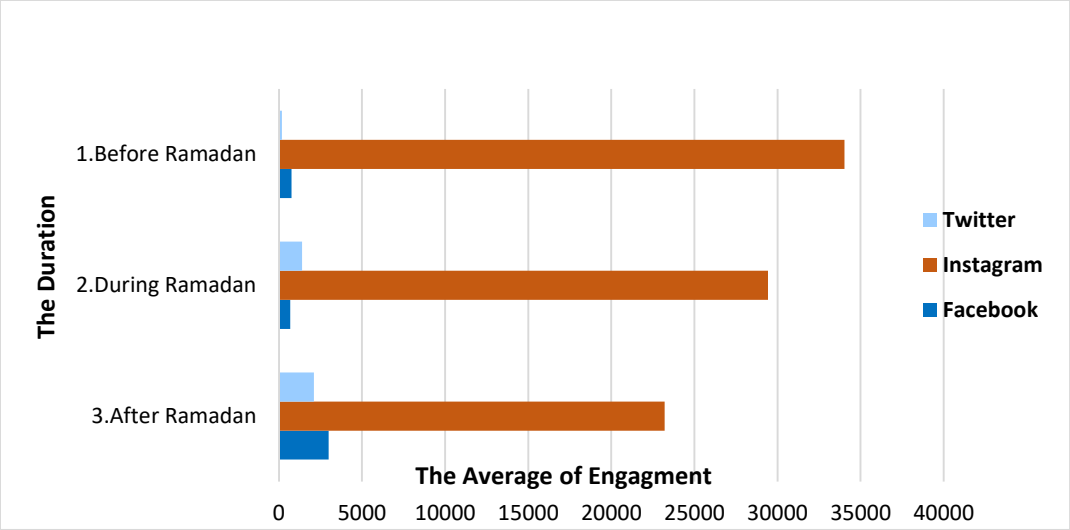


Figure-4: The average of engagement for each social network in the duration

.

4.2 Media Analysis:

The media or content type is the most relevant feature that strongly affects users' interaction with posts on social media. This part illustrates a deep analysis for the content type (Photo, Link, Video, Status) and clarifies the average of engagement for each post type on the social network (before, during and after Ramadan) as well as how the post type is mutable before, after and in Ramadan. In addition, it shows the percentage rate of each media type in the social network.

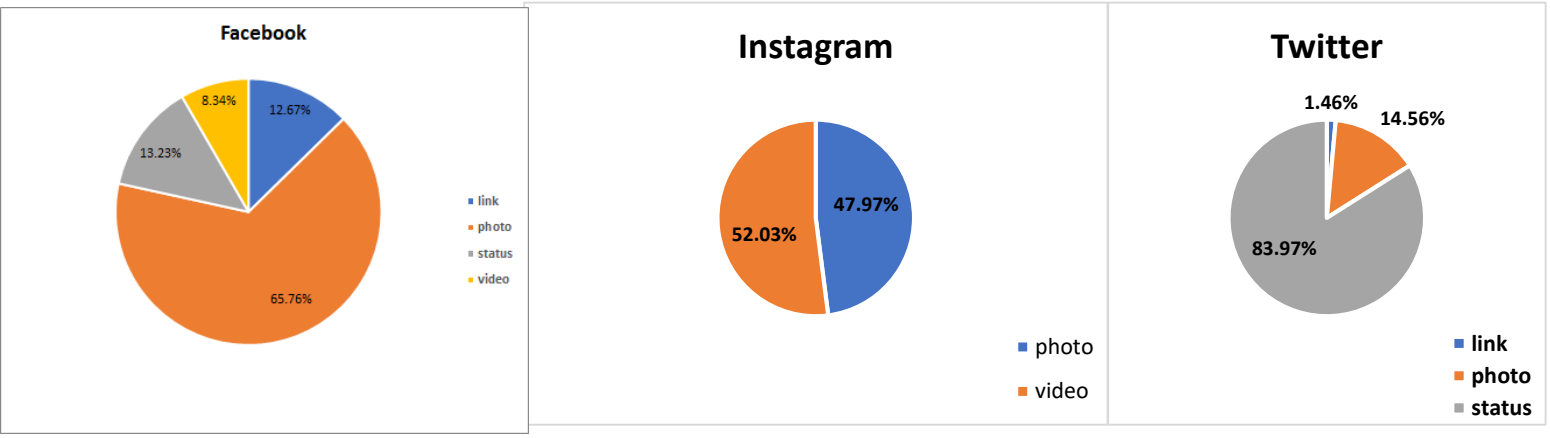


Figure-5: The percentage of media for the network.

Figure 5 demonstrates that photos have the largest portion on Facebook, with 65% of Facebook posts being photos and the rest being textual status updates, links and videos. Although the percentage of photos is totally greater than videos percentage (8.3%) users' reactions on videos ranging 5000-15000 in the three months surpassed the reactions on photos ranging 2000-5000 as identifies in Figure 6.

Instagram is primarily a photo gallery; the percentage of photo posts at around 48% and videos at 52% of content. The maximum engagement on Instagram photos exceeded the 130,000 before Ramadan as Figure 5 explains, then decreased to 120,000 during Ramadan. Afterwards engagement value dropped to 8000 Ramadan. Twitter had the lowest level of engagement and post numbers compared to Facebook and Instagram. The content of tweet or Twitter post can be a status, photo or link. Most of Twitter posts as shown in Figure 5 were status updates at 84%, 14.5% for photos and 1.5% for links. Most of the time the interaction of users on status was larger than on photos during and after Ramadan but photo engagement raised to1700-2100 in the time between 5pm -9pm before and after Ramadan in each time zone throughout the region.

Figure 6: The average engagement for media type, before, during and after Ramadan for each network.

4.3 Time of Post Analysis:

The analysis in this part illustrates the engagement of Facebook, Instagram and Twitter on weekends and weekdays before, during and after Ramadan. Furthermore, the analysis indicates to the hours which they have the maximum engagement rate in the mentioned period.

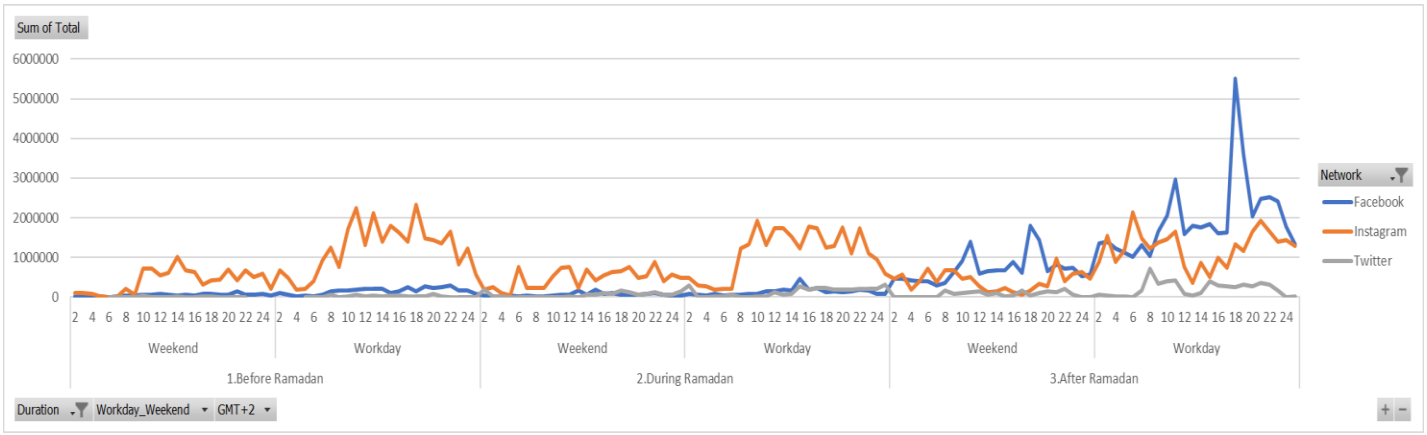


Figure 7: Time series of engagement rate in different networks.

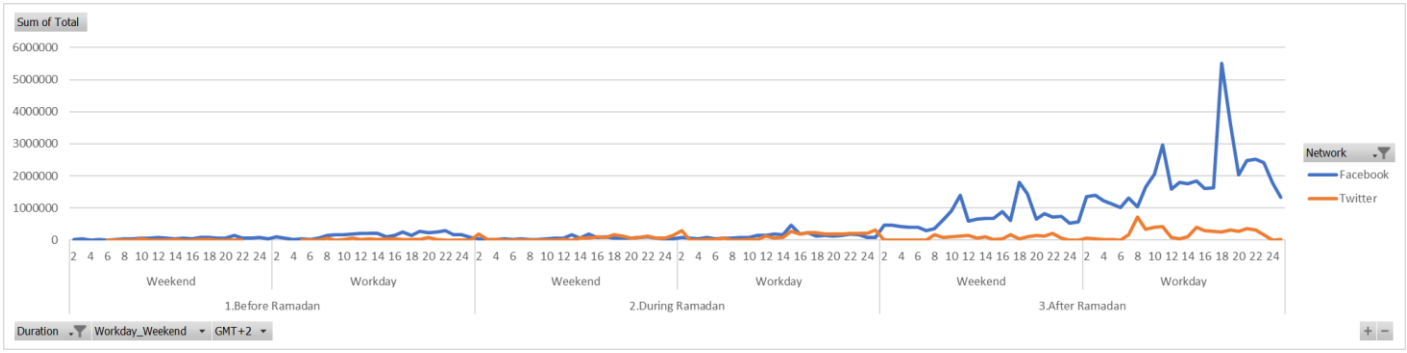


Figure 8: Time series of engagement rate Facebook and Twitter.

Figure 7 and Figure 8 describe the expected engagement of social network in the Arab world. It should be noted that social media users are more engaged and active during the workdays than the weekends. Use during daylight hours according to the analysis clarifies that on the weekends Instagram users start engaging with posts at 7am the engagement rapidly increases to 220K at 11am. Afternoon engagement starts at 1pm until 5pm to reach the peak of 230K, then the value starts to drop to 180K at 11pm. This behavior almost identical in the period before and during

Ramadan. After Ramadan, when the summer vacation starts, the interaction is different. Engagement starts between 12am-6am to be 460K-210K, then the engagement gradually decreases to 35K by 1pm.

At the weekends, the general engagement is lower than on weekdays. User interaction starts at 6am and the engagement increases until 2pm to reach 101K. After Ramadan, the users' behavior on Instagram differs slightly on weekends; engagement starts at 4am and remains at a high level until 6am (19K-70K) then the value decreased to 5k (7am-20pm).

