Thesis for the Degree of MS in Information Security 30.0 credits

Title of the Master Thesis

Social Media, A Source For Cognizance Of Healthcare Informatics

Submitted by

Saqiba Tariq

Supervisor: Asst. Prof. Mian Muhammad Waseem Iqbal
National University of Sciences and Technology, Islamabad, Pakistan
Committee Member: Assoc. Prof. Dr. Muhammad Faisal Amjad
National University of Sciences and Technology, Islamabad, Pakistan
Committee Member: Lecturer Narmeen Shafqat
National University of Sciences and Technology, Islamabad, Pakistan

January 2020

Abstract

Social sites are a platform where colossal bulk of data is shared, individuals as well as organizations share their information on various online media platforms. Among these sites Facebook is a valuable source where tremendous set of information is erected and can be evaluated for helpful outcomes. Facebook is the most utilized public network by wide-ranging population for information distribution. Facebook can likewise be very helpful for making people cognizant of healthcare informatics. This research is comprised of two parts, in first part two approaches were used to constitute different datasets for perusal of how Facebook pages of health organizations impact in distribution of healthcare information. A manual dataset was generated by accumulating Facebook posts of different health associations working in Punjab Pakistan. This data was statistically analyzed and based upon this data predictions were made for upcoming health relevant issues. An automated dataset was amassed by collecting health related posts of medical management associations. This information was further processed and statistically examined, also a classification tree was designed for the prediction of impending health problems. After the predictive analysis end results of both datasets were compared for accuracy, manual data set gave the accuracy rate of about 70% and automated dataset's prediction was nearly 50% accurate.

In second part of the research some factors were contemplated that made gigantic impact while accessing data from Facebook. The most important factor was the changing privacy policies of Facebook, as the cyber laws are modified the access to data is also restricted which results in less accuracy of the results. When security policies of Facebook are changed the access to maximum number of attributes is restricted and data mining techniques get effected. Limited access to Facebook data badly effects the data mining as well as the analysis process which results in low accuracy of the study. Few more factors that can effect this type of study were the selection of specific health organization, bulk of irrelevant information on health related pages and deficient users' interaction to informative posts.

Contents

List of Figures v

List of Tables vii

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Problem Statement . . . 3
- 1.3 Research Objectives . . . 4
- 1.4 Scope of Study . . . 4
- 1.5 Significance of Research . . . 4
- 1.6 Author's Contribution . . . 5
- 1.7 Thesis Outline . . . 5

2 Literature Review 7
- 2.1 Introduction . . . 7
- 2.2 Analysis of literature . . . 7

3 Methodology 11
- 3.1 Introduction . . . 11
- 3.2 Paradigm of Research . . . 11
- 3.3 Research Approach . . . 12
- 3.4 Account's Identification . . . 13
- 3.5 Databases Formation . . . 13
- 3.6 Statistical Analysis . . . 13
- 3.7 Predictive Analysis . . . 14
- 3.8 Overview of Influencing Factors . . . 14

4 Data Gathering 15
- 4.1 Introduction . . . 15
- 4.2 Healthcare Organizations' Identification . . . 15
- 4.3 Manual Dataset . . . 17
- 4.4 Automated Dataset . . . 19

5 Analysis 23
- 5.1 Introduction . . . 23
- 5.2 Pre Processing of Manual Dataset . . . 23
  - 5.2.1 Organization's Type . . . 25
  - 5.2.2 Post's Type . . . 25

5.2.3 Post's Content . . . 25
5.2.4 City . . . 28
5.2.5 Cardinal Points . . . 28
5.2.6 Final View of Gathered Data . . . 29
5.3 Implementation/ Statistical Analysis . . . 30
5.3.1 Trending Post . . . 30
5.3.2 Analysis of Post Type . . . 30
5.3.3 Analysis of Post_Content . . . 31
5.4 Predictive Analysis of Manual Data Set . . . 32
5.5 Statistical Analysis of Automated Dataset . . . 33
5.5.1 Organization's Type . . . 34
5.5.2 Post's Content . . . 34
5.5.3 City . . . 35
5.5.4 Statistical Analysis . . . 36
5.6 Predictive Analysis of Automated Dataset . . . 37
5.7 Impacting Factors in Research Process . . . 39

6 Experimental Results **41**
6.1 Introduction . . . 41
6.2 Results for Manual Dataset . . . 41
6.3 Results of Automated Dataset . . . 48

7 Conclusion and Future Work **53**
7.1 Introduction . . . 53
7.2 Conclusion . . . 53
7.3 Future Work . . . 54

Bibliography **55**

Appendix **59**

List of Figures

3.1 Data collection and processing 12

4.1 Dataset of Healthcare Organizations 16
4.2 Sorted Dataset 17
4.3 Manual Dataset View 18
4.4 Diseases Categorized in Eye Health 19
4.5 Power BI View of Automated Dataset 21
4.6 Bar Chart of Power BI dataset 21
4.7 Pie Chart of Power BI Dataset 22

5.1 SPSS View of Manual Dataset 23
5.2 Imported Dataset to SPSS 24
5.3 Keywords Assigned to Organization's Type 25
5.4 Map of Province Punjab w.r.t. Districts 28
5.5 Variable View of Manual Dataset 29
5.6 Data View of SPSS Dataset 29
5.7 Final View of Manual Dataset 33
5.8 SPSS View of Automated Dataset 34
5.9 Keywords for Organization's Type in Automated Dataset 34
5.10 Variable View of Automated Dataset 36
5.11 Data View of Automated Dataset 36
5.12 Automated Dataset After Prediction 38

6.1 Graph Showing Post Type of Manual Dataset 43
6.2 Graph Showing Post Content of Manual Dataset 45
6.3 Post Content w.r.t. maximum score 45
6.4 Type of Organization w.r.t. maximum score 46
6.5 Classification Tree of Manual DataSet 47
6.6 Graph of Post Content in Automated Dataset 49
6.7 Post Content w.r.t. highest comments 50
6.8 Classification Tree of Automated Dataset 51

List of Tables

1.1 Social Media Stats Pakistan, November 2018- November 2019 2

3.1 Selection Criteria of Organizations 12

4.1 Attributes of First Dataset 16
4.2 Attributes of Manual Dataset 18

5.1 Keywords Assigned to Post's Type 25
5.2 Keywords Assigned to Post's Content 26
5.3 Keywords Assigned to Cities 28
5.4 Keywords Assigned to Cardinal Points 29
5.5 Attributes of Automated Dataset 33
5.6 Keywords for Post's Content in Automated Dataset 35
5.7 Keywords for Cities in Automated Dataset 35

6.1 Statistics of Attribute Post Type of Manual Dataset 42
6.2 Statistics of Attribute Post Content in Manual Dataset 44
6.3 Model summary of Manual Dataset 47
6.4 Statistics Automated Dataset's Post Content 48
6.5 Model Summary of Automated Dataset 50

1 Introduction

1.1 Background

Social media is a platform to divvy gigantic bulk of information to general public. It is compilation of communication channels that have network based information sharing and collaboration. If we look each word individually the social part is basically the interaction of people by sharing and consuming information while the media part is the communication tool such as internet. Social media has features such as profile pages, user accounts, personalization, newsfeed, friends, followers, information posting like button, comment section, review rating and so on. Social media is a layer of Information and Communication Technologies and till now in the history youth has been recorded as the most active community of ICT. According to a report of International Telecommunication Union the young people are the most active users of Information Communication Technologies all over the world [uni]. Striking augmentation has found in the percentage of older adults' utilization of social media. Motivational aspects are getting significant for grownups to function freely and adopt a strong sense of good health. Researchers and technologists may effectively accommodate the necessities of elder people if they comprehend this peculiar population [HB14].

Facebook is a famous American web-based social networking site based in Menlo Park, California. In 2004, its website was launched by Mark Zuckerberg and his fellow Harvard college students and roommates Eduardo Saverin, Andrew McCollum, Dustin Moskovitz, Chris Hughes. Now a days Facebook has become a part of people's lives and they create, share and exchange information using this platform. It has transformed the mindsets of people and how they use internet, now people read news in their Facebook 'newsfeed' more willingly than visiting the websites of newspapers. People read the articles that are posted on Facebook which they might not see otherwise. Facebook is accessible in almost 37 different languages with more than 1.49 Billion active clients on daily basis and approximately 2.27 Billion monthly active clients on average [Fac]. Within the social media Facebook is most extreme need of clients as 64.8% while just 21.1% are having twitter as vital source of news [Zul]. In Pakistan approximately 92.05% individuals utilize Facebook as a social media tool for sharing and exchanging information and reading news [GS].

Sr. No.	Platform	Usage
1	Facebook	92.05%
2	Twitter	3.64%
3	Pinterest	1.42%
4	YouTube	1%
5	Instagram	1.83%
6	Reddit	0.02%

TABLE 1.1: Social Media Stats Pakistan, November 2018- November 2019

Pakistan has the worst mortality rate and more than 80% of neonatal deaths can be averted with access to well-trained midwives, alongside demonstrated solutions like clean water, disinfectants, breastfeeding within the first hour, skin-to-skin contact and good nutrition. At least half a million people in the country perish annually because of medication "errors" which incorporate wrong remedy, overdose of medications, self-medication and adverse impacts of medicine [hea]. Facebook can be used as a valuable tool to spread information about healthcare issues to general public and make them aware of health informatics.

In this research the Facebook pages of healthcare organizations in Punjab are dissected. A dataset is generated manually to examine what kind of information medical services associations are posting and the present patterns of healthcare informatics. After that a statistical analysis is done to know the type of post as well as area of diseases and how much this data is shared further to apprise general public [ibm]. Beside this some attributes are analyzed in an automated way utilizing Power BI to gather information about Facebook posts of health institutions [mic]. M language is used to formulate some queries to amass data from Public pages of Facebook over a particular timeframe. Perusal of this data is done to enlighten how much this data is useful in cognizance of general public.

For big data formation social networking sites play a crucial role, big data is comprised of bulk of data used for the analysis and examine different trends. [social media and big data] Users share large amount of personal data on social media platforms without knowing the privacy concerns and big data collects this information from different social networking sites [Ban+18]. Among these platforms Facebook is an important source for analyzing different aspects of users' data, but due to privacy concerns as the cyber laws are modified the access to data is also modified. Changing cyber laws have a dynamic effect on the data gathered by social media, the patterns of data gathering changes as the policies of social networks change.

In terms of data mining and privacy when Facebook data was fetched and explored it was beheld that when the privacy was not restricted and there was access of maximum amount of data the analysis gave better results with higher accuracy rate. But when data access was restricted because of privacy concerns of users, the data mining techniques had to change for accessing the data from

Facebook. Still a limited number of attributes were accessible which resulted in reduction of accuracy level after the analysis of that data.

1.2 Problem Statement

Pakistan has worst new born mortality rate, a large part of the general population bite the dust because of maladies. Inadequate awareness level related to health protection issues are the whys and wherefores behind such situation. In Pakistan, numerous efforts are made to rectify the healthcare framework but at the same time it isn't promising. The fundamental medical problems include lack of awareness, no balanced diet, water contamination and cleanliness issues prompting communicable and non-communicable ailments. Because of the deficiency of attendants and specialists along with no or less awareness in rural regions, consistently a huge number of kids bite the dust amid birth or they experience the ill effects of a couple of infections. The expanding number of heart patients for the most part including youngsters experiencing congenital heart disease is disturbing. The death rate of ladies is additionally high since midwives treat them who have no professional knowledge. There is a need to do genuine actions to improve the medicinal services framework in our country. The solution is health awareness and quick access to medicinal care. The hygienic structure and drinking water additionally should be observed to evade sickness in individuals living in rural regions.

Greater part of the general population in Pakistan utilize Facebook as a social media tool, it can likewise be used to aware individuals about medical problems. As Punjab is the heavily populated province of Pakistan with an evaluated citizens of 110,012,442 approximately as of 2017, proffering awareness to such a huge number of residents will have great impact. This research contemplates the need of using social media for making awareness to overall population in regard to medical problems. The study distinguishes the Facebook pages of medicinal services associations working in Punjab Pakistan and accumulates datasets for examination that how human services associations are using social media for the better health of overall population. Along with that what kind of posts are shared by medicinal associations, what sort of ailments are addressed and how much response is given by people in general on such posts. Moreover, this study proposes a machine learning technique to anticipate forthcoming medical issues. By dissecting this dataset and its outcomes the author believes that it will be advantageous for healthcare organizations to utilize social media for cognizance of public and to diminish medical problems in Punjab Pakistan. Facebook data of health organizations may be used for the analysis of upcoming health problems, but privacy policy of Facebook effects the data collection and analysis process. The varying privacy policies of Facebook has a greater impact on data collection process as well as the results derived from the data.

1.3 Research Objectives

This research suggests healthcare organizations to avail social media as a tool for awareness of general public about health issues. The objectives that got the nod to attain in this study are:

- Identifying public pages of healthcare organizations within Province Punjab running actively on Facebook.
- Gathering data sets for analysis and identifying the challenges in Data Collection and Analysis being faced due to new (more stringent) privacy policies of Facebook.
- Proposing a machine learning technique for prediction of an upcoming health care issue.

1.4 Scope of Study

This research is applicable to healthcare associations functioning in Punjab Pakistan along with ministry of Health. The study also incorporates different data ware houses of healthcare, NGOs working in health sector and medical sciences to avail social media platforms for enlightenment of public regarding health issues. By exploring the informational collection in this study current trends of medical problems in Punjab can be noticed and it might likewise provide insights about the upcoming health problems in the region.

1.5 Significance of Research

There is a dire need to address the dilemma of cognizance regarding medical management concerns. It is essential to identify the patterns of diseases and also making overall population mindful of them. Ministry of health has to develop auxiliary policies for betterment of public. People need to know about the ailments and drugs used for their treatment. Besides people should be familiar with the epidemics, their symptoms, precautionary measures as well as their medications. No research has been made to propose health organizations that Facebook can be used as a mechanism to alert general public about health issues in Pakistan neither there are datasets of the social media accounts of healthcare organizations of Punjab.