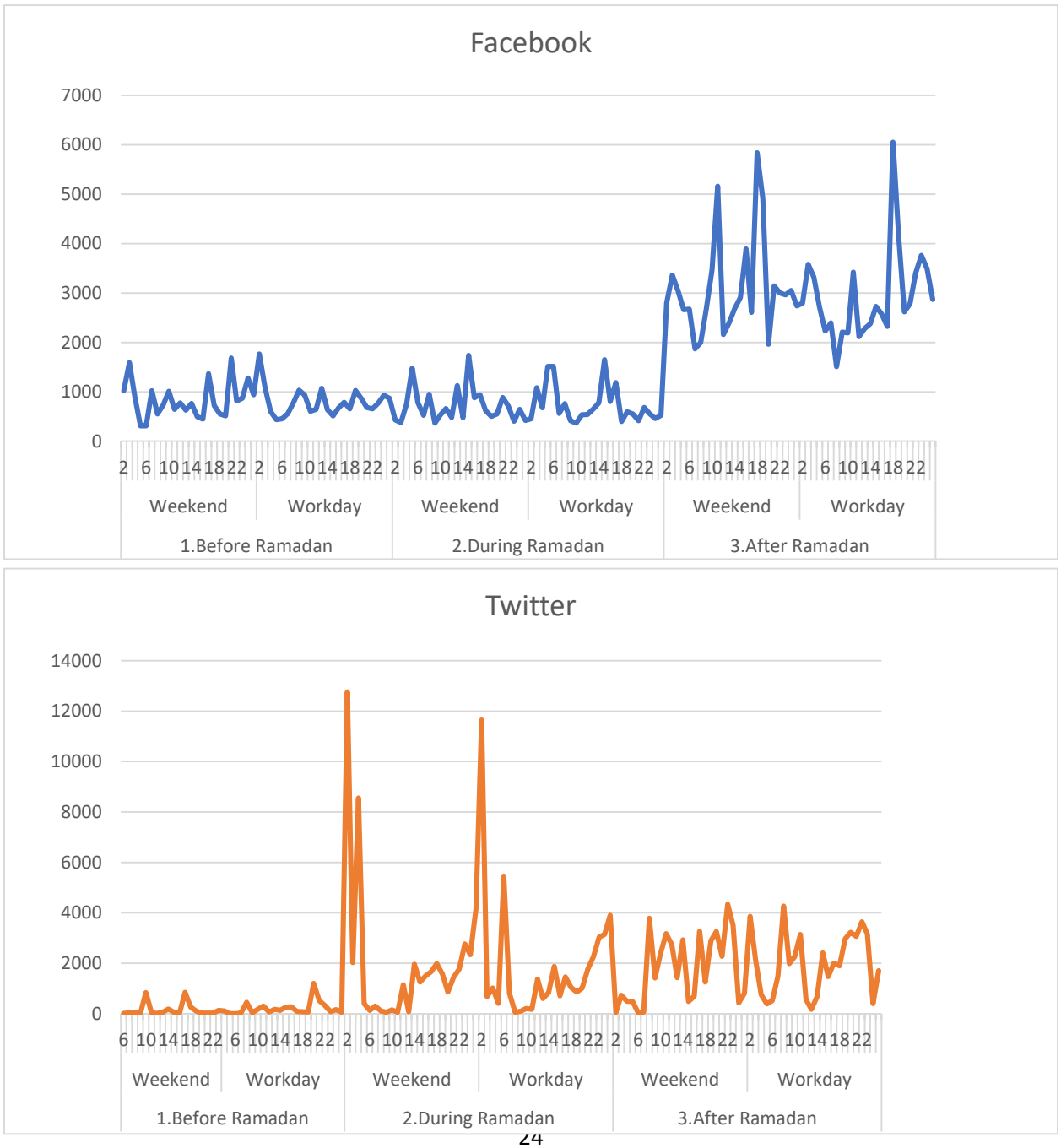


Figure 9: Engagement on Facebook and twitter through the day hours.

In Figure 9 there is a closer view for Facebook and Twitter engagement rate. On Facebook, the behavior is almost the same as during weekends and workdays in the period before and after Ramadan. Wherever the average engagement is unsteady, the interaction starts at 5 am in working days and the engagement is increases until 9am. After Ramadan, the engagement starts in the afternoon on weekends at 12 pm and increases to be 3.3K at 4am with the value decreasing to 1K at 7am. Users are back to engage more from 9am -5pm to reach the peak where the engagement exceeds the 5.8K level. Before Ramadan, the engagement rate on Twitter posts predominantly stable on the weekends and the weekdays, the engagement in this period increasing over the course of the afternoon, between 4pm and into mid evening or 8pm.

It is worth noting that during Ramadan users start engaging with Twitter posts at midnight because of the lifestyle of Ramadan and sahour time and the engagement levels achieve the maximum rate 12K at the weekend and 11K during weekdays. Figure 10 summarizes the average engagement in the different periods, Before Ramadan and During Ramadan Facebook engagement started at midnight and dramatically increased to 2500 at 4am the engagement decreased to 750 at 8am, after that it grew to be 2K at 22pm. Engagement on Instagram started early morning before and during Ramadan, but after Ramadan the maximum engagement reached at 2 am. Surprisingly Twitter posts had zero engagement on some tweets, this happened specially at early morning hours. it's good to know that the engagement on Twitter posts achieve the maximum from 12pm-2pm during Ramadan.

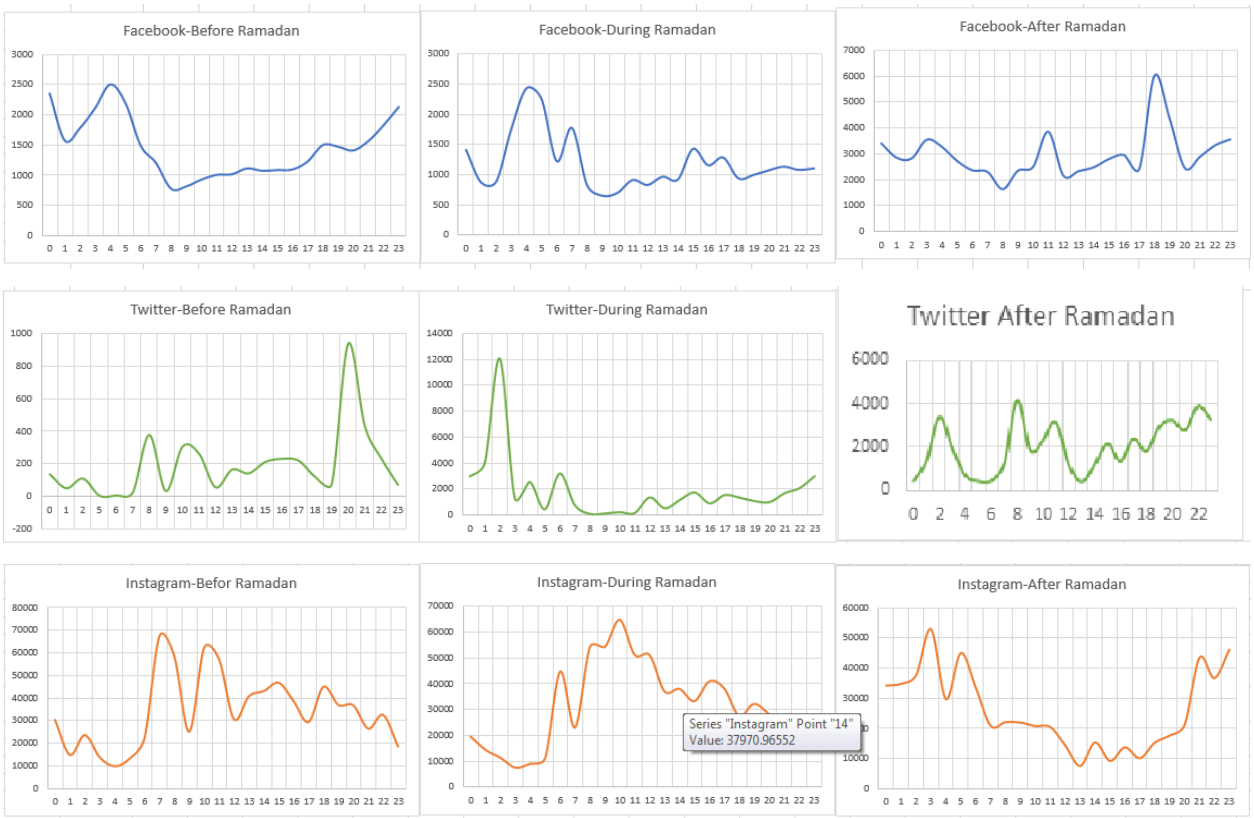


Figure 10 Average engagement in different periods for each network

4.4 Field Analysis:

The page type or the field is an essential factor that affects the engagement on the social network. The field analysis is the interactivity between social media users and the page type (Media, News, Social, Beauty, Health, Shopping, Fashion, Social, alike). This section describes the percentage of users engagement based on the page field before, during and after Ramadan and discusses the media type (Photo, Video, Status, Link) effects as well.

Media type is a crucial factor and strongly affects the engagement. In Figure 11 it is evident that Beauty pages were highly engaging before and after Ramadan (this can be explained due to Ramadan habits) makeup is not preferred through fasting. Furthermore, Celebrity and Fashion pages have high percentage rates in the same period. Moreover, during Ramadan Fashion pages dominate the engagement with the rate at 80% because of preparing for Al Eid and for the summer vacation; this also clarify why Shopping pages had more engagement before Ramadan and after

Ramadan. Health pages have a significant portion during Ramadan around 4% and the percentage after Ramadan clearly to be 7% whereas people after Ramadan started a healthy life style to lose the gained weight. Interestingly, celebrities were less active during Ramadan while the engagement rate is only at 6%.

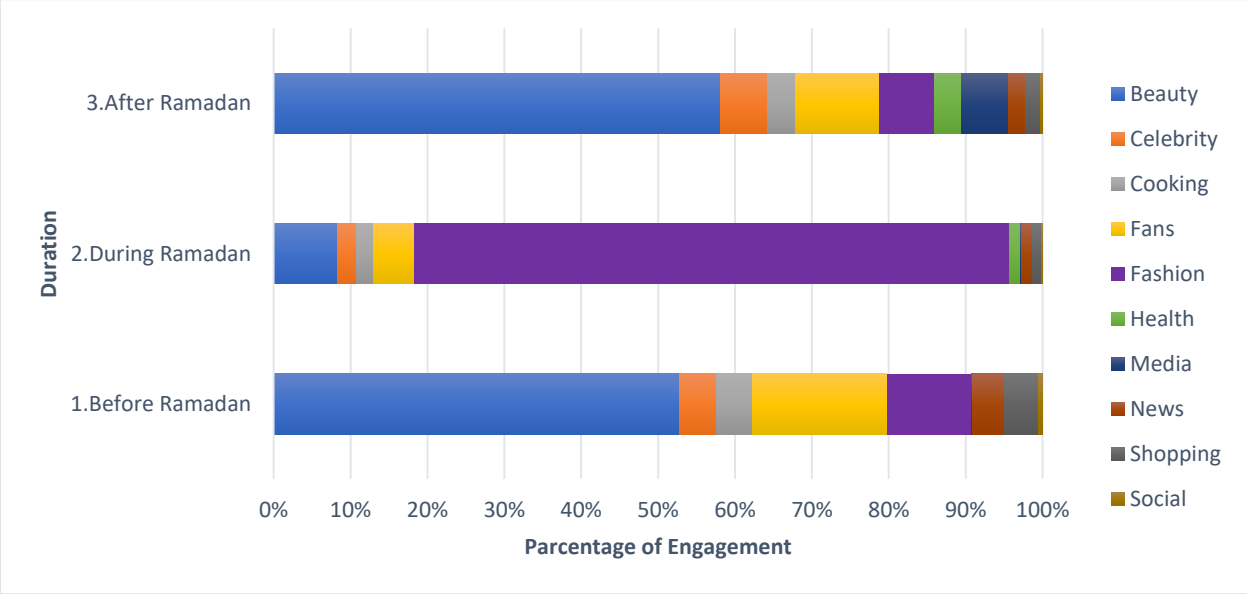


Figure 11 The percentage of engagement for the fields

In Figure 12 it's surprising that videos have the greatest engagement in most of the fields compared to photos, despite of the number of photo posts being greater across all networks as the table shows. Cooking, Beauty and News pages posted videos more than photos; maybe this explains why engagement with Instagram for videos is greater than photos. Wherever most of the pages for Instagram in the dataset are Beauty and Fashion pages. Referring to Figure 6 the highest engagement on videos was between (1pm-4pm), as cooking posts were videos this period the housewife engaging on social media for cooking ideas that time. also beauty (make up tutorials) are videos this reason explains the high percentage of videos in Beauty pages. Photos have larger engagement in Shopping, Media, Celebrity and Fashion pages compared to links and status updates. On the Health pages, the engagement on photos and videos were convergent. news pages. Finally, the status posts have engagement in news and media posts with almost a convergent percentage.