1.6 Author's Contribution

The conducted research recognizes the dynamic healthcare organizations working in Punjab Pakistan which incorporate government organizations, for profit organizations, not for profit organizations and non-governmental organizations. Datasets are generated for analysis, the content posted by these associations on their Facebook pages is explored and assessed to discern the prevailing patterns of diseases. The nature of posts health organizations are disseminating on Facebook accounts are examined along with the content of the posts. Likewise how much general public responds to such posts is observed by figuring out the number of likes, comments and shares. Two datasets are accumulated for perusal, one is gathered manually and the other one automatically. Both the datasets are explored from different perspectives and a machine learning technique is projected for the awareness of imminent health issue. By considering the previous data forthcoming maladies are prophesied so that general public get consciousness about that and respond accordingly. Moreover those factors were analyzed which had an immense effect on the whole research process in terms of varying cyber laws of Facebook.

1.7 Thesis Outline

The first chapter of the thesis narrates the overview along with the background of the topic, problem statement, scope of the study, significance of research and author's contribution. Chapter 2 encompasses literature review which features the performance of social media in healthcare informatics; what kind of information is posted by health relevant associations, how social platforms are assisting over-all community to deal with the health concerns and how social networks and search engine queries cooperates in discovering epidemics. Chapter 3 explains the approach used in the study, it provides the overview of modus operandi pursued in the further research procedure and overall view of the entire research progression. Chapter 4 elaborates that how data is amassed for the exploration, what attributes were chosen to design datasets and how datasets are generated from Facebook pages of medical management systems. How manual dataset is produced by means of precise ascribes and how automated dataset is created. Chapter 5 incorporates of the examination of both the datasets, statistical analysis of manual dataset is deduced and automated dataset is likewise explored for discerning the approaching health issues in the region. A predictive analysis for both datasets is carried out and the impact of privacy changes in Facebook policy influencing the analysis phase is discussed in detail. In chapter 6 experimental results of both datasets are presented, using the above modus operandi and analysis concluded results are delineated. Chapter 7 gives an overview of the conclusion of the study along with the explanation about the future work which can be done on the basis of this study.

2 Literature Review

2.1 Introduction

This chapter gives an overview of the research work executed previously in context of Facebook usage and how its data can be utilized for deriving useful information in terms of healthcare informatics. It also explains that how Facebook data is used for pragmatic predictive analysis of health relevant data. How search engine queries and social media's data can be utilized for predicting epidemics in certain regions is also elucidated in this chapter. Furthermore how varying cyber security laws and GDPR impact the data collection process of SNS is also described in this chapter.

2.2 Analysis of literature

In a paper by Clayton Hutto [HB14] it is stated that striking augmentation has found in the percentage of older adults' usage of web based life. How much this use gives social fulfillment to the general population beyond 50 years old is dissected. By investigating the correspondence patters between the users and non-users of social media it is concluded that directed communication is explicitly related with sentiments of social fulfillment among this community. Between older adults the features of users and non-users of social media are inspected on the basis of age, gender, race, marital status, education and income. Besides reasons behind not utilizing social networking sites are reviewed and how the users access these platforms alongside sort of data they share via web-based networking media, what kind of content they post on their accounts and along with Facebook what other SNSs are being used for the most part is observed. The consequence of the fast increment in aged people is of greater importance equally in science and technology. Along with mental and physical disabilities individuals are living longer. Inspirational variables for remaining associated is fundamental to help aged persons to work freely and grasp a sound feeling of prosperity. Comprehending this unique community will support specialists and researchers alike better oblige the necessities of the older.

Utilizing a numerous contextual analysis approach, usage of social media by patients, their loved ones, guardians, and associations regarding prostate and breast cancer is examined to perceive how they communicate to generate new insights [Bag+14]. Customized medication and patient's empowerment patterns for medical associations reveal the significance of information co-creation and to the cutting edge.

The utilization of social networking websites to dispense facts and figures associated to health is expanding briskly. To share prevailing information among health patrons Twitter is generally used social networking. Twitter accounts of 114 Australian health associated organizations were manually examined and categorized over a specific period of time. The work gave insights about the area as well as type of health related microblog postings as well as health informatics dissemination over social media. Alertness tweets and community health guidance were the common sort of health information that was dispersed on Twitter by the healthcare organizations of Australia. On the basis of some attributes like geographical location and type of organization different Twitter accounts of health relevant associations in working in Australia were examined. These organizations contained governmental and nongovernmental associations along with non- profit organizations. Their tweets were thoroughly studied to elaborate the context of the tweet and the area of disease mentioned in that tweet. Additionally an assessment was done to divulge the category of tweets by the Government bodies, NGOs and Non-Profit establishments [DS13].

For the researchers data gathering is a significant stage and online social networking platforms are the imperative hotspot for this. Distinctive modus operandi are utilized to amass information from various platforms of social media, relating APIs aids in getting facts and figures from these SNSs [AD14].

Social media and search engines have a prominent influence on healthcare professionals, patients and over-all public. The network where logs of health relevant questions are found is search engine, while social media is one more platform where individuals share their medical experiences, indication of diseases and treatment for those ailments [NML17]. By means of question logs of Google records and Twitter assists to collect data about healthcare informatics. Queries of search engines and tweets in conjunction with retweets are very useful in forecasting influenza outbreak in a particular geographical region. This data gives the insights about the causes, symptoms and treatments of diseases and specific medicines to cure those diseases effectively along with their side effects.

According to M. Ikram Ullah Lali et al. [Lal+16] search engine queries were utilized for looking through the flu related data and social platforms contained helpful links of data sources that were fruitful. This was intended for mining helpful data about reactions, meds and to track geological area of epidemic influenced zone.

Efficiency of predictive research with online networking information of a private medicinal services establishment utilizing classification and clustering algorithms is inspected. Together with that impact of framework explicit list of capabilities and client created highlights of SNS articles' to distinguish appropriate data mining techniques that are reasonable for recognizing authentic patterns in dataset is investigated. Bayesian Algorithm was best suited for classification while for clustering K-Means gave the better results. Data mining techniques are when applied on social media data of health relevant organization it gives the future insights of the organization in which likes and shares are the most important attributes [Mel+17].

A study by Nadiya Straton et al. [SMV17] demonstrates Facebook post's popularity exploration the study suggests healthcare associations to improve data scattering via web-based networking media by lessening mess and disarranged data. An approach is proposed to foresee how users interact with posts using attributes like: Type of Post, Season, Category of Timeline, and Country and so on. The research concluded that ANN forecasting showed more accurate results while deep neural networks indicated less accuracy in terms of healthcare related data of Facebook because of more time consumption in processing speed. According to social media experts the content of the post and the keywords matters a lot to engage people.

Approximately 300 Million people are suffering from depression around the world which is an immensely significant mental disorder. In a paper by Akkapon Wongkoblap a predictive model was delineated to identify people with depression on social media platform Facebook [9]. Different patterns were examined in the datasets of myPersonality between depression and fulfilment of life and a negative correlation was found. The predictive model reviewed the level of life fulfillment in people and then used the results for prediction of depression in people. A negative correlation was observed between the satisfaction of life and depression [WVC18].

Digital Security approaches have a colossal effect on data mining, data fusion and functionality of data driven applications. [cybersecurity] Varying cyber security policies modify the effectiveness and functionality of the data, such as changing in the law of freedom of speech can change the availability of data and publicly distributed information. Moreover the platforms that collect data from web are badly affected by the changing policies of cyber laws [Tuc+17].

Social networks are influenced by the General Data Protection Regulations, GDPR has an immense impact on social networking sites. [GDPR] When GDPR rules are implemented the way of organizations' treating information insurance changes. Social networking sites have a huge access of user's personal information and when these are made according to GDPR the way of the organization distributing and accessing user's information modifies.

3 Methodology

3.1 Introduction

This section provides an outline of the modus operandi adopted for this research. As in the previous section background studies regarding role of social media in healthcare informatics was discussed in detail, this chapter confers the kind of approach followed to conduct this research. Moreover, the insights of methodology used for collection of data and criteria for data collection is included in this section. An overview of datasets and analysis techniques used on those datasets is also explained in this chapter. How different datasets were analyzed in different ways and what were the outcomes for those specific datasets is illustrated. Likewise how prediction algorithm functioned and provided the enlightment of upcoming healthcare issues in Punjab Pakistan.

3.2 Paradigm of Research

An amalgamated methodology tactic is used to devise a solution for the research questions of this study. Exploratory research is used in combination with predictive research to explore under-researched problems of healthcare issues in Province Punjab Pakistan. The patterns and trends of diseases spreading in Punjab have been analyzed by exploring the social media accounts of health-care organizations. On the basis of existing data from the Facebook pages of health organizations, predictive analysis was carried out to know the future outcomes of upcoming healthcare issues in the region. The criteria for selecting a Facebook page is as under.

Sr. No.	Selection Criteria	Description
1	Nationality	Health organizations working in Pakistan
2	Administrative Division	In Pakistan only the organizations operating in different areas of Province Punjab
3	Substantial Worth	Facebook pages representing Government, Non-Government, For- Profit and Not-for-Profit Organizations
4	Post Content	Pages having posts relevant to healthcare data
5	Status	Pages that are currently active or at least active after February 2018

TABLE 3.1: Selection Criteria of Organizations

3.3 Research Approach

A mixed approach including quantitative and qualitative way is used to gather the data to overcome the constraints of a particular method [res]. By means of qualitative approach the type and content of a Facebook post of healthcare association was examined. In this way the type of organization along with name and area in which it is operating was fetched and stored. While quantitative data such as likes, comments and shares of a specific post gave insights about the current patterns and trending diseases in the region. More than two datasets are generated, which includes the dataset about the organizations and then datasets about the posts of those organizations.

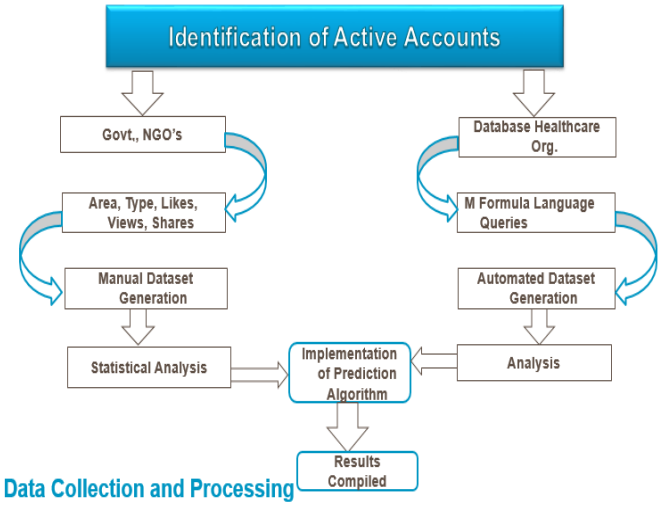


FIGURE 3.1: Data collection and processing

3.4 Account's Identification

Healthcare organizations working in Punjab were identified using search engine Google manually. It was examined that which organization are working in different districts of Punjab, then their official websites were explored [ngo]. Official websites of these organizations contained the links of social media accounts of the organizations. Many organizations have no official social media account associated with the website. Using links from the official websites the Facebook pages of the medical associations were analyzed. A database was generated having the information about the healthcare organizations working in Punjab along with their respective cities and services provided by them. Active accounts were identified manually to generate another dataset for keeping the record of the posts provided by the associations on their pages.

3.5 Databases Formation

First database was formed having the information about the healthcare organizations working in Punjab and their associated information [pun]. Besides another dataset was generated manually in which a Facebook page of one of the association was observed. It was studied that in a specific time period what type of health information was posted by an organization, in this way many organizations were examined. Datasets were generated manually because at that time Facebook was facing some issues and due to its strict privacy policy data mining was not permitted using R language. Furthermore an automated dataset was also generated to analyze the posts of Facebook pages of health associations using M formula language queries, it was generated with the help of Power BI. Again because of strict and updated privacy policies of Facebook only a few attributes were able to be fetched and studied.

3.6 Statistical Analysis

A statistical analysis was performed on the second dataset of health administrations in SPSS. Trending posts and current patterns of diseases were explored and recorded. The frequencies and percentages of maladies found were calculated along with the type of posts. A detailed quantitative analysis was carried out to inspect that what type of content was posted by the health sector of Punjab. Different graphs were produced to get perceptions of the data and the ailments found in that.

Different graphs were generated form the automated dataset to know the trending posts based on the number of comments on each post of a specific organization over a particular period of time. The posts from January 2018 to January 2019 were mined and investigated in Power BI to get

insights about the highest commented post. The top five posts from each organization were than statistically examined in SPSS to get most posted ailments.

3.7 Predictive Analysis

Predictive analysis of SPSS's dataset was done to get an overview of the upcoming healthcare issues in the region. A decision tree was made to predict about the future outcomes of maladies depending upon the data gathered previously. It also predicted the diseases that may arise in future by discerning the previous data assembled from specific organizations of explicit areas. Based upon the collected datasets a predictive value was calculated to have a clear idea about the upcoming health issues.

3.8 Overview of Influencing Factors

Different factors impacting the research process were observed and analyzed. Many factors which play a vital role in the study and effect the results as well as the accuracy rate of the final results were taken into account. It was espied that how the general public interacts with health associations and play a part in concluding fruitful results relevant to healthcare informatics. Furthermore how privacy concerns of Facebook effects the data collection process was analyzed [Fb].

4 Data Gathering

4.1 Introduction

Chapter three has outlined the modus operandi followed for this study and given a bird's eye view of the whole the data collection process in this study. In this chapter the detailed process of data gathering and datasets assembling is discussed in detail. Section 4.2 describes the inclusion criteria of the healthcare organizations gathered and section 4.3 explains how the second manual dataset was derived from the first dataset. Section 4.4 gives an overview of the mining process for automated dataset and how it was processed in graphical form.

4.2 Healthcare Organizations' Identification

By means of google search engine district wise healthcare organizations were explored, websites of different health organizations were fetched. As there were a lot of organizations working in Punjab but all of them were not active on social platforms, from the official websites the social media accounts were noted down, Facebook pages of these associations were scrutinized [ngo]. Different attributes of the Facebook page were put on record like organization's city, nature of the organization, services provided and the contact details of Facebook along with other some details. The list of attributes included in this dataset are in the given table.

Sr. No.	Attribute	Description
1	Organization	Contains the name of the organization mentioned on Facebook Page.
2	City	Name of the city in which the organization is currently working.
3	User Name	The user name mentioned under the Page's name on the Facebook public page.
4	Status	The status of the organization either it is active or inactive.
5	Nature	Contains the information about the nature of the organization either it's a governmental or non-governmental organization.
6	Review	The reviews given by the general public about the organization.
7	Services	The basic services provided by the organization mentioned in the About section of Facebook Page.
8	Contact Details	Contains the Facebook Messenger link of the organization.
9	Official Website	Contains Official Website of the Health Organization.

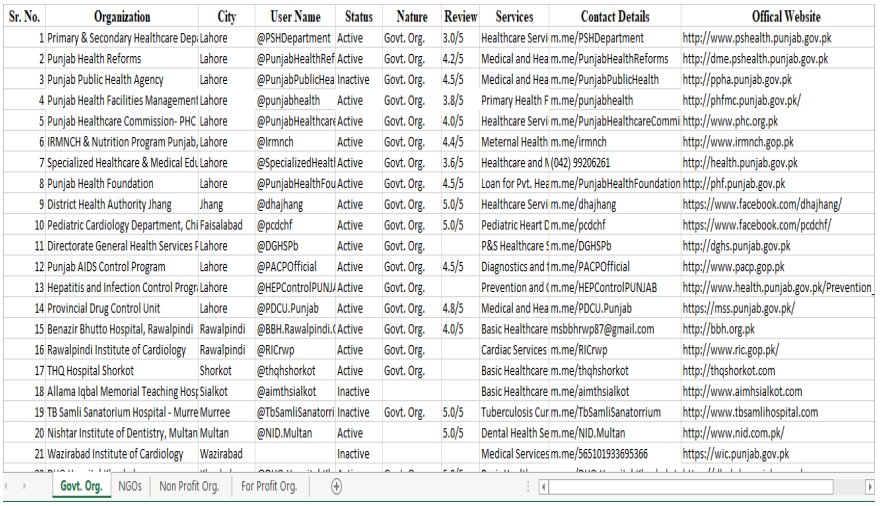
TABLE 4.1: Attributes of First Dataset

Below is a screenshot of the initial and first dataset having the information about healthcare organizations.

A	B	C	D	E	F	G	H	I	J	K	L	M
Sr. No.	Organization	City	User Name	Status	Nature	Review	Services	Contact Details	ical Website			
1	Primary & Secondary Healthcare Department	Lahore	@PSHDepartment	Active	Govt.	3.0/5	Healthcare Serv	m.me/PSHDepar	http://www.pshealth.punjab.gov.pk			
2	Okara Patient Welfare Association	Okara	@pwadhqokara	Active	NGO	4.2/5	Medical and He	m.me/pwadhqo	http://www.opwacare.com/			
3	UNICEF Pakistan	H.O Islamabad	@unicefpakistan	Active	Non-profit	4.5/5	Health, Polio, N	m.me/unicefpak	https://www.unicef.org/pakistan/			
8	Sundas Foundation	Gujranwala	@sundasfoundationgujran	Active	NGO		Blood and Bloo	m.me/sundasfo	http://www.sundasfoundation.org/			
4	Musa Memorial Hospital	Bhakkar	@Musahospital	Active	Non-profit	4.8/5.0	Medical Service	m.me/Musahos	http://mmhbhakkar.wixsite.com/musaho			
9	District Health Authority Jhang	Jhang	@dhajhang	Active	Govt.	5.0/5	Healthcare Serv	m.me/dhajhang	https://www.facebook.com/dhajhang/			
5	Mukhtar A. Sheikh Hospital	Multan	@MASMWH	Active	Non-profit	5.0/5	Basic Healthcar	m.me/MASMWH	http://mashospital.org			
6	Capital Health Hospital	Chakwal	@Capitalhealth	Active	For Profit	4.8/5	Basic Healthcar	m.me/Capitalhe	http://www.capitalhealth.com.pk			
26	Spring Clinic - Institute of Psychiatry	Multan	@springclinicmultan	Active	Govt.	5.0/5	Mental Health S	m.me/springclin	http://fatima-group.com/mast/			
7	Shahid Hospital	Sargodha	@shahidhospitalsargodha	Avctive	For Profit	4.3/5	Child and Wom	m.me/shahidho	http://shahidhospital.pk/			
3	LRBT Free Eye Care Trust- Pakistan	Chiniot	@Lrbt.Pakistan	Active	NGO	4.7/5	Eye Care Servic	m.me/Lrbt.Pakis	http://www.lrbt.org.pk			
8	Winner Health Care	Gujranwala	@winnerhealthcare	Active	For Profit	4.4/5	Dialysis, Dental	m.me/winnerhe	http://www.winner.com.pk			

FIGURE 4.1: Dataset of Healthcare Organizations

Then this database was sorted out and Government, NGOs, Non-Profit and For-Profit-Organizations were put in different worksheets of an excel workbook. Data of 72 health organizations working in Punjab was collected and organized in one database. This database was produced to proceed for the further research process and for creating other datasets.



Sr. No.	Organization	City	User Name	Status	Nature	Review	Services	Contact Details	Offical Website
1	Primary & Secondary Healthcare Dep	Lahore	@PSHDepartment	Active	Govt. Org.	3.0/5	Healthcare Servi	m.me/PSHDepartment	http://www.pshealth.punjab.gov.pk
2	Punjab Health Reforms	Lahore	@PunjabHealthRef	Active	Govt. Org.	4.2/5	Medical and Hea	m.me/PunjabHealthReforms	http://dme.pshealth.punjab.gov.pk
3	Punjab Public Health Agency	Lahore	@PunjabPublicHea	Inactive	Govt. Org.	4.5/5	Medical and Hea	m.me/PunjabPublicHealth	http://ppha.punjab.gov.pk
4	Punjab Health Facilities Management	Lahore	@punjabhealth	Active	Govt. Org.	3.8/5	Primary Health F	m.me/punjabhealth	http://phfmc.punjab.gov.pk/
5	Punjab Healthcare Commission- PHC	Lahore	@PunjabHealthcare	Active	Govt. Org.	4.0/5	Healthcare Servi	m.me/PunjabHealthcareCommi	http://www.phc.org.pk
6	IRMNCH & Nutrition Program Punjab,	Lahore	@Irmnch	Active	Govt. Org.	4.4/5	Meternal Health	m.me/irmnch	http://www.irmnch.gop.pk
7	Specialized Healthcare & Medical Edu	Lahore	@SpecializedHealt	Active	Govt. Org.	3.6/5	Healthcare and N	(042) 99206261	http://health.punjab.gov.pk
8	Punjab Health Foundation	Lahore	@PunjabHealthFou	Active	Govt. Org.	4.5/5	Loan for Pvt. Hea	m.me/PunjabHealthFoundation	http://phf.punjab.gov.pk
9	District Health Authority Jhang	Jhang	@dhajhang	Active	Govt. Org.	5.0/5	Healthcare Servi	m.me/dhajhang	https://www.facebook.com/dhajhang/
10	Pediatric Cardiology Department, Chi	Faisalabad	@pcdchf	Active	Govt. Org.	5.0/5	Pediatric Heart D	m.me/pcdchf	https://www.facebook.com/pcdchf/
11	Directorate General Health Services F	Lahore	@DGHSPb	Active	Govt. Org.		P&S Healthcare S	m.me/DGHSPb	http://dghs.punjab.gov.pk
12	Punjab AIDS Control Program	Lahore	@PACPOfficial	Active	Govt. Org.	4.5/5	Diagnostics and t	m.me/PACPOfficial	http://www.pacp.gop.pk
13	Hepatitis and Infection Control Progr	Lahore	@HEPControlPUNJ	Active	Govt. Org.		Prevention and (	m.me/HEPControlPUNJAB	http://www.health.punjab.gov.pk/Prevention_
14	Provincial Drug Control Unit	Lahore	@PDCU.Punjab	Active	Govt. Org.	4.8/5	Medical and Hea	m.me/PDCU.Punjab	https://mss.punjab.gov.pk/
15	Benazir Bhutto Hospital, Rawalpindi	Rawalpindi	@BBH.Rawalpindi.(	Active	Govt. Org.	4.0/5	Basic Healthcare	msbbhrwp87@gmail.com	http://bbh.org.pk
16	Rawalpindi Institute of Cardiology	Rawalpindi	@RICrwp	Active	Govt. Org.		Cardiac Services	m.me/RICrwp	http://www.ric.gop.pk/
17	THQ Hospital Shorkot	Shorkot	@thqhshorkot	Active	Govt. Org.		Basic Healthcare	m.me/thqhshorkot	http://thqshorkot.com
18	Allama Iqbal Memorial Teaching Hos	Sialkot	@aimthsialkot	Inactive			Basic Healthcare	m.me/aimthsialkot	http://www.aimhsialkot.com
19	TB Samli Sanatorium Hospital - Murre	Murree	@TbSamliSanatorri	Inactive	Govt. Org.	5.0/5	Tuberculosis Cur	m.me/TbSamliSanatorrium	http://www.tbsamlihospital.com
20	Nishtar Institute of Dentistry, Multan	Multan	@NID.Multan	Active		5.0/5	Dental Health Se	m.me/NID.Multan	http://www.nid.com.pk/
21	Wazirabad Institute of Cardiology	Waziabad		Inactive			Medical Services	m.me/565101933695366	https://wic.punjab.gov.pk

Govt. Org. | NGOs | Non Profit Org. | For Profit Org.

FIGURE 4.2: Sorted Dataset

4.3 Manual Dataset

Based upon the dataset generated another dataset was designed manually comprising the posts of health organizations. A healthcare organization was chosen manually and its Facebook page was evaluated to get the posts from that page. Fifty posts were assembled and classified in excel sheets from February 2018 onwards. A post of a specific health protection association was examined manually to investigate that what type of post is this and the content of the post is addressing which sort of disease.

The randomly selected organizations for analysis were:

- Alshifa Eye Hospital
- Bilal Hospital
- DHA Jhang
- Provisional Drug Control Unit
- KKT Orthopedic Spine Center
- Layton Rahmatullah Benevolent Trust
- Punjab AIDS Control Program
- Punjab Health Facilities Management Company
- Primary and Secondary Health Department
- Sundas Foundation

- Zhong-Ba Hospital

Fifty posts from each of these organizations were explored and put on record in excel files for further analysis of the data.

The attributes included are mentioned in the table below.

Sr. No.	Attribute	Description
1	Views	Contains the number of views of a video post.
2	Date	Contains the date of the post.
3	Post's Type	Contains the type of the post e.g. awareness, feedback etc.
4	Likes	Number of likes on the post.
5	Comments	Number of comments on the post.
6	Shares	Number of times a post has been shared.
7	Post's Link	Facebook link of the post.
8	Post's Content	Contains the content of the post i-e the area of disease targeted in that post.

TABLE 4.2: Attributes of Manual Dataset

After gathering these attributes the excel dataset looked like the screenshot given below.

		Zhong-Ba Hospital					
Views	Date	Post's Type	Post's Content	Likes	Comments	Shares	Post's Link
	23-Feb-18	Advertisement	Muscle Pain	3	0	0	https://www.facebook.com/zhongbahospital/posts/954
	23-Feb-18	Awareness	Sacrum Pain	5	0	2	https://www.facebook.com/zhongbahospital/posts/954
	24-Feb-18	Awareness	Thyroid Disorder	5	0	0	https://www.facebook.com/zhongbahospital/posts/954
	24-Feb-18	Advertisement	Andrology	2	0	0	https://www.facebook.com/zhongbahospital/posts/954
	24-Feb-18	Awareness	Anorectal Disorders	2	0	0	https://www.facebook.com/zhongbahospital/posts/954
	25-Feb-18	Awareness	Back Pain	1	0	0	https://www.facebook.com/zhongbahospital/posts/953
	25-Feb-18	Awareness	Andrology	2	2	0	https://www.facebook.com/zhongbahospital/posts/953
	26-Feb-18	Awareness	Gynae and Obs	1	0	0	https://www.facebook.com/zhongbahospital/posts/953
	26-Feb-18	Awareness	Gynae and Obs	2	0	0	https://www.facebook.com/zhongbahospital/posts/950
	27-Feb-18	Advertisement	Anorectal Disorders	4	0	0	https://www.facebook.com/zhongbahospital/posts/950
	27-Feb-18	Advertisement	STD	1	0	0	https://www.facebook.com/zhongbahospital/posts/950
	27-Feb-18	Awareness	Migraine	1	0	0	https://www.facebook.com/zhongbahospital/posts/950
	27-Feb-18	Awareness	Migraine	5	0	0	https://www.facebook.com/zhongbahospital/posts/950
	28-Feb-18	Advertisement	Gynae and Obs	5	0	0	https://www.facebook.com/zhongbahospital/posts/957
	28-Feb-18	Advertisement	Gynae and Obs	2	0	0	https://www.facebook.com/zhongbahospital/posts/957
	28-Feb-18	Advertisement	Anorectal Disorders	1	0	0	https://www.facebook.com/zhongbahospital/posts/957
	28-Feb-18	Awareness	Frozen Shoulder	3	1	0	https://www.facebook.com/zhongbahospital/posts/957
	1-Mar-18	Advertisement	Sciatica	4	0	0	https://www.facebook.com/zhongbahospital/posts/957

FIGURE 4.3: Manual Dataset View

Different diseases were categorized in one keyword for post content. Eleven organizations were selected randomly from different cities and different backgrounds and their posts were collected to get an idea about the trending diseases of that time period. The categorization of different diseases under specific keywords are as under:

Eye Health				
Blindness	Lost Vision	Cornea Problems	Oculoplastics	Congenital Cataract
Eye Grafting	Refractive Error	Tumor in Eyes	Glaucoma	Visual Impairment

FIGURE 4.4: Diseases Categorized in Eye Health

Other diseases relevant to eyes were also put under the category of eye health. Besides other diseases were also categorized using specific keywords.