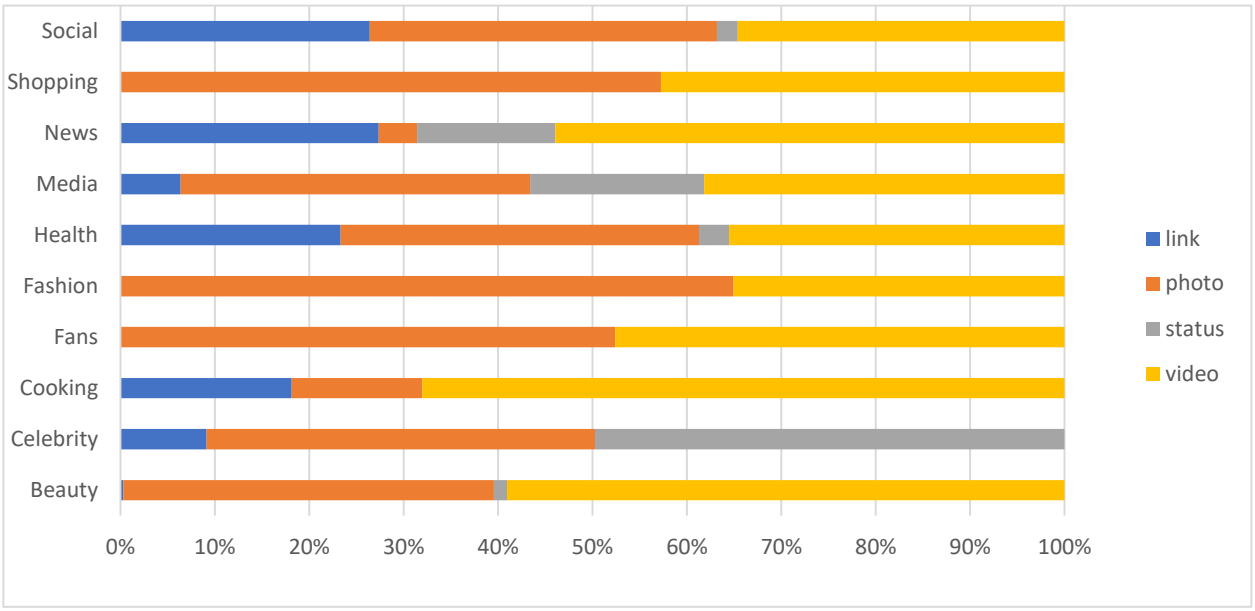


Figure 12 The Percentage of users' engagement cross the field type and media.

4.5 Facebook Reactions Analysis:

Figure 13: Facebook reactions.

To better understand how users emotionally engaged with posts, Facebook reactions provide this opportunity. This analysis will provide better understanding for the different emotions and the impact of reactions on the post content. The analysis in this part covers two parts, in the first part it will discuss the total of reactions based on pages fields, while the second part shows the total reactions based on day hours before, during and after Ramadan. Firstly, as Figure 13 shows there are six types of post reactions in Facebook this study analyses two reactions Positive (Like, Love, Haha, Wow) reactions and Negative (Sad, Angry) reactions.

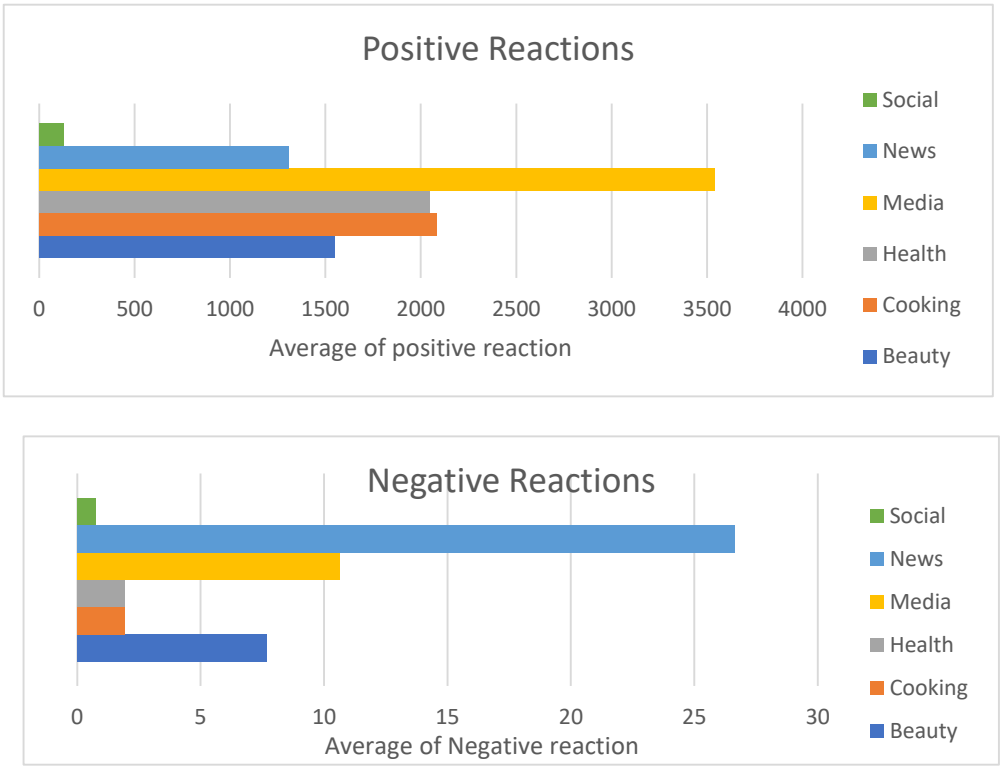


Figure 14: The average reaction based on the field.

In general, the Positive engagement on the posts is greater than negative engagement. Figure 14 shows that the average of Positive reactions on the media pages reach 3500 in Media pages, 2000 in Health and Cooking posts. From the other side the maximum negative reactions on the News posts 26.6, Media posts also obtain 10.6 negative reaction. It's noticed that most of users are Positive on Health, Cooking and Beauty pages.

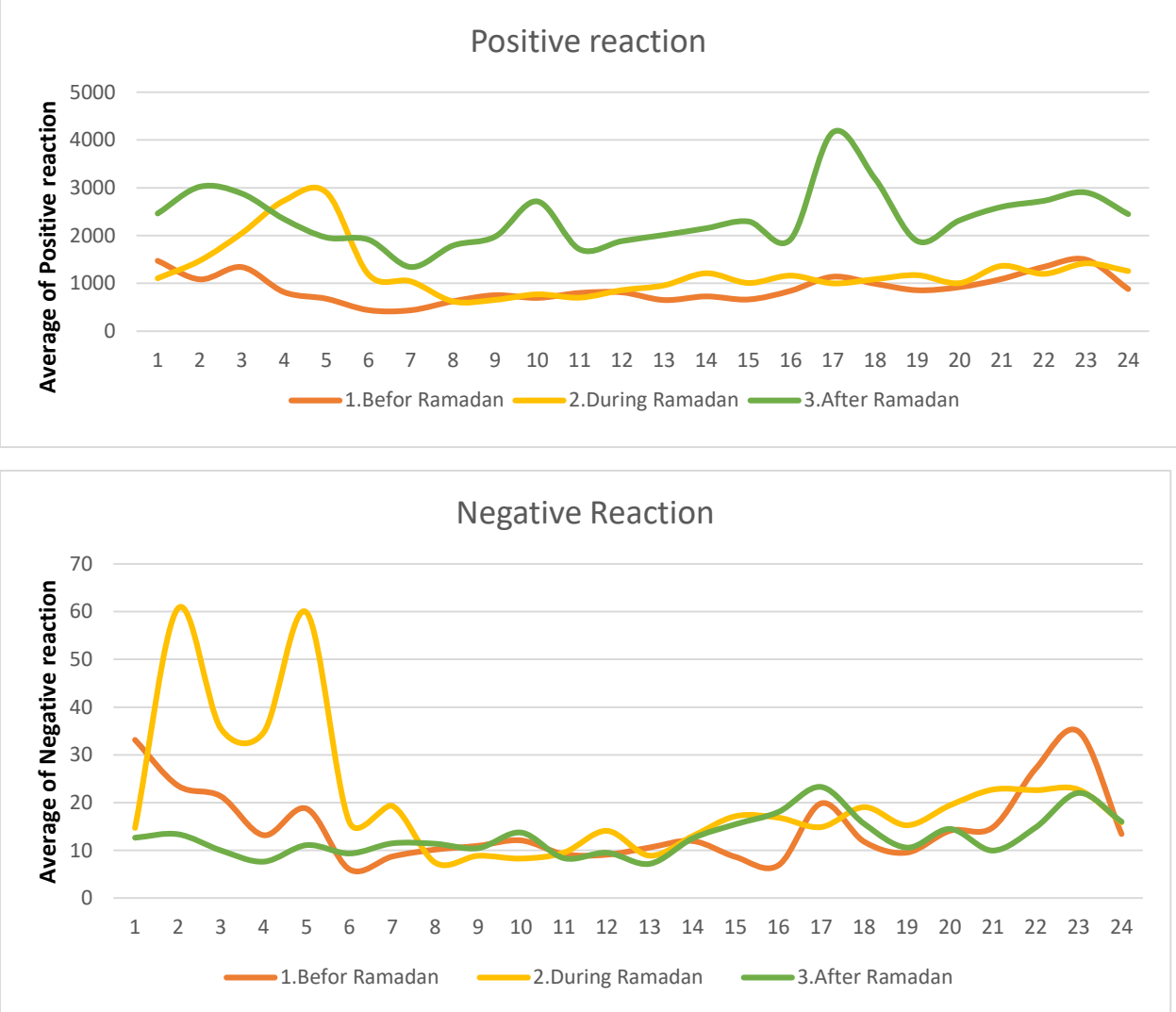


Figure 15: Average of reaction cross the time and duration

As Figure 15 clarifies, it is before and during Ramadan the Positive reaction almost have the same value 500-1200 starting from 8am till the end of the day. Users have a somewhat unstable mood during Ramadan at the period (1am-12pm), where the negative reaction rises and drops. Before

Ramadan, the sadness drops slightly from (1am to 6am). From 7am-1pm the behavior almost steady.