4.4 Automated Dataset

For automated dataset Facebook posts of thirty different health organizations working in different cities of Punjab were recorded automatically. Each organization was working in different health area and different city, the organizations selected randomly for analysis are as follows:

- Al-Khidmat Foundation
- Al-Shifa Trust Eye Hospital
- Benazir Bhutto Hospital
- Capital Health Hospital
- Directorate General Health Services Punjab
- District Health Authority Jhang
- Ghurki Trust Teaching Hospital
- Hepatitis and Infection Control Program, Punjab
- The Indus Hospital
- IRMNCH and Nutrition Program, Punjab
- Pakistan Kidney and Liver Institute and Research Center
- LRBT Free Eye Care Trust, Pakistan
- Muslim Medical Mission
- Punjab Aids Control Program
- Pediatric Cardiology Department, Children Hospital Faisalabad
- Provisional Drug Control Unit

- Punjab Healthcare Commission PHC
- Punjab Health Foundation
- Punjab Health Facilities Management Company
- Punjab Public Health Agency
- Primary and Secondary Healthcare Department
- Punjab Health Reforms
- Okara Patient Welfare Association
- Rawalpindi Institute of Cardiology
- Specialized Healthcare and Medical Education Department
- Spring Clinic – Institute of Psychiatry
- Sundas Foundation
- UNICEF
- WHO
- Zhong-Ba Hospital

Graph API of Facebook was utilized to get the data from public pages of Facebook, for this purpose Power BI was utilized to design a dataset automatically [fac]. Facebook pages of healthcare organizations were opened and then the name of the organizations were connected to Power BI and loaded in that. To connect with the Facebook page and mine the data for that the snippet used is:

let
Source = Facebook.Graph("https://graph.facebook.com/v2.8/punjabhealth/posts

Using this type of snippets the data from different pages of healthcare organizations of Punjab was mined and recorded. After getting the data it was refined and pre-processed for extraction of he required attributes from this bulk of data. Imported data was then transformed to get the required information for further analysis, data and time was separated by delimiter and then connections were established to get the number of comments for each post using the snippet:

 # "Expanded object_link" = Table.ExpandRecordColumn(#"Renamed Columns", "object_link", "connections", "object_link.connections")

After refining following attributes are gathered and sorted further.

message	created_date	created_time	id	Number f Comments	story
This cute baby girl Naila was having tumor in her eye, Was	*Wednesday, May 23, 2018*	17:00:47+0000	393530784080887_1362969760470313	*99*	
Of all the blind children in Pakistan, 80% can't see becaus	*Friday, July 1, 2016*	05:42:48+0000	393530784080887_816055785161716	*92*	
Al-Shifa Trust Eye Hospital, Rawalpindi, Pakistan is struggli	*Friday, May 11, 2018*	14:41:15+0000	393530784080887_1353112128122743	*57*	
A Beacon of Light, Al-Shifa Trust Eye Hospital This Ramada	*Tuesday, June 6, 2017*	08:33:30+0000	393530784080887_1076309475803011	*40*	
This Ramadan, Help Al-Shifa Trust Eye Hospital in treatmer	*Thursday, June 22, 2017*	13:07:49+0000	393530784080887_1089827654451193	*38*	
Please Pray for Sami Ullah, Serious injury in both Eyes (Cor	*Wednesday, July 4, 2018*	08:56:21+0000	393530784080887_1409951749105447	*38*	
President of Al-Shifa Trust Eye Hospital Gen (R) Hamid Jave	*Monday, May 30, 2016*	08:49:17+0000	393530784080887_800284990072129	*38*	
@alshifaeyetrustpak Mission, to eradicate and fight agains	*Wednesday, June 13, 2018*	14:58:44+0000	393530784080887_1381385378628751	*35*	
First Ever Eye Bank in Pakistan, Established in Al-Shifa Trust	*Thursday, August 2, 2018*	14:51:05+0000	393530784080887_1451049748328980	*34*	
"Har Aankh Roshan" with Dr. Abdullah Naeem Syed, of Al-S	*Monday, July 10, 2017*	07:49:45+0000	393530784080887_1104720362961922	*32*	Al-Shifa Trust Eye Hospital was live.
"Al-Shifa was the only beacon of light in that disappointing	*Friday, May 18, 2018*	13:20:00+0000	393530784080887_1358428197591136	*28*	
#GreenPakistan #Plant4Pakistan Plantation campaign in Al	*Wednesday, September 5, 2018*	04:38:05+0000	393530784080887_584719251926382	*27*	Al-Shifa Trust Eye Hospital was live.
Arshad Mehmood sharing his views about Al-Shifa Trust Ey	*Thursday, May 31, 2018*	10:14:05+0000	393530784080887_1369452236488732	*23*	
Jahandad Khan with his son Salman Khan (From Afghanista	*Wednesday, June 6, 2018*	12:30:00+0000	393530784080887_1374427399324549	*23*	
Fateh Muhammad is 78 Years Old having Diabetic Retinop	*Tuesday, May 29, 2018*	05:23:55+0000	393530784080887_1367612803339342	*22*	
Al-Shifa Trust Eye Hospital Team of School Screening visite	*Saturday, June 2, 2018*	15:15:27+0000	393530784080887_1371375229629766	*20*	
""It was very painful when Zuhra's father passed away in a	*Sunday, May 20, 2018*	13:00:04+0000	393530784080887_1360284877405468	*19*	
Al-Shifa Trust Eye Hospital, Celebrating Jashn e Azadi...	*Wednesday, August 16, 2017*	05:33:19+0000	393530784080887_1134580439975914	*18*	Al-Shifa Trust Eye Hospital was live.
Al-Shifa Trust Eye Hospital's Dedicated team of Al-Shifa Ce	*Tuesday, April 18, 2017*	14:30:19+0000	393530784080887_1035315389902420	*17*	
Pakistan Institute of Ophthalmolgy (PIO), of Al-Shifa Trust I	*Saturday, December 2, 2017*	11:21:02+0000	393530784080887_1217726424994648	*17*	Al-Shifa Trust Eye Hospital added 38 new photo
ONE DAY WORKSHOP Teachers awareness and orientation	*Tuesday, September 27, 2016*	06:48:25+0000	393530784080887_870575343043093	*16*	
"Har Aankh Roshan" Live on Riphah FM 102.2, with Brig. D	*Monday, August 21, 2017*	06:18:06+0000	393530784080887_1138029039631054	*16*	Al-Shifa Trust Eye Hospital was live.
Fariha, 4 months old infant didn't open her eyes straight a	*Wednesday, May 11, 2016*	15:25:18+0000	393530784080887_791575450943083	*16*	
"Your Kindness Helps Stop Blindness" This Ramadaan, Sup	*Monday, May 29, 2017*	04:10:44+0000	393530784080887_1070119256412033	*16*	

FIGURE 4.5: Power BI View of Automated Dataset

Transformed data was then imported and was further processed for visualization, number of comments for all posts were set to decreasing order. Different type of bar and pie charts were made to get a better understanding of the data gathered. Each part of the chart showed number of comments on each post in descending order. Below is the bar chart of AlShifa Eye Hospital showing the number of comments along with the post.

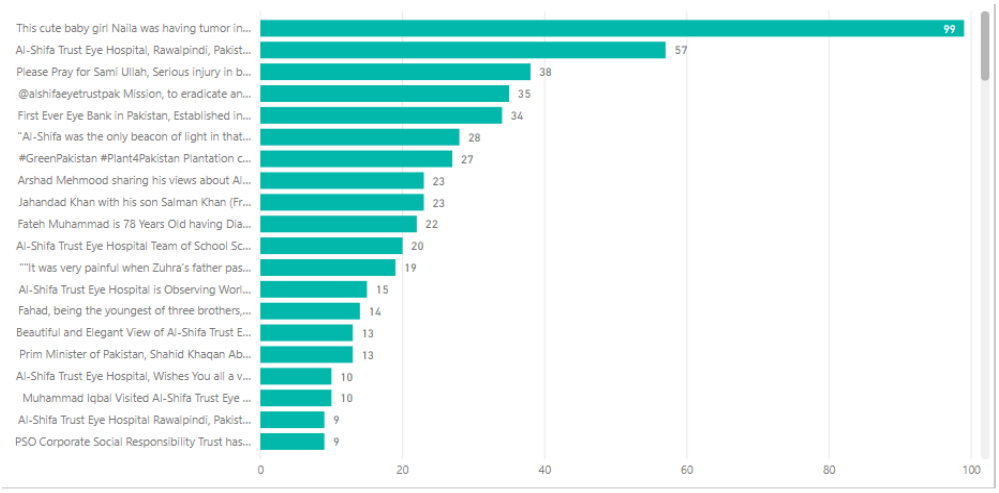


FIGURE 4.6: Bar Chart of Power BI dataset

Following is the screenshot of the pie chart for District Health Authority Jhang's Facebook posts.

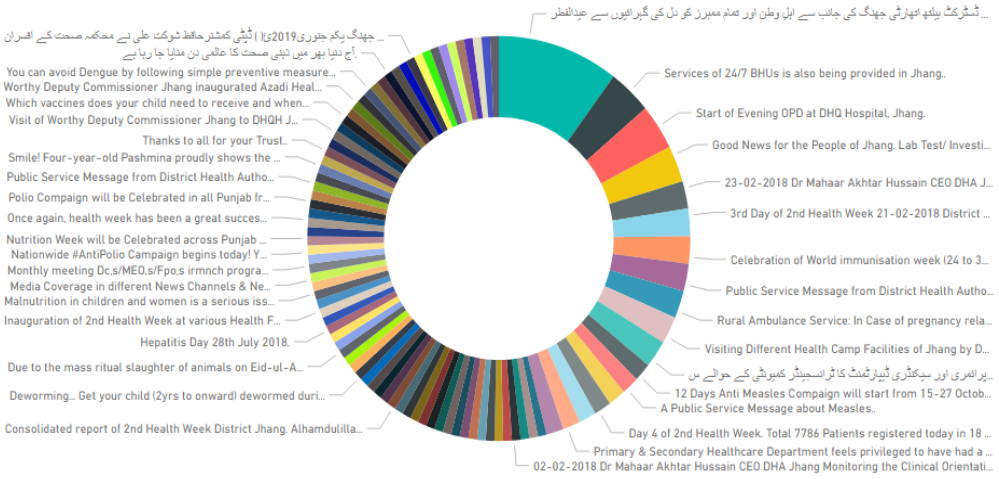


FIGURE 4.7: Pie Chart of Power BI Dataset

This data was of one year that was from January 2018 to January2019. Earlier the data was mined as a bunch but for the graphical representation one year's data was included in the pictorial representation.

5 Analysis

5.1 Introduction

This chapter gives a detailed explanation of how the datasets were preprocessed for analysis, and how the analysis phase was accomplished. Section 5.2 describes the preprocessing of manual dataset and the details about the keywords used for representation of different attributes. Further in this chapter the statistical and predictive analysis of manual as well as automated dataset was discussed in detail. In section 5.8 those factors were elaborated which had impact on this study and can impact further researches of this type in detail.

5.2 Pre Processing of Manual Dataset

With the end goal of factual examination of the dataset SPSS was used, as it only takes scale, nominal and ordinal data so the data from the datasets was first changed over into these data types [IBM]. Making use of the manual dataset generated in xlsx documents it was put into SPSS in more precise form and attributes were added for detailed analysis of data. Many new keywords were added to categorize the diseases to get precise results, eleven excel files were imported to eleven SPSS files.

Views	Date	Post_Type	Post_Content	Likes	Comments	Shares
.	23-Feb-2018	Advertisem...	Spinal Disorders	3	0	0
.	23-Feb-2018	Awareness	Spinal Disorders	5	0	2
.	24-Feb-2018	Awareness	Thyroid Disorder	5	0	0
.	24-Feb-2018	Advertisem...	Andrology	2	0	0
.	24-Feb-2018	Awareness	Digestive Disorders	2	0	0
.	25-Feb-2018	Awareness	Spinal Disorders	1	0	0
.	25-Feb-2018	Awareness	Andrology	2	2	0
.	26-Feb-2018	Awareness	Gynae and Obs	1	0	0
.	26-Feb-2018	Awareness	Gynae and Obs	2	0	0
.	27-Feb-2018	Advertisem...	Digestive Disorders	4	0	0
.	27-Feb-2018	Advertisem...	STD	1	0	0
.	27-Feb-2018	Awareness	Mental Health	1	0	0
.	27-Feb-2018	Awareness	Mental Health	5	0	0

FIGURE 5.1: SPSS View of Manual Dataset

In this way eleven SPSS files were created by importing data from excel files to SPSS. Then the data of these files was merged in a single SPSS file, one active file was merged in other SPSS file first using this code.

```
GET DATA /TYPE=XLSX
  /FILE='D:\Manual DataSet\BilalHospital.xlsx'
  /SHEET=name 'BilalHospital'
  /CELLRANGE=full
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.
```

FIGURE 5.2: Imported Dataset to SPSS

After merging the file some more attributes were added to the file and some extra attributes such as post's link were removed for better understanding of data, eleven attributes included in this file were.