Users during Ramadan express the maximum Positive engagement at midnight and reaches 3000 at 5am. The Positive engagement is the dominant response after Ramadan, with the average number of reaction starting at 2500 in the night rising to 3000 at 2pm. Positive engagement gradually decreases to 1900 at 4pm. Suddenly the engagement reaches the peaks to 4000 at 5pm. Conversely, the negative reaction stabilizes during the day, however the sadness reaction lightly increases in the afternoon between 2pm-5pm and at night between 9pm to 11pm.

Chapter 5: Clustering and Classification Models with performance evaluation:

This chapter also presents the classification techniques to predict the engagement of post based on post type (Photo, Video, Status and Link) and network. Classification is used to predict the possible consequences of user's engagement on the different selected pages from the Arab world. this section covers the performance of different classification models that applied on the available data set. It also describes different used models to detect the best predictive model for the data set. A decision tree with different criteria such as Gini index, gain ratio and accuracy. Naïve Bayes, K-Nearest Neighbor. Each model is evaluated based on the accuracy and performance of each model.

5.1 Data Classification:

Classification is a machine learning task in data mining that is used to define which set category of the new observations belong to, depending on known categories and the training set of data that contains the observations (J Tang, S Alelyani, H Liu 2014).classification is used to classify the items into features with respect to the predefined classes (Tina, Patil, & Sherekar 2013).Supervised classifiers are usually used and designed with labeled data set, classifiers assumed the training data set is correctly labeled and available. The algorithm generates a model to minimize the error in prediction as much as possible (Mukhopadhyay et al. 2014). In Figure 16, the general process of data classification is usually consisting two phases, training phase and predicting (testing) phase. In the training phase data is labeled and analyzed based on feature (attribute) selection; these

attributes may either be ordinal, categorical integer number or real number. After extracting these attributes data will be representing to labels. In the testing phase data will be represented by the

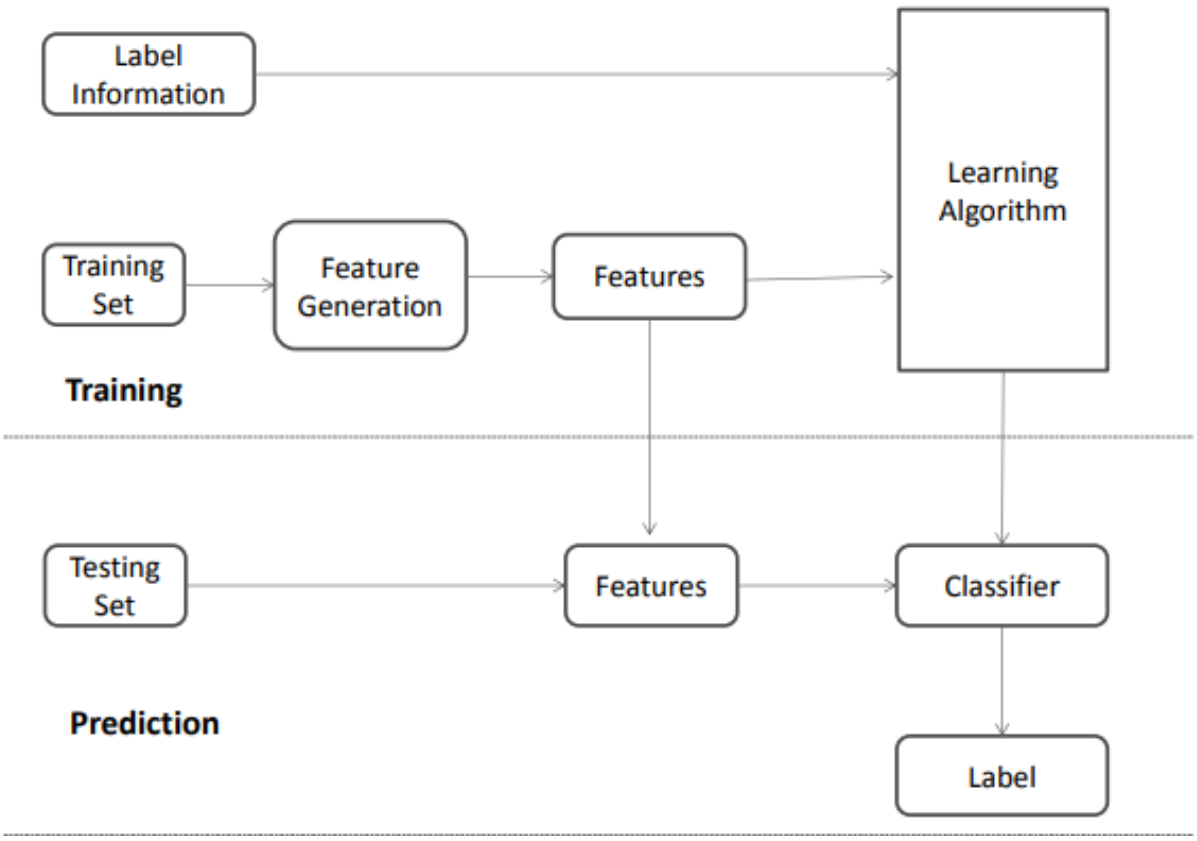


Figure 16 General process of data classification (Adapted from J Tang, S Alelyani, H Liu 2014, p.3)

5.2 Feature selection (extraction):

Feature selection is the process of finding relevant and meaningful input or reducing the inputs for better processing and analysis. It is considered an important part of machine learning because it extracts useful information from the existing data to build a valid model. A data set usually contains more information and attributes than is needed, so applying an unsupervised algorithm like clustering helps to summarize and collate the data into groups (or clusters). Data in the same cluster behave in similar ways in contrast to that in another cluster.

5.3 Confusion Matrix:

Confusion Matrix: Is a table layout that visualizes the performance of the classification model by showing the number of predicted and the actual set. A Confusion matrix helps to define the correct predicted sets and the incorrect sets, as Figure 17 shows. The accuracy of the model can be calculated from the Confusion matrix by the following formula:

Accuracy = Correctly Classified Points / Total Number of Points.

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (TP + TN) / (P + N).

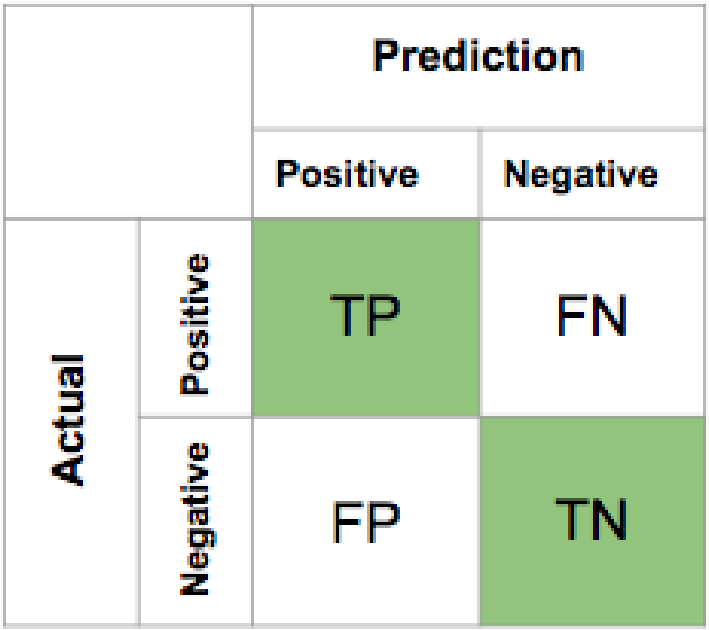


Figure 17: Confusion matrix

The results contribute in choosing the best model for classification to meet the research objective.

K-cross validation is a familiar evaluation technique for machine learning and data mining. It splits the dataset randomly into k parts one part for testing the classifier and the other parts for training. This process is repeating for k-1 times, each time the dataset is splitting randomly. Then the average accuracy is evaluated on average of each iteration (Wa'el, 2015).

5.4 Clustering:

Clustering is a technique for extracting information from unlabeled data. It is used to describe the data set and to arrange the cluster, based on the engagement in this study. Clustering is useful in grouping the objects that are similar to each other and with the dissimilar conferred to another group. K-means is a widely used algorithm for clustering the objects by measuring the Euclidian distance between objects. Usually K-means algorithm applies on the numerical values, so that I used the operator “Nominal – Numerical” to change the non-numerical attributes into numeric type. It also maps all values of these attributes to numeric values. On the other hand “Replace Missing Values” operator is used to fill the empty cells of selected attributes by a specified replacement with minimum, maximum or average value. “Optimize Selection” operator also used to enhance the selection of the most relevant attributes. In Figure 18 shows that data is clustered among the engagement into four groups, the scatter displays the clusters among the engagement. Cluster-3 shows the lowest engagement rate, cluster-1 shows the moderate engagement, while cluster 5 indicates to the highest engagement and cluster-0 contains the low engagement rates.

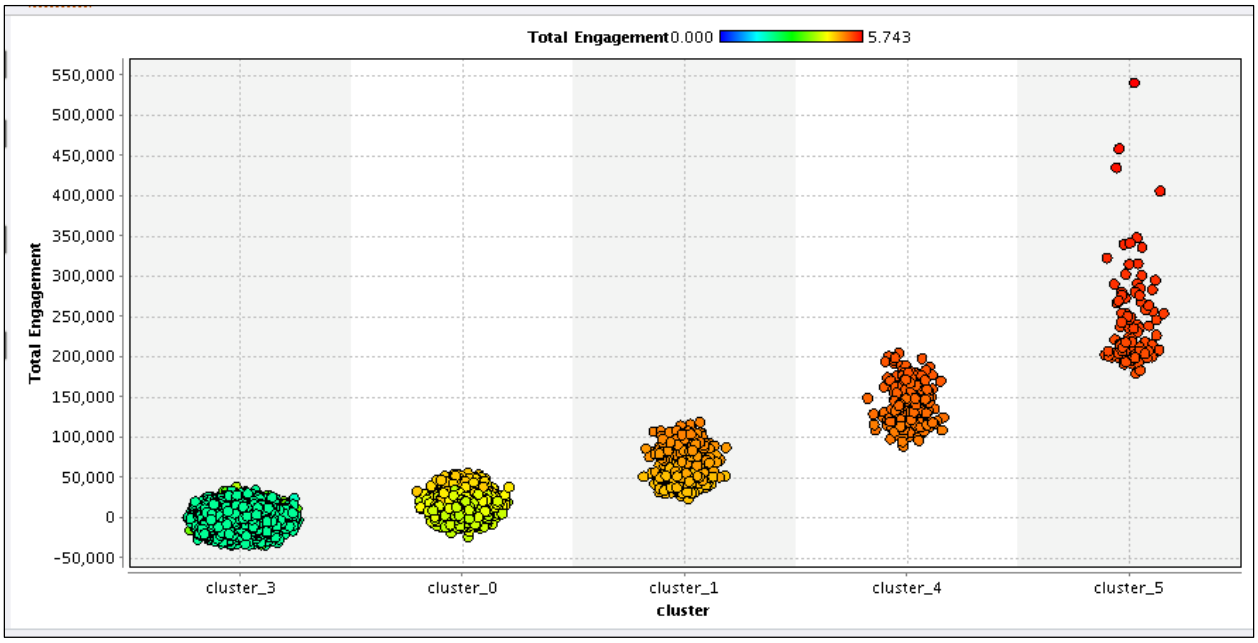


Figure 18: Scatter plotter for clustering.

5.5

Decision tree is a familiar classifier that used widely in the classification task based on choosing the attribute that maximize and fix the data division. This attribute is used to split the data into branches and recursively repeat the splitting until the classification reached (Ray et al. 2014).

Decision tree can be performed and interrupted easily, while the attributes are tested on each node it's also considered as a visualization technique. There are three hierarchal nodes in the decision tree:

- Root node: the beginning attribute that used to start the division of data
- Internal node: the attribute that used to divide the data set
- The terminal node (leaf): presents the predictive variable.

Overfitting is a problem in the generated decision tree, it may be low accuracy in classifying the testing set or the tree has a lot of branches due to the outlier value or noise data. To optimize the resulted tree and solve the overfitting problem pre-pruning or post-pruning the only approaches to avoid overfitting. Pre-pruning: not to split the node if weight below the threshold, post-pruning: to remove the branches form the tree and each time to check the accuracy (like removing the small values). Data is splitting in decision tree based on the homogeneity of data. There are four criteria to split the data for decision tree: Information gain, Gini index, Gain ratio and the Accuracy.

Information gain: is the difference between the information before splitting and after splitting. Information gain is choosing the attribute with the higher gain values as leaves nodes.

Gini Index: is the impurity measurement for the categorical variable that is used in building the tree. The attributes with high uncertainty in reducing gain value through splitting will not be selected. Therefore, it takes the size and number of branches into account.

Gain ratio: is an adjustment for information gain to reduce the bias by

Accuracy: enhance the attributes that increase the accuracy.

5.5.1 Decision tree for the dataset:

Because of the variance between the total engagement on Instagram posts comparing to total engagement on Facebook and Twitter posts I split the dataset into two sets (Instagram set, Facebook and Twitter set). I used Rapidnier tool to apply the decision tree on Instagram dataset to predict the total engagement category (Very low, Low, Moderate, High, Very high). This model support manager ‘s decisions on whether to publish the post, when to publish and using which network. In Figure 19 it is clear that most of Instagram’s posts got very high engagement or low engagement. While 80% of the posts got very low engagement.

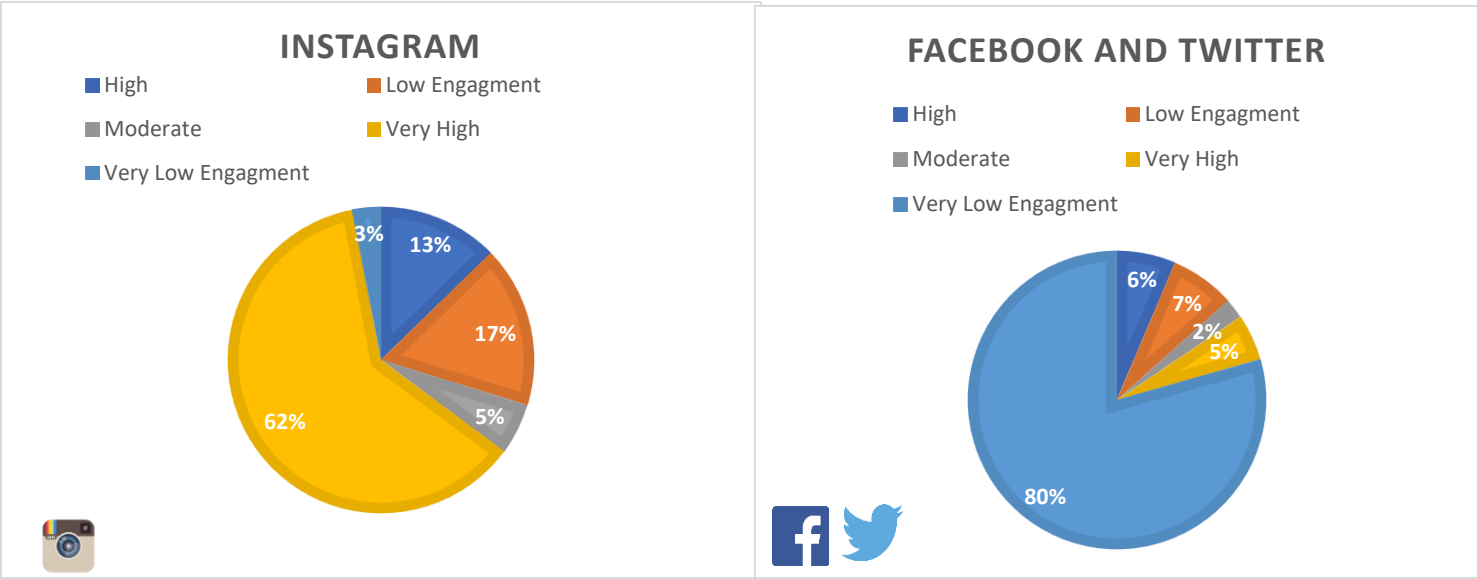


Figure 19 The networks engagement among the categories

The results of decision tree in Figure 20 shows that Beauty and Fans pages got the very high engagement, however the Fashion pages got high engagement on videos after Ramadan and moderate level after Ramadan. Media posts had high engagement; cooking pages had very high engagement after Ramadan on Instagram and low level after Ramadan as well as shopping pages.

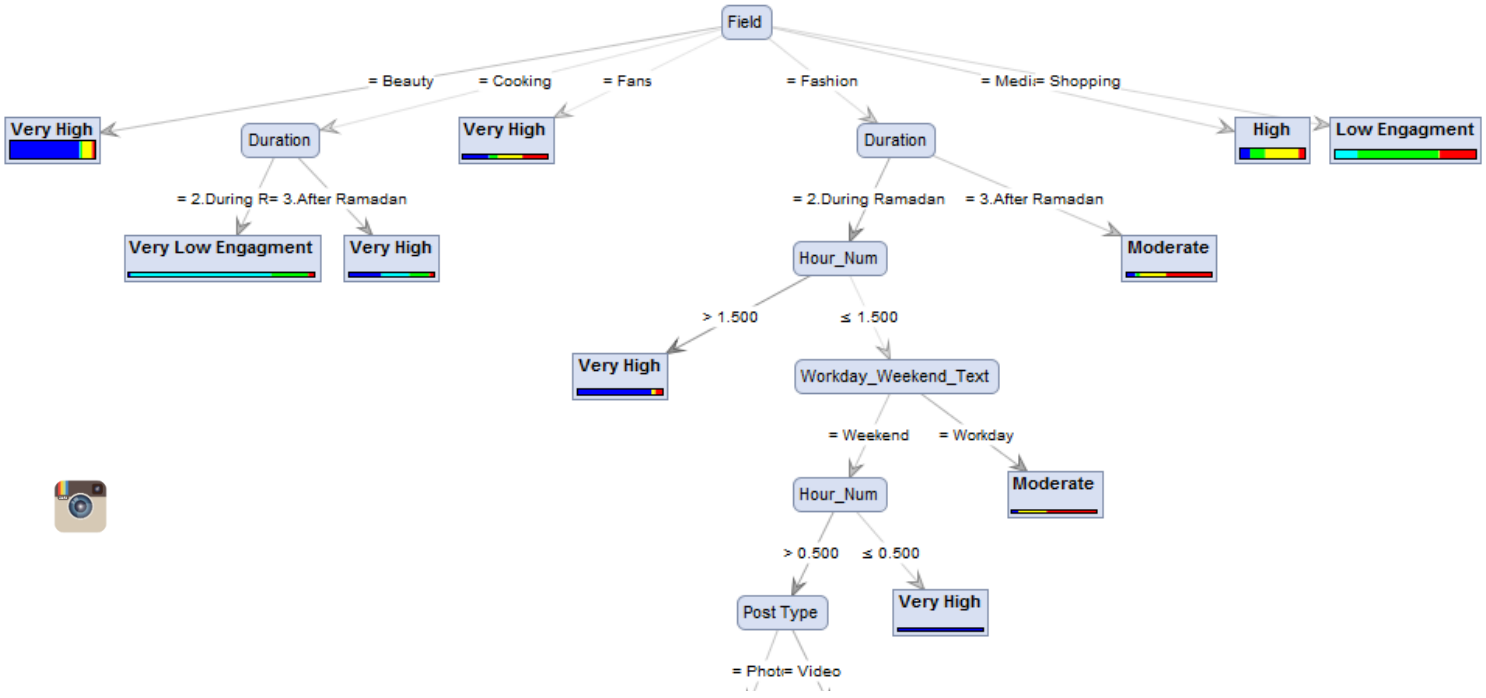


Figure 20 : Decision tree for Instagram dataset

The Decision tree for Facebook and Twitter explains in Figure 21 that in general Beauty pages had low engagement on Twitter and Facebook. In Figure 21 cooking videos had high engagement on Facebook before, during and after Ramadan. Followers on Twitter were very active especially on their photos and status, so the engagement is very high as Figure 21 shows. In particular users interact highly on Health pages in Facebook mostly on links type during and after Ramadan. Also, the videos of news pages were effective on Facebook rather than Twitter. Appendix II shows the full decision tree for Facebook and Twitter