- Views
- Date
- Post_Type
- Post_Content
- Likes
- Comments
- Shares
- Org_Type
- City
- Org_Name
- Cardinal_Points

For the final analysis, different attributes were assigned different types according to their nature. The attributes Date, Views, Likes, Comments and Shares were assigned the data type "Scale" because all of these attributes were consisted of numerical values. The remaining attributes were assigned "Nominal" data type because they were stored as text and had no quantitative value. Later on some quantitative values were assigned to the data of these attributes as text cannot be analyzed quantitatively. The attributes along with assigned values and details are given below.

5.2.1 Organization's Type

As this was quantitative analysis so numbers were assigned to nominal data for clear and better results. First the organization's type column was given the following values.

Organization's Type			
1.00	2.00	3.00	4.00
Govt.	NGO.	Non Profit	For Profit

FIGURE 5.3: Keywords Assigned to Organization's Type

5.2.2 Post's Type

In the same way the other attribute that was Post's_Type was also assigned values for each entry so that quantitative analysis could be done easily. The values assigned to the type of the post are mentioned in the table below.

Post Type	Number Assigned	Post Type	Number Assigned
Alert	5	Jobs	50
Advertisement	10	Meeting	55
Awareness	15	Notification	60
Camps	20	Organizational News	65
Events	25	Promotional	70
Feedback	30	Public Health Advice	75
Fundraising	35	Social Cause	80
General	40	Visits	85
Health Week	45		

TABLE 5.1: Keywords Assigned to Post's Type

5.2.3 Post's Content

The next attribute was Post's_Content, different numbers were assigned to different health areas to have a better understanding of the quantitative results. A specific number had a specific keyword and the keyword has many diseases associated with it. The numbers assigned to divergent keywords are shown in the table below.

Keyword	Number Assigned	Keyword	Number Assigned
Eye Health	6	Drug Control	32
Digestive Disorders	7	Mental Health	33
Spinal Disorder	8	Andrology	37
Cancer	9	Gynae and Obs	39
Cardiac Health	11	STD	41
Hepatitis	12	Obesity	43
Pain and Fever	14	UTI	44
Nutrition	16	Thyroid Disorder	46
Kidney Diseases	17	Bacterial Infections	47
Dermatology	18	General	48
Diabetes	19	General Health	49
Viral Infections	23	Blood Disorders	51
Child Health	26		

TABLE 5.2: Keywords Assigned to Post's Content

Each of these keywords are comprised of many diseases the diseases covered in each of the keyword is as follows:

- Eye Health

  The keyword "Eye Health" is comprised of the diseases Glaucoma, Blindness, Refractive error, Tumor in Eye, Corneal opacity, Lost vision, Diabetic Retinopathy in Eyes, Cornea Perforation, Cornea Problem, Eye Grafting and Medial Canthal Reconstruction, Congenital Contract, Low Vision, Diseases of Oculoplastics, Eye Bank, Eye Diseases, Visual Impairment, Corneal Diseases and Eye health issues due to Smog.

- Digestive Disorders

  The keyword "Digestive Disorders" include Diarrhoeal Diseases, Foodborne Diseases, Vomiting and Abdominal Distress, Irritable bowel Syndrome (IBS), Liver Problem, Gastrointestinal Disorders and Anorectal Disorders.

- Spinal Disorder

  Many diseases are put under the category of "Spinal Disorder" including Back Pain, Leg Pain, Spinal Diseases, Arthritis, Lower Back Pain, Neck Pain, Knee Osteoarthritis, Rheumatism, Spinal Wellbeing, Spinal Disc Herniation, Disc Bulge, Text Neck, Scoliosis, Cervical Spinal Asymmetry, Spondylitis, Slip Disc, Deformity, Joint Pain, Muscle Pain, Sacrum Pain, Frozen Shoulder and Osteoporosis.

- Cancer

  Different types of Cancer is incorporated with the keyword "Cancer" such as Breast Cancer, Cervical Cancer, Cancer caused by Tobacco, Cancer and HIV.

- Cardiac Health

  Cardiac Health includes Angina, Heart Diseases, High BP, Hypertension and Cholesterol.

- Hepatitis

  This keyword includes Hepatitis A, Hepatitis B, Hepatitis C, Hepatitis D, Hepatitis E, HIV and Hepatitis.

- Pain and Fever

  The diseases included under the category of this keyword are Stroke, Pain and Fever.

- Nutrition

  Under the category of "Nutrition" keyword the posts about Nutrition, Malnutrition and Balanced Diet are merged.

- Bacterial Infections

  The posts related to Tuberculosis and Antibiotics are placed under the keyword "Bacterial Infections".

- Viral Infections

  Posts relevant to HIV/AIDS, Seasonal Influenza, Polio, Congo Virus, Flu, Dengue, Typhoid, Measles, Cough, Cold and Flu are put down under the category of "Viral Infections".

- Gynae and Obs

  All posts regarding female health such as maternal health and health issues of unmarried girls are unified under the keyword "Gynae and Obs".

- Blood Disorders

  Different posts about blood disorders such as Blood Cancer, Thalassemia and Hemophilia are placed in this category.

- General Health

  All posts having information about Health Week, Drugs Inspection, Monitoring of different Health Organizations, Facilities available in Health Organizations like information about Ultrasound machine or ambulances, Briefings about health issues and other posts regarding healthcare sector are included in "General Health".

- General

  All other posts on the pages of health organizations such as wishes of Eid, Jumma, Labour Day and these sort of days are put under "General" section. Moreover admissions, job opening and all the posts that were vague and not giving information about a specific disease were put under this category. Some common examples of general posts are Labor Day, Eid ul Fitr, Admissions, Eid Mubarik, , New Year, Pakistan Resolution Day, Award ceremony, Job opening and Jumma Mubarik.

5.2.4 City

Different cities were assigned different numbers for better understanding of the statistical data. The cities included in this file where the healthcare organizations were currently established along with the numbers assigned to those cities are as under.

City				
101	102	103	104	110
Lahore	Rawalpindi	Jhang	Chiniot	Gujranwala

TABLE 5.3: Keywords Assigned to Cities

Different organizations working on different health issues were examined and those were from the above mentioned cities.

5.2.5 Cardinal Points

For a better understanding about the area of the organization working for specific health issue a map of districts of Punjab was made.

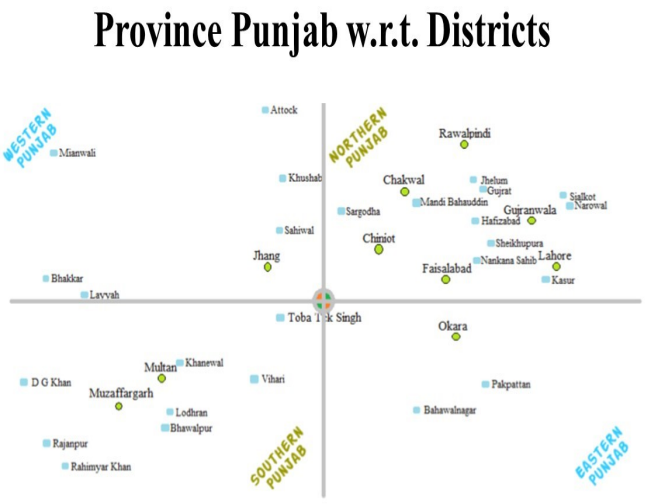


FIGURE 5.4: Map of Province Punjab w.r.t. Districts

Cardinal points of all the districts were noted down and each cardinal point was assigned a specific value to get a better understanding about the geo location of the organization w.r.t. city. The values assigned to different cardinal points are in the given table.

Cardinal Points			
333	555	777	999
North	South	West	East

TABLE 5.4: Keywords Assigned to Cardinal Points

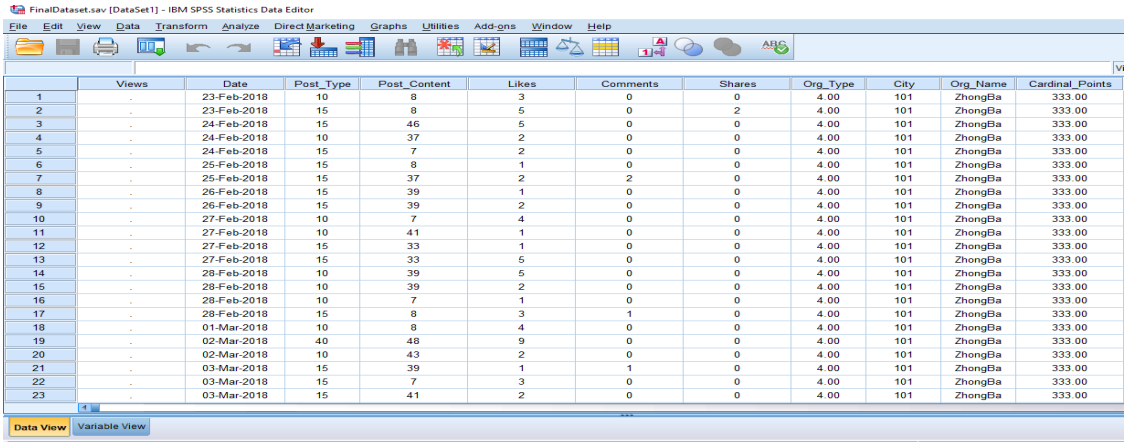
5.2.6 Final View of Gathered Data

Here data was refined and each attribute was assigned different data type such as numeric, date and string. Moreover different values were assigned to the nominal data in these attribute and all other specified labels were given. The variable view of this collection of data is given below.

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Views	Numeric	12	0		None	None	12	Center	Scale	Input
Date	Date	11	0		None	None	11	Center	Scale	Input
Post_Type	String	13	0		{10, Adverti...	None	8	Center	Nominal	Input
Post_Content	String	19	0		{11, Cardiac...	None	12	Center	Nominal	Input
Likes	Numeric	12	0		None	None	12	Center	Scale	Input
Comments	Numeric	12	0		None	None	12	Center	Scale	Input
Shares	Numeric	12	0		None	None	12	Center	Scale	Input
Org_Type	Numeric	8	2		{1.00, Govt}...	None	8	Center	Nominal	Input
City	String	33	0		{101, Lahor...	None	8	Center	Nominal	Input
Org_Name	String	8	0		None	None	8	Center	Nominal	Input
Cardinal_Po...	Numeric	8	2		{333.00, Nor...	None	12	Center	Nominal	Input

FIGURE 5.5: Variable View of Manual Dataset

The final data view of this refined set of data is as under.



	Views	Date	Post_Type	Post_Content	Likes	Comments	Shares	Org_Type	City	Org_Name	Cardinal_Points
1	.	23-Feb-2018	10	8	3	0	0	4.00	101	ZhongBa	333.00
2	.	23-Feb-2018	15	8	5	0	2	4.00	101	ZhongBa	333.00
3	.	24-Feb-2018	15	46	5	0	0	4.00	101	ZhongBa	333.00
4	.	24-Feb-2018	10	37	2	0	0	4.00	101	ZhongBa	333.00
5	.	24-Feb-2018	15	7	2	0	0	4.00	101	ZhongBa	333.00
6	.	25-Feb-2018	15	8	1	0	0	4.00	101	ZhongBa	333.00
7	.	25-Feb-2018	15	37	2	2	0	4.00	101	ZhongBa	333.00
8	.	26-Feb-2018	15	39	1	0	0	4.00	101	ZhongBa	333.00
9	.	26-Feb-2018	15	39	2	0	0	4.00	101	ZhongBa	333.00
10	.	27-Feb-2018	10	7	4	0	0	4.00	101	ZhongBa	333.00
11	.	27-Feb-2018	10	41	1	0	0	4.00	101	ZhongBa	333.00
12	.	27-Feb-2018	15	33	1	0	0	4.00	101	ZhongBa	333.00
13	.	27-Feb-2018	15	33	5	0	0	4.00	101	ZhongBa	333.00
14	.	28-Feb-2018	10	39	5	0	0	4.00	101	ZhongBa	333.00
15	.	28-Feb-2018	10	39	2	0	0	4.00	101	ZhongBa	333.00
16	.	28-Feb-2018	10	7	1	0	0	4.00	101	ZhongBa	333.00
17	.	28-Feb-2018	15	8	3	1	0	4.00	101	ZhongBa	333.00
18	.	01-Mar-2018	10	8	4	0	0	4.00	101	ZhongBa	333.00
19	.	02-Mar-2018	40	48	9	0	0	4.00	101	ZhongBa	333.00
20	.	02-Mar-2018	10	43	2	0	0	4.00	101	ZhongBa	333.00
21	.	03-Mar-2018	15	39	1	1	0	4.00	101	ZhongBa	333.00
22	.	03-Mar-2018	15	7	3	0	0	4.00	101	ZhongBa	333.00
23	.	03-Mar-2018	15	41	2	0	0	4.00	101	ZhongBa	333.00

FIGURE 5.6: Data View of SPSS Dataset

5.3 Implementation/ Statistical Analysis

5.3.1 Trending Post

For each row a score was calculated to observe the trending post, it was calculated by adding the number of likes, comments and shares. So the post having the highest number of score was on trending among these organizations. It showed that public interact with highest scoring posts more so these diseases are on trending and it showed maximum public involvement.

5.3.2 Analysis of Post Type

For the column of Post_Type different facet were analyzed to have a clear idea about what type of posts are in trending among the healthcare organizations examined. Missing and non-missing values for the attribute of post's type were calculated first. After that frequency table was designed to calculate different values of the attribute post's type, the frequency table consisted of the following values.