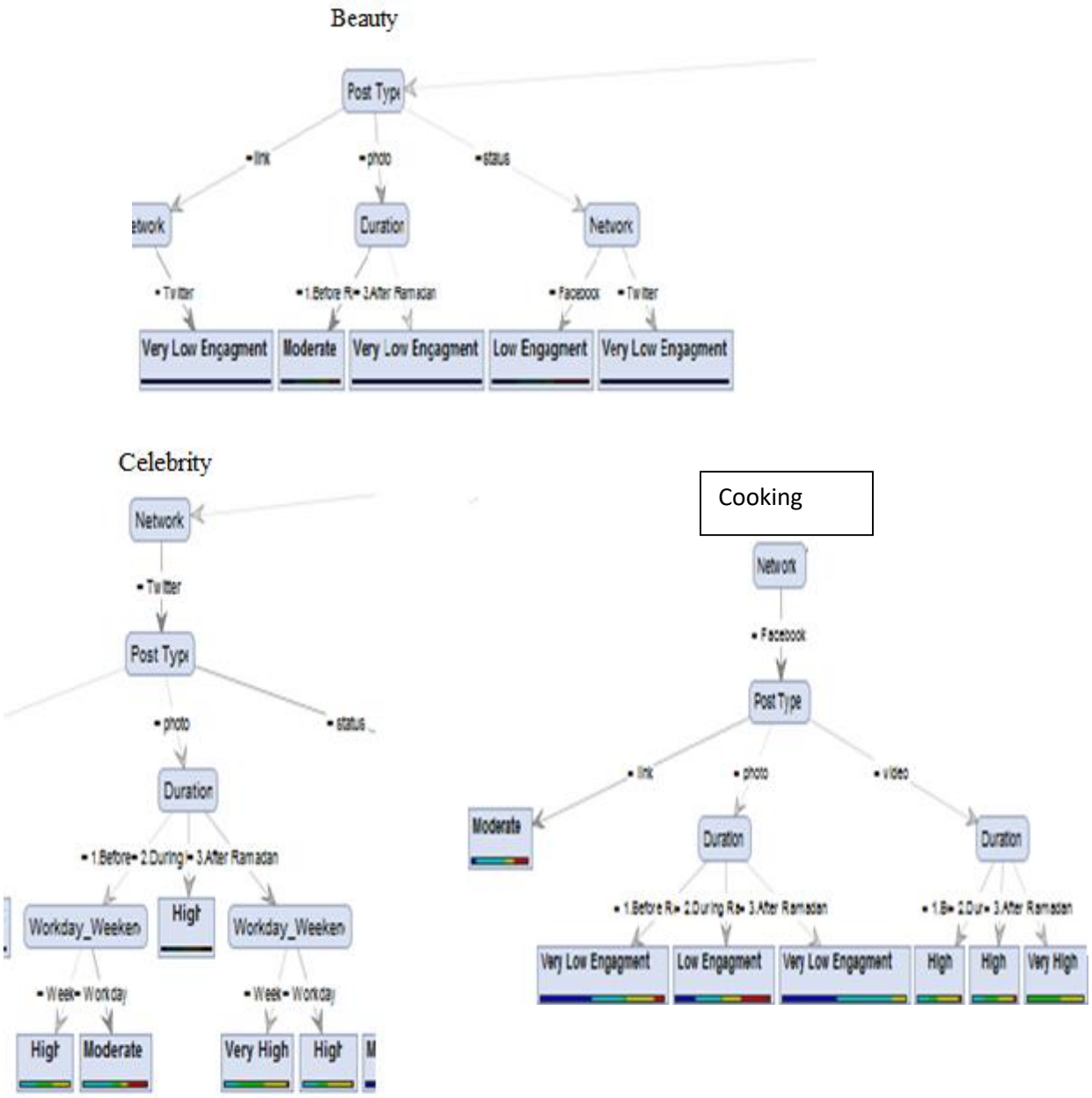


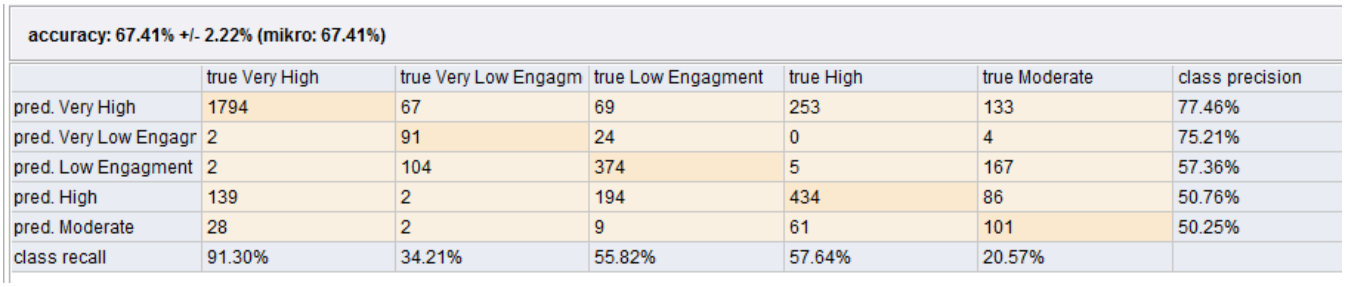
Figure 21 Decision tree examples

5.5.2 The performance evaluation for Decision tree model:

Evaluation for decision tree is depending on the accuracy of built model, as mentioned in part 6.5 decision tree algorithm has four important criterions to split the dataset into training and testing set. In the classification model. I applied the four criterions on each dataset (Instagram dataset, Facebook and Twitter dataset) to choose the maximum accuracy for each set, as long as accuracy affects the model performance. In table-4 it's clear that the accuracy of Instagram with gain ratio criteria has the best accuracy comparing to the other criterions, so the model performed using gain ratio and the accuracy of the model clear in Figure 22a For Facebook and Twitter dataset Gini index was the best criterion to perform in the model with accuracy 84.2% as Figure 22b describes.

Network	**The Accuracy of classification model**			
Decision tree criterions	Gain Ratio	Information Gain	Gini index	Accuracy
Instagram	67.4%	67.3%	64.8%	66%
Facebook &Twitter	83.8%	83.8%	84.2%	81.2%

Table 4 The Accuracy of Decision tree with different criterion



accuracy: 67.41% +/- 2.22% (mikro: 67.41%)

	true Very High	true Very Low Engagm	true Low Engagment	true High	true Moderate	class precision
pred. Very High	1794	67	69	253	133	77.46%
pred. Very Low Engagr	2	91	24	0	4	75.21%
pred. Low Engagment	2	104	374	5	167	57.36%
pred. High	139	2	194	434	86	50.76%
pred. Moderate	28	2	9	61	101	50.25%
class recall	91.30%	34.21%	55.82%	57.64%	20.57%	

Figure 22a: Model Performance for Instagram.

accuracy: 84.29% +/- 0.17% (mikro: 84.29%)

	true Very Low Engagm	true Moderate	true Very High	true High	true Low Engagment	class precision
pred. Very Low Engagr	69025	536	270	472	586	97.37%
pred. Moderate	601	3537	2085	2600	1070	35.75%
pred. Very High	106	283	1344	1061	148	45.68%
pred. High	612	787	785	1333	371	34.28%
pred. Low Engagment	645	872	44	313	1197	38.98%
class recall	97.23%	58.80%	29.68%	23.07%	35.50%	

Figure 22b: Model Performance for Facebook and Twitter.

5.6 Naive Bayes Classifier:

Is a statistical algorithm for predicting the probability of given sample to which class belongs? Naive Bayes assumes that the effect of features values for a class is independent of the values of the other features. It is effective in large dataset and exhibit high accuracy and the performance of the model can be comparable with decision tree performance. When applying the model with Naïve Bayes algorithm on Instagram dataset the accuracy was 65% and on Facebook Twitter dataset 75.9%

5.7 K-Nearest Neighbor Classifier:

Is unsupervised learning model that use the similarity between objects for classification. The similarity is measured based on Euclidian distance between the training objects and the testing objects. It is simple and direct classification method. It is simply find the k training examples that are closest to the unseen example. The performance of K-NN is evaluated also through the cross validation to check the accuracy of the model. 55.2% was the accuracy for the model on Instagram dataset and 78.9% the accuracy for Facebook Twitter dataset.

Network	**Classification models Accuracy**		
	Decision tree	Naïve Bayes	K-NN
Instagram	67.4%	65.4%	55.2%
Facebook & Twitter	84.2%	75.9%	78.9%

Table 5 The Accuracy of the model using different algorithms

Chapter 6 Discussion and Research questions answers

This chapter discusses the results for posts analysis from different perspectives and shows the effects on users' engagement in different times. In addition, it describes how data mining algorithms used to predict the users' interaction on the posts. Research questions are answered in this part as well.

6.1 Post analysis results and Research questions answers:

Research Question 1: Which social network application is the best for advertising?

The post analysis contributes in understanding the effects of published posts using social media on users' engagement and interaction. In addition, this analysis represents base for marketing persons to advertise on social media based on knowledge. After analyzing the posts the results are listed below:

As table 6 shows the Average engagement differs between the social networks with respect to durations. Instagram has the largest portion from user's engagement in the Arab world, despite of the count number of posts comparing to the other network Instagram was the dominant social application. Facebook became in the second level, users respond before and after Ramadan more than Twitter users. Twitter users were active in Ramadan more than at any other time.

Duration	Network	Count of posts	Avg. Engagement
Before Ramadan	Facebook	6470	764.5
	Twitter	3192	180.9
	Instagram	1152	34028
During Ramadan	Facebook	6971	675
	Twitter	3250	1400
	Instagram	1302	29422
After Ramadan	Facebook	21390	2998
	Twitter	3256	2099

	Instagram	1713	32215

Table 6 Average engagement for different networks

Research Question 2: What is the best media type for the post attracts the users?

- In addition to social network effects on engagement, post media type considered as the most effective feature. Table 7 summarized each media type with average engagement on each type. Videos were the most attractive post at all, although videos are not type of Twitter ‘s posts. Photos were in the second level, Instagram was the most application that posted images. Were the average of engagement reached 36K.links status were used by Facebook more than Twitter, Twitter links had more engagement in Ramadan rather than Facebook ‘links. Users on Twitter engaged mostly on status posts. While Facebook users engaged highly on status after Ramadan.

Duration	Network	Photo	Video	Link	Status
Before Ramadan	Facebook	1703	3039.2	842.5	53.3
	Twitter	123	NA	21.3	311.6
	Instagram	36620	3215	NA	NA
During Ramadan	Facebook	1285	3747.3	1172.3	47.3
	Twitter	817	NA	133.6	1979.6
	Instagram	28340	30470	NA	NA
After Ramadan	Facebook	4036	NA	1448.9	2064.7
	Twitter	1484	4715.5	122.2	3061.6
	Instagram	19158	30382	NA	NA

Table 7 Average engagement for media post in different networks

Research Question 3: When is the best time to publish the post for maximum engagement?

- Network types and post media types were the most effective factors on users’ engagement; also, the analysis above included the role of time and page field. As a result, for the analysis

users were active in working days more than weekends. On the other hand, pages fields had significant impact; Beauty pages had the maximum engagement in the duration before Ramadan and after Ramadan, while fashion pages reached the highest engagement in Ramadan. Celebrities' pages had the maximum engagement before Ramadan.

- Analyzing Facebook reaction showed that the average of positive reactions was much more than the negative reaction, after analyzing the reaction based on page field the results displayed that positive reactions were more in media, health and cooking fields. Regarding to negative reactions, the engage was on news posts, media posts had negative reactions too.

Research Question 4: How to predict the engagement before publishing the post?

- In the classification phase the experiment was done using decision tree with different creations, KNN and Naïve Bayes algorithms and based on the accuracy decision tree algorithm was the best to use in this model.

- After splitting the dataset into two sets one for Instagram and the other for Facebook and Twitter together, the predictive model for user's engagement was successful to predict 83% from the testing set accuracy for Facebook and Twitter data using the decision tree algorithm using Gini index criteria was 84.2%. And for Instagram dataset the accuracy was 67.4% using the gain ration criteria.

Chapter 7 Conclusion and Future Work

7.1 Conclusion

This research focused on analyzing social media posts from different networks to predict the engagement on the post using data mining. More over decision tree was a successfully employed in this model. Around 48K posts were used in this model but the process accomplished by splitting the data into two sets because the variance between Instagram data was larger than Facebook and Twitter data. The models were evaluated using confusion matrix with accuracy of 67.4% for Instagram and 84.2% for Facebook and Twitter.

This study showed how post type, post media and the posting time analysis provide insights about social media engagement. The advantage of using these analyses linked to business marketing.

7.2 Future work

As a future work, it is planned to consider another data mining technique like text mining to analyze the content of comments to improve the engagement and the prediction model. The proposed analysis considered the comment as engagement regardless the comment content. This study can be extended by including test classification, to classify the comments positively or negatively. On the other hand, the accuracy for Instagram dataset was 67%, the model can be modified by discretizing the total engagement to more than five categories to reduce the percentage of errors in prediction and increase the accuracy.

References:

Moro, S., Rita, P. & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, vol. 69 (9), pp. 3341-3351

Turnbull, S. & Jenkins, S. (2016). Why Facebook Reactions are good news for evaluating social media campaigns. Journal of Direct, Data and Digital Marketing Practice, vol. 17 (3), pp. 156-158.

C, S., M, C. & G, G. (2013). An Analysis on the Performance of Naive Bayes Probabilistic Model Based Classifier for Cardiotocogram Data Classification. International Journal on Computational Science & Applications, vol. 3 (1), pp. 17-26.

Tang, J., Alelyani, S. and Liu, H., (2014). Feature selection for classification: A review. Data Classification: Algorithms and Applications, p.37.

Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S. & Coello, C. (2014). A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I. IEEE Transactions on Evolutionary Computation, vol. 18 (1), pp. 4-19.

"Feature Selection (Data Mining)". (2017). [Accessed 21 October 2017]. Available at: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/feature-selection-data-mining

Ray, P., Mohanty, S., Kishor, N. & Catalao, J. (2014). Optimal Feature and Decision Tree-Based Classification of Power Quality Disturbances in Distributed Generation Systems. IEEE Transactions on Sustainable Energy, vol. 5 (1), pp. 200-208.

Data Mining Techniques Implementation To Improve Healthcare Among Diabetic Patients.. (2016). Master. The British University in Dubai.

Model Evaluation Measures - Machine Learner 1.0.0 - WSO2 Documentation". (2017). [Accessed 22 October 2017]. Available at: https://docs.wso2.com/display/ML100/Model+Evaluation+Measures#ModelEvaluationMeasures-confusionmatrix

de Vries, L., Gensler, S. & Leeflang, P. (2012). Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. Journal of Interactive Marketing, vol. 26 (2), pp. 83-91.

Moro, S., Rita, P. & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, vol. 69 (9), pp. 3341-3351.

Schroeder, R., Everton, S., & Shepherd, R. (2017). Mining twitter data from the Arab Spring.

Agarwal, S., Sureka, A., & Goyal, V. (2015, December). Open source social media analytics for intelligence and security informatics applications. In International Conference on Big Data Analytics (pp. 21-37). Springer, Cham.

Lavanya, D. & Rani, K. (2011). Performance Evaluation of Decision Tree Classifiers on Medical Datasets. International Journal of Computer Applications, vol. 26 (4), pp. 1-4.

Lovejoy, K. & Saxton, G. (2012). Information, Community, and Action: How Nonprofit Organizations Use Social Media*. Journal of Computer-Mediated Communication, vol. 17 (3), pp. 337-353.

Trainor, K., Andzulis, J., Rapp, A. & Agnihotri, R. (2014). Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM. Journal of Business Research, vol. 67 (6), pp. 1201-1208.

He, W., Zha, S. & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, vol. 33 (3), pp. 464-472.

Turnbull, S., & Jenkins, S. (2016). Why Facebook Reactions are good news for evaluating social media campaigns. Journal of Direct, Data and Digital Marketing Practice, 17(3), 156-158.

"Implementation of Data Mining in Analyzing Social Media Users Personality with Naïve Bayes Classifier: A Case Study of Instagram Social Media". (2016). International Journal of Computer Science Issues, vol. 13 (4), pp. 76-82.

Cvijikj, I. P., Spiegler, E. D., & Michahelles, F. (2011, October). The effect of post type, category and posting day on user interaction level on Facebook. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 810-813). IEEE..

Jayakameswaraiah, M., Babu, M. M. V., Ramakrishna, S., & Yamuna, M. P. (2016). Computation Accuracy of Hierarchical and Expectation Maximization Clustering Algorithms for the Improvement of Data Mining System.

MBRSG - Welcome to the Mohammed Bin Rashid School of Government's Portal". (2017). [Accessed 23 October 2017]. Available at: http://www.mbrsg.ae/home.aspx

Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. IEEE Transactions on Learning Technologies, 7(3), 246-259.

Jussila, J., Kärkkäinen, H. & Aramo-Immonen, H. (2014). Social media utilization in business-to-business relationships of technology industry firms. Computers in Human Behavior, vol. 30, pp. 606-613.

Agnihotri, R., Dingus, R., Hu, M. & Krush, M. (2016). Social media: Influencing customer satisfaction in B2B sales. Industrial Marketing Management, vol. 53, pp. 172-180.

Saxton, G. & Waters, R. (2014). What do StakeholdersLikeon Facebook? Examining Public Reactions to Nonprofit Organizations' Informational, Promotional, and Community-Building Messages. Journal of Public Relations Research, vol. 26 (3), pp. 280-299.

Guo, C. & Saxton, G. (2014). Speaking and Being Heard: How Nonprofit Advocacy Organizations Gain Attention on Social Media. Academy of Management Proceedings, vol. 2014 (1), pp. 16174-16174.

Zuber, M. (2014). A survey of data mining techniques for social network analysis. International Journal of Research in Computer Engineering & Electronics, 3(6).

Hays, S., Page, S. & Buhalis, D. (2013). Social media as a destination marketing tool: its use by national tourism organisations. Current Issues in Tourism, vol. 16 (3), pp. 211-239

Gikas, J. & Grant, M. (2013). Mobile computing devices in higher education: Student perspectives on learning with cellphones, smart phones & social media. The Internet and Higher Education, vol. 19, pp. 18-26.

Orenga-Roglá, S. & Chalmeta, R. (2016). Social customer relationship management: taking advantage of Web 2.0 and Big Data technologies. SpringerPlus, vol. 5 (1).

Sin, K., & Muthu, L. (2015). APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS--A LITERATURE REVIEW. ICTACT journal on soft computing, 5(4).

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

Williams, G., & Mahmoud, A. (2017, May). Analyzing, classifying, and interpreting emotions in software users' tweets. In Proceedings of the 2nd International Workshop on Emotion Awareness in Software Engineering (pp. 2-7). IEEE Press.

Chen, X., Vorvoreanu, M. & Madhavan, K. (2014). Mining Social Media Data for Understanding Students' Learning Experiences. IEEE Transactions on Learning Technologies, vol. 7 (3), pp. 246-259.

Mostafa, M. & El-Masry, A. (2013). Citizens as consumers: Profiling e-government services' users in Egypt via data mining techniques. International Journal of Information Management, vol. 33 (4), pp. 627-641.

Hadi, W. E. (2015). Classification of Arabic Social Media Data. Advances in Computational Sciences and Technology, 8(1), 29-34.

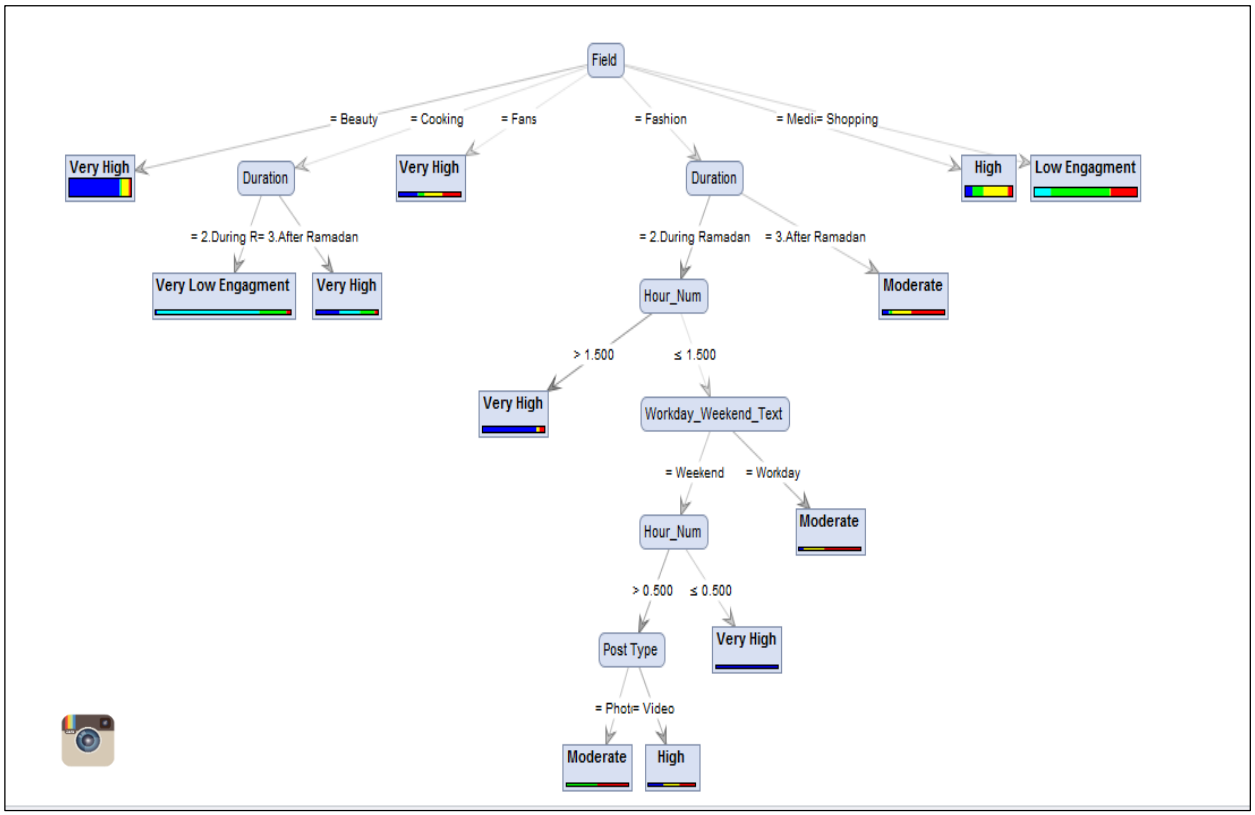
Inoue, G., Shindo, H., & Matsumoto, Y. (2017). Joint prediction of morphosyntactic categories for fine-grained Arabic part-of-speech tagging exploiting tag dictionary information. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017) (pp. 421-431).