Frequency value that was calculated by using tally marks method and it represented that how many times this specific post type occurred in the whole sample.

Percent value which was basically the percentage of that specific frequency and was derived by dividing the frequency value by total number of sample consisting of all missing and non-missing values.

Valid percent value which was basically the value calculated via dividing the frequency value by number of non-missing values in the dataset. It was basically the valid percentage of non-missing values of the dataset.

Cumulative percent value was the value that represented the dataset's total percentage of each row. It was considered by adding the valid percentage of each row to the value of valid percentage of the above row. So as for the first row the value remained same as valid percent value but after that the calculated value was added to the valid percentage of the above row and so on. In this way frequency table for the attribute Post_Type was calculated.

FREQUENCIES VARIABLES=Post_Type /ORDER=ANALYSIS.

After the frequency table a pictorial representation of attribute Post_Type was prepared in the form of graph. A pie chart was made to show the percentages of each category of the post type attribute, different colors represented different categories of post type along with percentage and name of the category. The snippet for making pie chart for post type is

** Chart Builder. GGRAPH /GRAPHDATASET NAME="graphdataset" VARIABLES=Post_Type COUNT()[name="COUNT"] MISSING=LISTWISE REPORTMISSING=NO*

5.3.3 Analysis of Post_Content

The procedure to calculate the values of post's content was same as used for post's type. Missing and non-missing values were examined on the first hand and then a frequency table was calculated for the attribute of Post_Content. The values calculated in the frequency table were Frequency value, Percent value, Valid percent value and Cumulative percent value. The representation and the process of calculating these values were same as for the values of frequency table in post type attribute. The snippet used here for the frequency table was

FREQUENCIES VARIABLES=Post_Content /ORDER=ANALYSIS.

After the frequency table a pie chart was made for the pictorial representations of the post content categories. Different colors represented different ailments along with their percentages in the dataset for a clear idea about the most occurring diseases in the dataset. The snippet used for the pie chart is

** Chart Builder. GGRAPH /GRAPHDATASET NAME="graphdataset" VARIABLES=Post_Content COUNT()[name="COUNT"] MISSING=LISTWISE REPORTMISSING=NO*

Another pie chart was made to observe the maximum number of score with respect to post content, this chart represented that on what type of diseases had maximum number of likes, comments and shares. The maximum user interaction was on the posts about Eye Health, after that Blood Disorder Diseases and then came General posts.

Likewise one more observation in the form of pie chart was analyzed which represented that which type of organization has the highest number of scores. It showed that general public interact with what type of organizations the most and which organization type had maximum user's likes, comments and shares. The snippet used for calculating this is

** Chart Builder. GGRAPH /GRAPHDATASET NAME="graphdataset" VARIABLES=Org_Type SUM(Score)[name="SUM_Score"] MISSING=LISTWISE REPORTMISSING=NO*

The results of the pie chart showed that users were maximum active on the posts of different NGOs and Non Profit Organizations, as these type of organizations have frequent postings on social media. Government and For Profit organizations are less active on social media so users visits those pages on random basis and were not very interactive on those posts.

5.4 Predictive Analysis of Manual Data Set

A Classification Tree was used for the predictive analysis of the whole data set, a tree based model for classification was designed by using decision tree. It predicted the value that was the dependent value based upon independent values by making some rules. All the types of data was included in this analysis such as nominal, ordinal and scale data. The dependent variable was Post_Content which was nominal data.

In independent variables Views, Date, Post_Type, Likes, Comments, Shares, Org_Type, City, Org_Name, Cardinal_Points and Score were included. These were of nominal and scale data types' variables.

CHAID that is "Chi-Squared Automatic Interaction Detection" method was used for the predictive analysis. It merges dependent variables if they are not significantly dissimilar from each other, and calculates a predicted value that is dependent variable based upon the predictors that are independent variables. The growth limits and criteria was set, Pearson method was used for fast calculations and scale intervals for fixed variables was set. To save the information in the data file from the given model a predicted value was selected which was produced for the dependent variable by analyzing all independent variables in the model.

A model summary was derived having information about the method used in the model along with information about the variables included for prediction and also the variables specified but were excluded from the model. A risk and classification table was made to have an idea about the predictive accuracy of tree which showed the risk estimate of the cases. The tree merged the values if independent nodes in one node if they have same values. For each node there was a frequency table showing the percentage and count of diseases that was the dependent variable. The snippet for the classification tree is:

** Decision Tree. TREE Post_Content [n] BY Views [s] Date [s] Post_Type [n] Likes [s] Comments [s] Shares [s] Org_Type [n] City [n] Org_Name [n] Cardinal_Points [n] Score [s] /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO /DEPCATEGORIES USEVALUES=['11' '12' '14' '16' '17' '18' '19' '23' '26' '32' '33' '37' '39' '41' '43' '44' '46' '47' '48' '49' '51' '6' '7' '8' '9'] /PRINT MODELSUMMARY CLASSIFICATION RISK /SAVE PREDVAL /METHOD TYPE=CHAID /GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=100 MINCHILDSIZE=50 /VALIDATION TYPE=NONE OUTPUT=BOTHSAMPLES /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05 SPLITMERGED=NO CHISQUARE=PEARSON CONVERGE=0.001 MAXITERATIONS=100 ADJUST=BONFERRONI INTERVALS=10 /COSTS EQUAL /MISSING NOMINALMISSING=MISSING.*

After the analysis of the manual dataset another column that was predicted value was added to the data. This column was showing the predicted value of dependent variable based upon the

independent variables. All these predicted values were compared to the original values of post content variable to check the accuracy level of the predicted value. After the final analysis the accuracy level for the predicted value based upon the tree model was approximately 70%.

A screenshot of the final data representing the score along with the predicted value is as follows.

	Views	Date	Post_Type	Post_Cont...	Likes	Comments	Shares	Org_Type	City	Org_Name	Cardinal_Points	Score	PredictedValue_1
1	.	23-Feb-2018	10	8	3	0	0	4.00	101	ZhongBa	333.00	3.00	39
2	.	23-Feb-2018	15	8	5	0	2	4.00	101	ZhongBa	333.00	7.00	39
3	.	24-Feb-2018	15	46	5	0	0	4.00	101	ZhongBa	333.00	5.00	39
4	.	24-Feb-2018	10	37	2	0	0	4.00	101	ZhongBa	333.00	2.00	39
5	.	24-Feb-2018	15	7	2	0	0	4.00	101	ZhongBa	333.00	2.00	39
6	.	25-Feb-2018	15	8	1	0	0	4.00	101	ZhongBa	333.00	1.00	39
7	.	25-Feb-2018	15	37	2	2	0	4.00	101	ZhongBa	333.00	4.00	39
8	.	26-Feb-2018	15	39	1	0	0	4.00	101	ZhongBa	333.00	1.00	39
9	.	26-Feb-2018	15	39	2	0	0	4.00	101	ZhongBa	333.00	2.00	39
10	.	27-Feb-2018	10	7	4	0	0	4.00	101	ZhongBa	333.00	4.00	39
11	.	27-Feb-2018	10	41	1	0	0	4.00	101	ZhongBa	333.00	1.00	39
12	.	27-Feb-2018	15	33	1	0	0	4.00	101	ZhongBa	333.00	1.00	39
13	.	27-Feb-2018	15	33	5	0	0	4.00	101	ZhongBa	333.00	5.00	39
14	.	28-Feb-2018	10	39	5	0	0	4.00	101	ZhongBa	333.00	5.00	39
15	.	28-Feb-2018	10	39	2	0	0	4.00	101	ZhongBa	333.00	2.00	39
16	.	28-Feb-2018	10	7	1	0	0	4.00	101	ZhongBa	333.00	1.00	39
17	.	28-Feb-2018	15	8	3	1	0	4.00	101	ZhongBa	333.00	4.00	39
18	.	01-Mar-2018	10	8	4	0	0	4.00	101	ZhongBa	333.00	4.00	39
19	.	02-Mar-2018	40	48	9	0	0	4.00	101	ZhongBa	333.00	9.00	39
20	.	02-Mar-2018	10	43	2	0	0	4.00	101	ZhongBa	333.00	2.00	39
21	.	03-Mar-2018	15	39	1	1	0	4.00	101	ZhongBa	333.00	2.00	39
22	.	03-Mar-2018	15	7	3	0	0	4.00	101	ZhongBa	333.00	3.00	39
23	.	03-Mar-2018	15	41	2	0	0	4.00	101	ZhongBa	333.00	2.00	39

FIGURE 5.7: Final View of Manual Dataset

5.5 Statistical Analysis of Automated Dataset

As the data was in PowerBi and graphs were made for the data having the maximum number of comments. That graphical data was then loaded in SPSS for further statistical analysis, so a new dataset was generated into SPPS having the following attributes.

Sr. No.	Attribute	Description
1	Org_Type	Describes the type of the organization i-e Govt, NGOs, For Profit and Not for Profit organizations.
2	Post_Content	The content of the post i-e information about specific ailments.
3	No_of_Comments	The total number of comments of the post.
4	Org	The name of the organization.
5	City	The city where the organization exists.

TABLE 5.5: Attributes of Automated Dataset

The data view after importing the graphical data in SPSS is in the screenshot below.

	Org_Type	Post_Content	No_of_Comments	Org	City
1	NGO	General	216.00	Alkhidmat	Lahore
2	NGO	General	129.00	Alkhidmat	Lahore
3	NGO	General	108.00	Alkhidmat	Lahore
4	NGO	General	71.00	Alkhidmat	Lahore
5	NGO	General	54.00	Alkhidmat	Lahore
6	NGO	General	48.00	Alkhidmat	Lahore
7	NGO	General	33.00	Alkhidmat	Lahore
8	NGO	General	33.00	Alkhidmat	Lahore
9	NGO	General	29.00	Alkhidmat	Lahore
10	NGO	General	26.00	Alkhidmat	Lahore
11	NGO	General	26.00	Alkhidmat	Lahore
12	NGO	General	21.00	Alkhidmat	Lahore
13	NGO	General	19.00	Alkhidmat	Lahore
14	NGO	General	18.00	Alkhidmat	Lahore
15	NGO	General	17.00	Alkhidmat	Lahore
16	Non Profit	Eye Health	99.00	Alshifa	Rawalpindi
17	Non Profit	Eye Health	57.00	Alshifa	Rawalpindi
18	Non Profit	Eye Health	38.00	Alshifa	Rawalpindi
19	Non Profit	Eye Health	35.00	Alshifa	Rawalpindi
20	Non Profit	Eye Health	34.00	Alshifa	Rawalpindi
21	Non Profit	Eye Health	28.00	Alshifa	Rawalpindi
22	Non Profit	General	27.00	Alshifa	Rawalpindi

FIGURE 5.8: SPSS View of Automated Dataset

As the data here was nominal like the data of manual dataset, except the number of comments which had numeric data. The keywords assigned to the different attributes in this dataset are explained below.

5.5.1 Organization's Type

As for statistical analysis text data can't be processed in SPSS so these keywords were assigned to the type of organizations in the dataset.

Organization's Type			
1.00	2.00	3.00	4.00
Govt.	NGO.	Non Profit	For Profit

FIGURE 5.9: Keywords for Organization's Type in Automated Dataset

5.5.2 Post's Content

For different diseases different keywords were used for a better understanding of data which are mentioned in the table below.

Keyword	Number Assigned	Keyword	Number Assigned
Eye Health	6	Child Health	26
Digestive Disorders	7	AIDS/HIV	27
Spinal Disorder	8	Mental Health	33
Cancer	9	Andrology	37
Cardiac Health	11	Gynae and Obs	39
Hepatitis	12	Bacterial Infections	47
Nutrition	16	General	48
Kidney Diseases	17	General Health	49
Diabetes	19	Blood Disorders	51
Viral Infections	23		

TABLE 5.6: Keywords for Post's Content in Automated Dataset

Each of the keyword is comprised of many diseases relevant to the keyword such as eye health represents many sort of eye diseases. All of the keywords have same set of ailments defined for the manual dataset.

5.5.3 City

The organizations selected for the analysis were from the different cities of Punjab Pakistan, list of the cities along with the keyword by which they are represented are as follows.

City	Keyword	City	Keyword
Lahore	101	Chakwal	106
Rawalpindi	102	Muzaffargarh	107
Jhang	103	Faisalabad	108
Chiniot	104	Okara	109
Gujranwala	105	Multan	111

TABLE 5.7: Keywords for Cities in Automated Dataset

The final variable view of the gathered and processed data is as under.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Org_Type	Numeric	8	2		{1.00, Govt}...	None	18	Center	Nominal	Input
2	Post_Content	Numeric	33	0		{6, Eye Hea...	None	28	Center	Nominal	Input
3	No_of_Com...	Numeric	8	2		None	None	14	Center	Scale	Input
4	Org	String	9	0		None	None	18	Center	Nominal	Input
5	City	Numeric	17	0		{101, Lahor...	None	8	Center	Nominal	Input

FIGURE 5.10: Variable View of Automated Dataset

The final data view of the gathered and processed data is as under.

	Org_Type	Post_Content	No_of_Comments	Org	City
1	2.00	48	216.00	Alkhidmat	101
2	2.00	48	129.00	Alkhidmat	101
3	2.00	48	108.00	Alkhidmat	101
4	2.00	48	71.00	Alkhidmat	101
5	2.00	48	54.00	Alkhidmat	101
6	2.00	48	48.00	Alkhidmat	101
7	2.00	48	33.00	Alkhidmat	101
8	2.00	48	33.00	Alkhidmat	101
9	2.00	48	29.00	Alkhidmat	101
10	2.00	48	26.00	Alkhidmat	101
11	2.00	48	26.00	Alkhidmat	101
12	2.00	48	21.00	Alkhidmat	101
13	2.00	48	19.00	Alkhidmat	101
14	2.00	48	18.00	Alkhidmat	101
15	2.00	48	17.00	Alkhidmat	101
16	3.00	6	99.00	Alshifa	102
17	3.00	6	57.00	Alshifa	102
18	3.00	6	38.00	Alshifa	102
19	3.00	6	35.00	Alshifa	102
20	3.00	6	34.00	Alshifa	102
21	3.00	6	28.00	Alshifa	102
22	3.00	48	27.00	Alshifa	102

FIGURE 5.11: Data View of Automated Dataset

5.5.4 Statistical Analysis

Because of the limited attributes in the automated dataset, the main focus of analysis was the post content and number of comments attribute. A frequency table of attribute Post_Content was designed using the snippet.