Tillmann, C., Mansour, S., & Al-Onaizan, Y. (2014). Improved sentence-level Arabic dialect classification. In Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (pp. 110-119).

Gao, Y., Wang, F., Luan, H., & Chua, T. S. (2014, April). Brand data gathering from live social media streams. In Proceedings of International Conference on Multimedia Retrieval (p. 169). ACM.

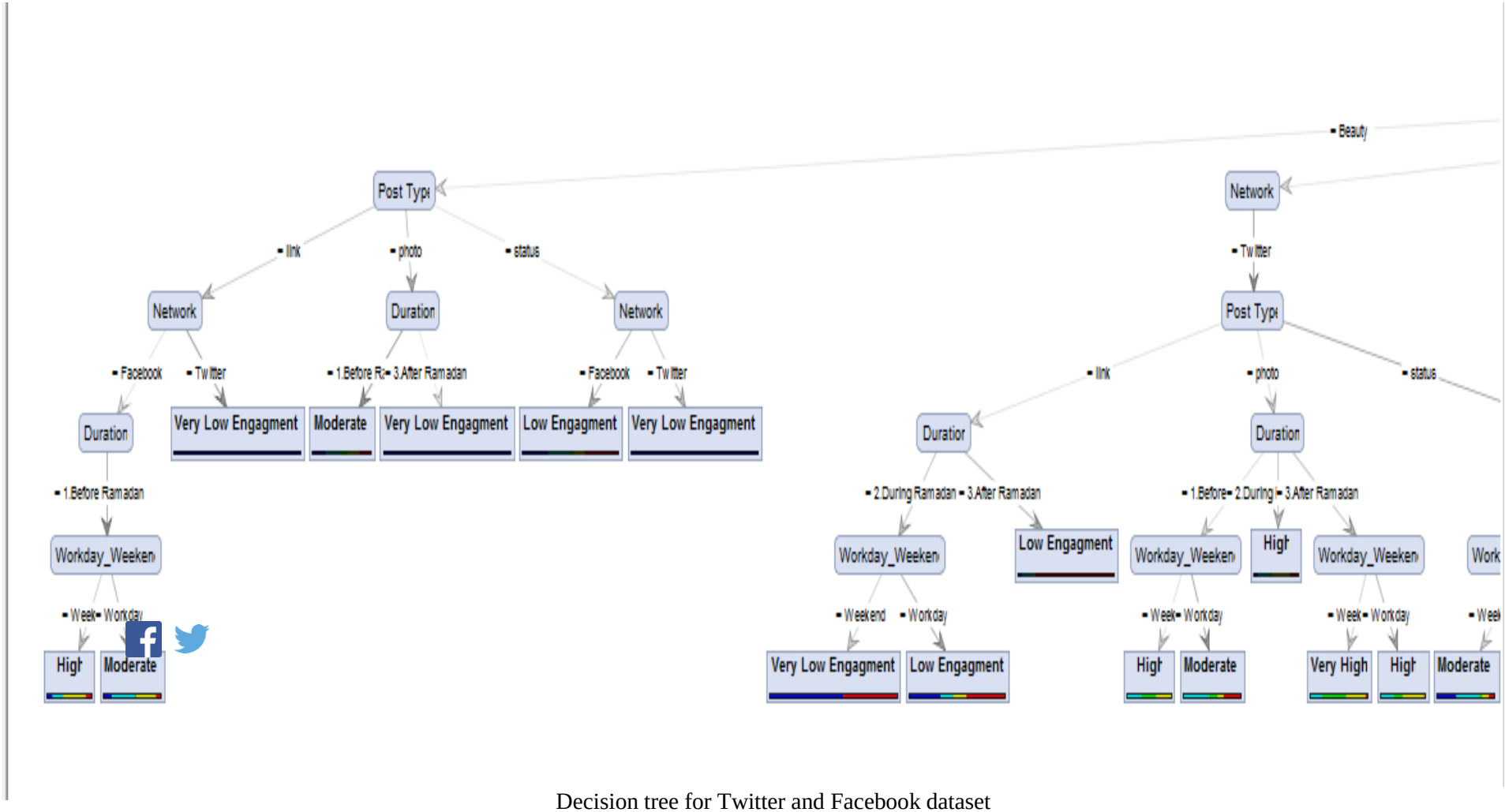
Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2017). A survey of text mining in social media: Facebook and Twitter perspectives. Adv. Sci. Technol. Eng. Syst. J, 2(1), 127-133.

Appendix I

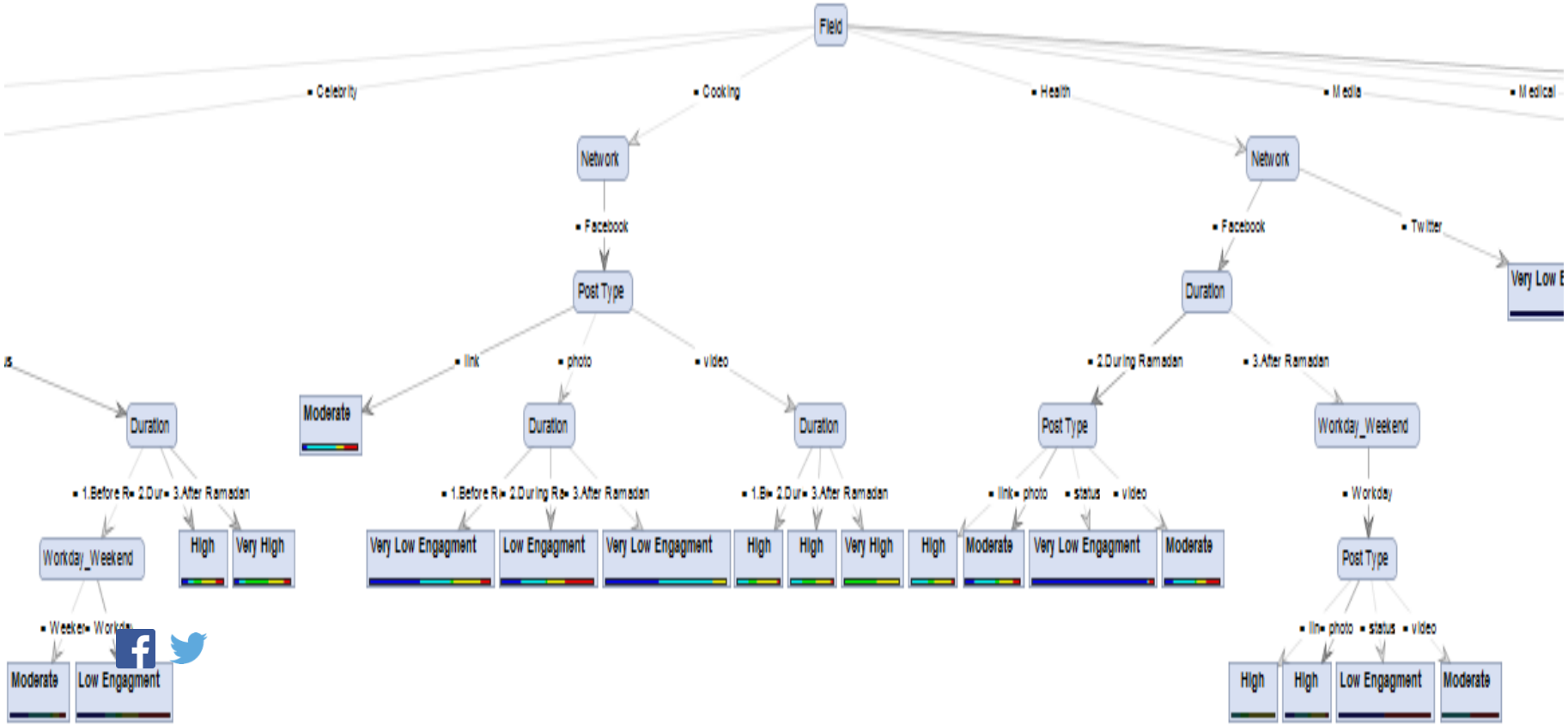


Decision tree for Instagram dataset

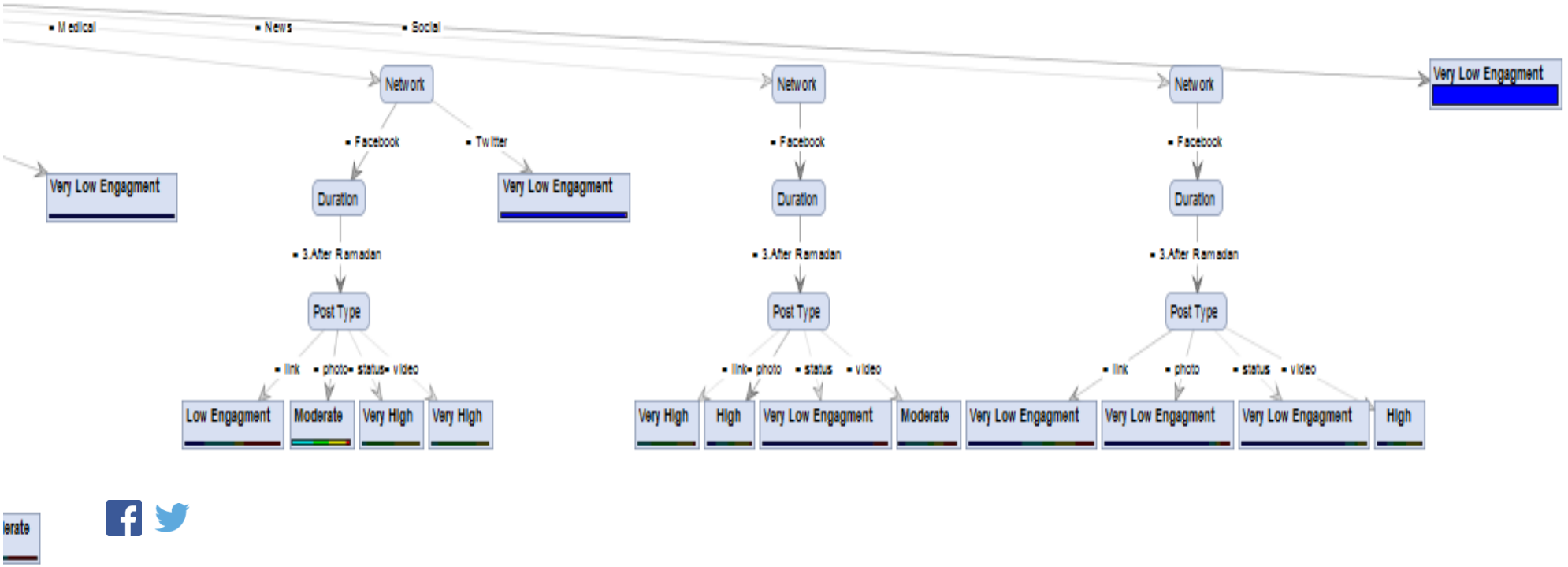
Appendix II



Decision tree for Twitter and Facebook dataset



Decision tree for Twitter and Facebook dataset



Decision tree for Twitter and Facebook dataset