FREQUENCIES VARIABLES=Post_Content /ORDER=ANALYSIS.

For the post content missing and non-missing values were calculated to have a clear about the valid and missing entries of the data. Frequency for each category of the sample was computed by means of tally mark method. Percent Value was calculated that was primarily the percentage of the

frequency, obtained by dividing the frequency value over sum of missing and non-missing values in the sample set. Valid percent was also determined for post content variable by dividing the frequency value by non-missing values of the sample data. Cumulative percent was also retrieved by adding the valid percent to the percent of the row above. After creating the frequency table of this data a pie chart was made to represent the percentage of different diseases present in the dataset. Different colors showed the percentage of different maladies present in the given set of data using the snippet.

** Chart Builder. GGRAPH /GRAPHDATASET NAME="graphdataset" VARIABLES=Post_Content COUNT()[name="COUNT"] MISSING=LISTWISE REPORTMISSING=NO /GRAPHSPEC SOURCE=INL*

Another chart was also made to analyze the post content according to maximum number of comments on that post. The chart showed what sort of ailments had highest number of comments in the given dataset.

** Chart Builder. GGRAPH /GRAPHDATASET NAME="graphdataset" VARIABLES=Post_Content SUM(No_of_Comments)[name="SUM_No_of_Comments"] MISSING=LISTWISE REPORTMISSING=NO /GRAPHSPEC SOURCE=INLINE.*

5.6 Predictive Analysis of Automated Dataset

A Classification Tree was used for the predictive analysis of the whole data set, a tree based model for classification was designed by using decision tree. It predicted the value that was the dependent value based upon independent values by making some rules. The data types included were nominal and scale data.

The dependent variable was Post_Content which was nominal data.

In independent variables Org_Type, No_of_Comments, Org and City were included. These were of nominal and scale data types' variables. CHAID that is "Chi-Squared Automatic Interaction Detection" method was used for the predictive analysis. It merged those independent variables that were not significantly dissimilar from each other, and calculated a predicted value that was dependent variable based upon the predictors that are independent variables. The growth limits and criteria was set, Pearson method was used for fast calculations and scale intervals for fixed variables was set. To save the information in the data file from the given model a predicted value was selected which was produced for the dependent variable by analyzing all independent variables in the model.

A model summary was derived having information about the method used in the model along with information about the variables included for prediction and also the variables specified but were excluded from the model. A risk and classification table was made to have an idea about the

predictive accuracy of tree which showed the risk estimate of the cases. The tree merged the values if independent nodes in one node if they have same values. For each node there was a frequency table showing the percentage and count of diseases that was the dependent variable. The snippet for the classification tree is:

** Decision Tree. TREE Post_Content [n] BY Org_Type [n] No_of_Comments [s] Org [n] City [n] /TREE DISPLAY=TOPDOWN NODES=STATISTICS BRANCHSTATISTICS=YES NODEDEFS=YES SCALE=AUTO /DEPCATEGORIES USEVALUES=[6 7 8 9 11 12 16 17 19 23 26 27 33 37 39 47 48 49 51] /PRINT MODELSUMMARY CLASSIFICATION RISK /SAVE PREDVAL /METHOD TYPE=CHAID /GROWTHLIMIT MAXDEPTH=AUTO MINPARENTSIZE=100 MINCHILDSIZE=50 /VALIDATION TYPE=NONE OUTPUT=BOTHSAMPLES /CHAID ALPHASPLIT=0.05 ALPHAMERGE=0.05 SPLITMERGED=NO CHISQUARE=PEARSON CONVERGE=0.001 MAXITERATIONS=100 ADJUST=BONFERRONI INTERVALS=10 /COSTS EQUAL /MISSING NOMINALMISSING=MISSING.*

After the analysis of the automated dataset another column that was predicted value was added to the data. This column was showing the predicted value of dependent variable based upon the independent variables. All these predicted values were compared to the original values of post content variable to check the accuracy level of the predicted value. After the final analysis the accuracy level for the predicted value based upon the tree model was approximately 48%. A screenshot of the final data representing the score along with the predicted value is as follows.

	Org_Type	Post_Content	No_of_Comments	Org	City	PredictedValue_Final
1	2.00	48	216.00	Alkhidmat	101	48
2	2.00	48	129.00	Alkhidmat	101	48
3	2.00	48	108.00	Alkhidmat	101	48
4	2.00	48	71.00	Alkhidmat	101	48
5	2.00	48	54.00	Alkhidmat	101	48
6	2.00	48	48.00	Alkhidmat	101	48
7	2.00	48	33.00	Alkhidmat	101	48
8	2.00	48	33.00	Alkhidmat	101	48
9	2.00	48	29.00	Alkhidmat	101	48
10	2.00	48	26.00	Alkhidmat	101	48
11	2.00	48	26.00	Alkhidmat	101	48
12	2.00	48	21.00	Alkhidmat	101	48
13	2.00	48	19.00	Alkhidmat	101	48
14	2.00	48	18.00	Alkhidmat	101	48
15	2.00	48	17.00	Alkhidmat	101	48
16	3.00	6	99.00	Alshifa	102	6
17	3.00	6	57.00	Alshifa	102	6
18	3.00	6	38.00	Alshifa	102	6
19	3.00	6	35.00	Alshifa	102	6
20	3.00	6	34.00	Alshifa	102	6
21	3.00	6	28.00	Alshifa	102	6
22	3.00	48	27.00	Alshifa	102	6
23	3.00	6	23.00	Alshifa	102	6

FIGURE 5.12: Automated Dataset After Prediction

5.7 Impacting Factors in Research Process

Social media is a platform used by vast majority of overall population to share their views and organizations to share their information. Most popular SNS utilized worldwide as for dynamic users are Facebook, Youtube, WhatsApp, Facebook Messenger and so on [sta].

During the research phase it was observed that many factors have affected the research process, all those aspects were analyzed briefly. Following are some of the factors after analysis:

a. Changing Cyber Laws

Varying cyber security laws have a greater influence on social networking sites in regard of privacy concerns. The changing policies have an immense effect on the data mining techniques, when social media is used for data mining and analysis of certain type of data slight changes in the privacy policy have a prominent consequences on the results. In this research for social networking site Facebook, earlier there was access of huge amount of data through graph API including the number of likes, comments, shares, created date, time of a post from a public page. Number of related attributes for a post of public page were available for mining and analysis. But after the Facebook Cambridge Analytica Data Scandal the privacy policy for Facebook was changed and made according to the latest cyber laws [USA].

The privacy policy for Facebook was changed and changes were made according to the rules of GDPR. When these changes were made the access to the data was restricted, through data mining only some attributes were accessible for analysis. For Facebook the attributes mined were the posts of public pages along with the number of comments on those posts. So it was analyzed that if privacy of SNS is lose and users may access huge amount of data using data mining techniques, the accuracy rate after analysis of data is more as compared to the restricted access. When due to privacy concerns data access is restricted then because of getting less attributes from data mining the accuracy level after analysis is low. So changes in the privacy policy of Facebook may have a huge impact on the results of data mined from it and its accuracy.

b. Majority of Irrelevant Posts

It was observed that even on the pages of healthcare organizations, majority posts were in general. The posts regarding specific events and days along with general issues were more than posts with health information.

c. Less Interaction of Public

Another factor noticed was that the general public do not interact more on the posts regarding health issues. Specifically in terms of number of comments people do not comment most of the

time instead they just like or view the post, this thing has a great impact while analysis phase especially when attributes restrict due to privacy issues.

d. Frequency of Posts

The thing that has impacted the results of the research is the frequency of posts by healthcare organizations. NGOs are very active on social media and they post stuff frequently, but other governmental and for profit organizations do not post frequently. When for analysis random organizations are selected for analysis the samples for NGOs are greater than the samples of other organizations and it effects the results of the research.

6 Experimental Results

6.1 Introduction

In this chapter the experimental results derived after the implementation and analysis of both datasets are shown and explained. The end results after the statistical analysis and predictive analysis of manual and automated dataset are presented in this chapter. The tabular and graphical form of the results computed after the exploration are expounded in this chapter.

6.2 Results for Manual Dataset

After the analysis of manual dataset it was perceived that some attributes have integral role in the results. Those attributes were statistically analyzed and results were drawn, the first attribute of manual dataset in this regard was Post_Type.
First of all the missing and non-missing values were calculated and there was no missing values in this data. Further a frequency table was designed to calculate the frequency of each category in post type.

Post_Type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Advertisement	19	3.5	3.5	3.5
	Awareness	79	14.7	14.7	18.2
	Camps	14	2.6	2.6	20.8
	Events	37	6.9	6.9	27.7
	Feedback	34	6.3	6.3	34.0
	Fundraising	37	6.9	6.9	40.9
	General	15	2.8	2.8	43.7
	Health Week	10	1.9	1.9	45.5
	Alert	12	2.2	2.2	47.8
	Jobs	17	3.2	3.2	50.9
	Meeting	14	2.6	2.6	53.5
	Notification	6	1.1	1.1	54.6
	Org. News	89	16.5	16.5	71.2
	Promotional	44	8.2	8.2	79.4
	Public Health Advice	60	11.2	11.2	90.5
	Social Cause	5	.9	.9	91.4
	Visits	46	8.6	8.6	100.0
	Total	538	100.0	100.0	

TABLE 6.1: Statistics of Attribute Post Type of Manual Dataset

After the frequency table this data was put into graphical representation to have a clear idea about the types of the posts of healthcare organizations.

From the randomly selected organization's data it was analyzed the maximum posts were about the organizational news mostly about the activities carried out in organizations. Then came the posts of public awareness that contained data about different health issues and some information to aware public of certain health problems.

After awareness the majority post content was about public health advice and visits of health organization people to different areas along with visits of prominent personalities to hospitals. There were also promotional posts promoting the organization by posting about the facilities available at their organizations. Other categories which had low frequency in this sample set of data are also shown in the figure.

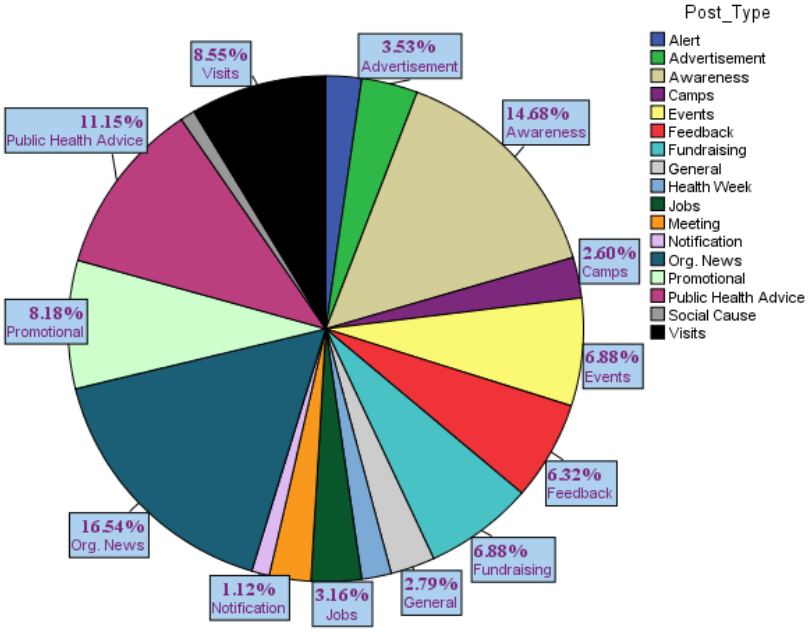


FIGURE 6.1: Graph Showing Post Type of Manual Dataset

The next and the most important attribute of the dataset was Post_Content, this was basically the type of disease mentioned in the posts. Content was directly about the disease or it gave clues or some relevant information about specific diseases.

Missing and non-missing values for this attribute were calculated and a frequency table was designed to show the frequency and the percentage of that frequency of each category included in the post content.

The information contained in the Frequency Table was.

Post Content

Frequency

Percent

Valid Percent

Cumulative Percent

Post_Content

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Cardiac Health	12	2.2	2.2	2.2
	Hepatitis	11	2.0	2.0	4.3
	Pain and Fever	3	.6	.6	4.8
	Nutrition	5	.9	.9	5.8
	Kidney Diseases	13	2.4	2.4	8.2
	Dermatology	2	.4	.4	8.6
	Diabetes	1	.2	.2	8.7
	Viral Infections	61	11.3	11.3	20.1
	Child Health	7	1.3	1.3	21.4
	Drug Control	20	3.7	3.7	25.1
	Mental Health	6	1.1	1.1	26.2
	Andrology	5	.9	.9	27.1
	Gynae and Obs	23	4.3	4.3	31.4
	STD	9	1.7	1.7	33.1
	Obesity	2	.4	.4	33.5
	UTI	2	.4	.4	33.8
	Thyroid Disorder	1	.2	.2	34.0
	Bacterial Infections	10	1.9	1.9	35.9
	General	77	14.3	14.3	50.2
	General Health	55	10.2	10.2	60.4
	Blood Disorder	39	7.2	7.2	67.7
	Eye Health	97	18.0	18.0	85.7
	Digestive Disorders	14	2.6	2.6	88.3
	Spinal Disorders	56	10.4	10.4	98.7
	Cancer	7	1.3	1.3	100.0
	Total	538	100.0	100.0	

TABLE 6.2: Statistics of Attribute Post Content in Manual Dataset

This frequency table was put under the pictorial representation to examine the maximum occurring disease in the data. This representation was for only the selected organizations' specific data, according to this data the most occurring disease among these organizations was Eye Diseases and the majority posts were about the eye health issues. After this category maximum posts were from general category those included posts about the days celebrated in Pakistan and quotes, people mostly interact on general posts instead of informative posts. Then comes the posts about viral infections, these posts included influenza, polio, measles, smog and all such type of viral infection in the region. Many ailments were put under the category of viral infection and are discussed in detail in chapter 5. After this category spinal disorder and general health posts were in majority among these organizations. Other maladies addressed in the posts of health care organizations are also shown on the graph.

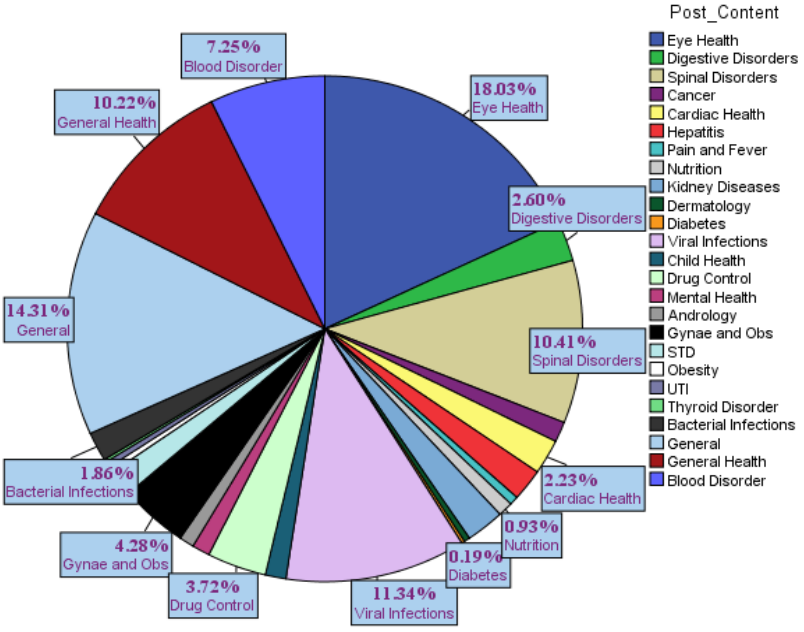


FIGURE 6.2: Graph Showing Post Content of Manual Dataset

Another graph was made to show the trending posts, these were basically the posts where maximum users interacted on posts. An attribute was derived on the basis of previous attributes that was score, score was the sum of number of likes, comments and shares. Score showed that on which post general public has made maximum number of likes, comments and shares and it depicted maximum user interaction. The posts where maximum number of users interacted were the posts about eye health issues, then came general and blood disorder problems. Post content with respect to maximum number of score is shown below.

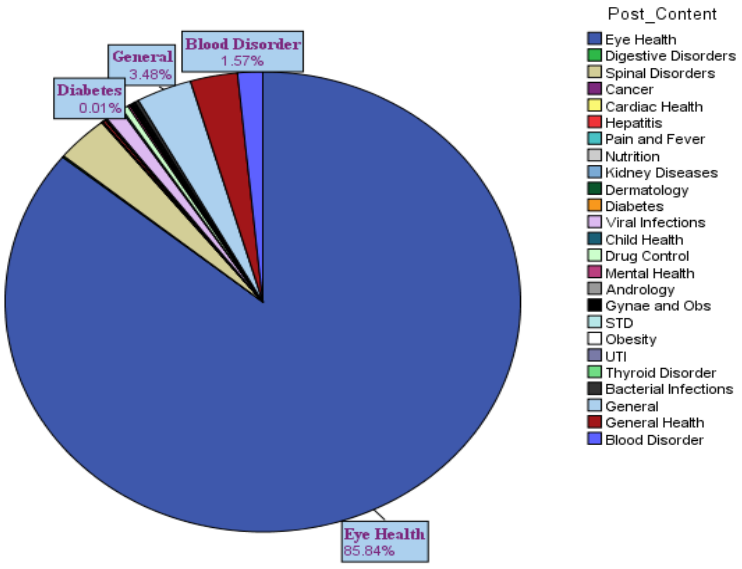


FIGURE 6.3: Post Content w.r.t. maximum score

After the analysis it was discerned that what type of organizations are mostly active on social media and general public mostly interact with what sort of organizations among these. On the basis of the score calculated it was examined that the majority active users are on the posts of NGOs and also NGOs are actively posting on Facebook about health informatics. Then comes the non profit organizations followed by governmental organizations. A chart showing this result is given below.

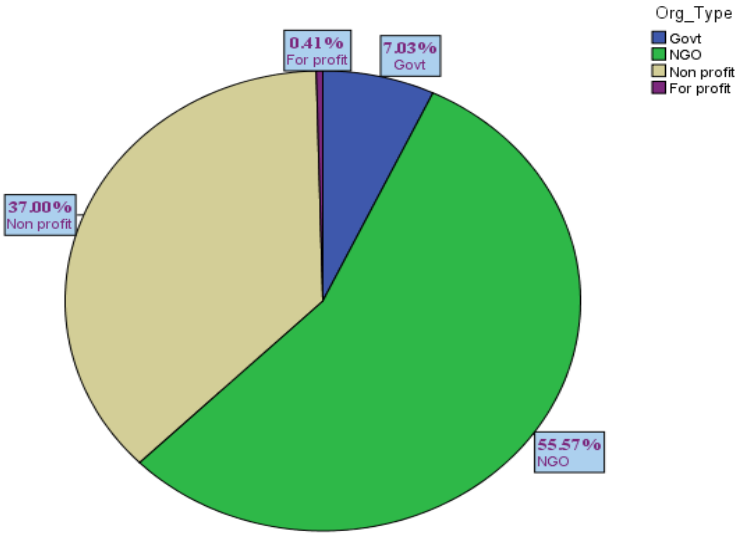


FIGURE 6.4: Type of Organization w.r.t. maximum score

Subsequently a predictive analysis was carried out on this dataset where classification tree was made to predict a value against the post content. Chi square Automatic Interaction Detection was used as a growing method for the tree classification. Post Content was the dependent variable in this tree and all other attributes in the manual dataset were independent variables or predictors of the data. The tree merged the data of same category and it gave the results according to the organization name with respect to post content. A model summary of the classification tree is in the table below.

Model Summary

	Growing Method	CHAID
	Dependent Variable	Post_Content
	Independent Variables	Views, Date, Post_Type, Likes, Comments, Shares, Org_Type, City, Org_Name, Cardinal_Points, Score
Specifications	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
	Independent Variables Included	Org_Name
Results	Number of Nodes	10
	Number of Terminal Nodes	9
	Depth	1

TABLE 6.3: Model summary of Manual Dataset

The classification tree along with all the relevant information derived is as follows.

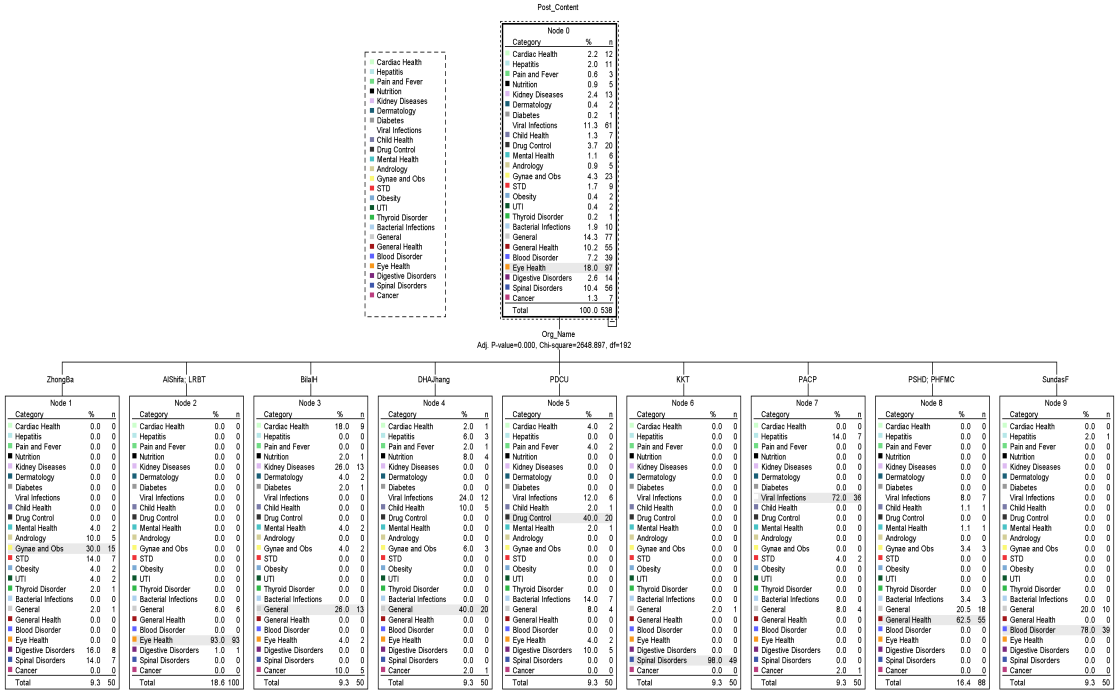


FIGURE 6.5: Classification Tree of Manual DataSet

This tree gave a predicted value against each value of post content, the predicted value was then compared to the original value of post content to check the accuracy level. The accuracy for this tree for this specific dataset was approximately 70%.

6.3 Results of Automated Dataset

Because of the restriction of the policies of Facebook, the attributes for automated dataset were less. The main and the most impacting attribute in automated dataset was Post_Content, missing and non-missing values for this attribute were calculated and there was no missing value in this sample set. A frequency table was made to calculate the frequency and percentage against each category included in post content variable. Valid percent for each category was also computed and then cumulative percent value was derived from these values. Below is the frequency table for the attribute post content.

Post_Content

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Eye Health	28	6.2	6.2	6.2
	Digestive Disorders	2	.4	.4	6.7
	Spinal Disorders	10	2.2	2.2	8.9
	Cancer	4	.9	.9	9.8
	Cardiac Health	3	.7	.7	10.4
	Hepatitis	11	2.4	2.4	12.9
	Nutrition	4	.9	.9	13.8
	Kidney Diseases	8	1.8	1.8	15.6
	Diabetes	1	.2	.2	15.8
	Viral Infections	10	2.2	2.2	18.0
	Child Health	24	5.3	5.3	23.3
	AIDS/HIV	14	3.1	3.1	26.4
	Mental Health	18	4.0	4.0	30.4
	Andrology	5	1.1	1.1	31.6
	Gynae and Obs	24	5.3	5.3	36.9
	Bacterial Infections	7	1.6	1.6	38.4
	General	110	24.4	24.4	62.9
	General Health	153	34.0	34.0	96.9
	Blood Disorders	14	3.1	3.1	100.0
	Total	450	100.0	100.0	

Table 6.4: Statistics Automated Dataset's Post Content

This table was put in graphical form to have a clear idea about the content of the posts. From the organizations included in the automated dataset majority posts were about general health issues, these were basically the posts about the health week and other such weeks to give information about health issues to general public. It also included the monitoring and inspection of different health care issues in organizations along with information about the primary and secondary health. Afterwards general posts were high in number having the information of different religious days

celebrated and quotes etc. Then came the eye health and child health issues followed by HIV/AIDS and blood disorder posts. Other posts having information of some other ailments are also shown on the graph.

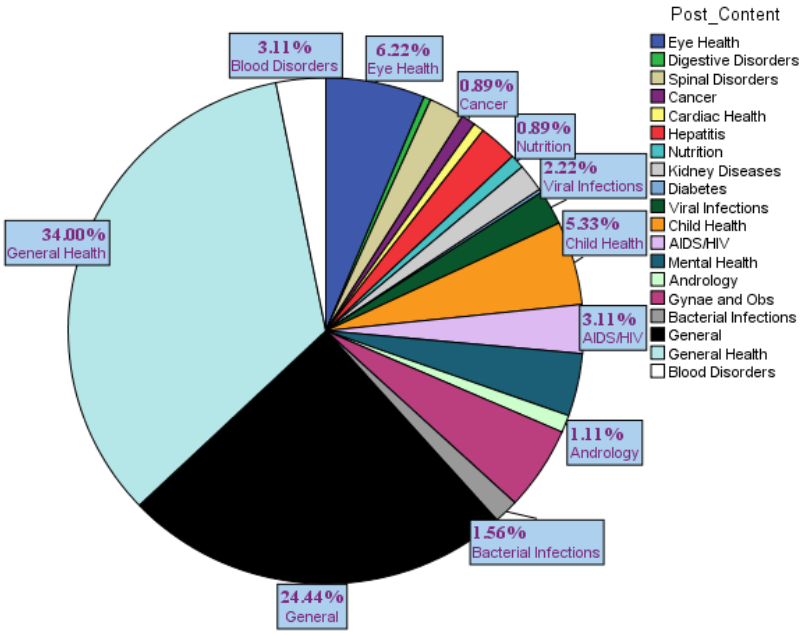


FIGURE 6.6: Graph of Post Content in Automated Dataset

One more graph was made to analyze the ailments with respect to highest number of comments. The posts about general health had the highest number of commented posts, people mostly appreciate the activities of health week and information distributed through these kind of activities. Then came the general posts as these were about some days like Jumma Mubarak or Eid Day, people mostly comment on such posts and interact with these posts.

Further posts having more comments were about Gynae and obs followed by bacterial infections, kidney diseases and mental health problems. Other posts having minimum user interaction are also shown on the graph below.

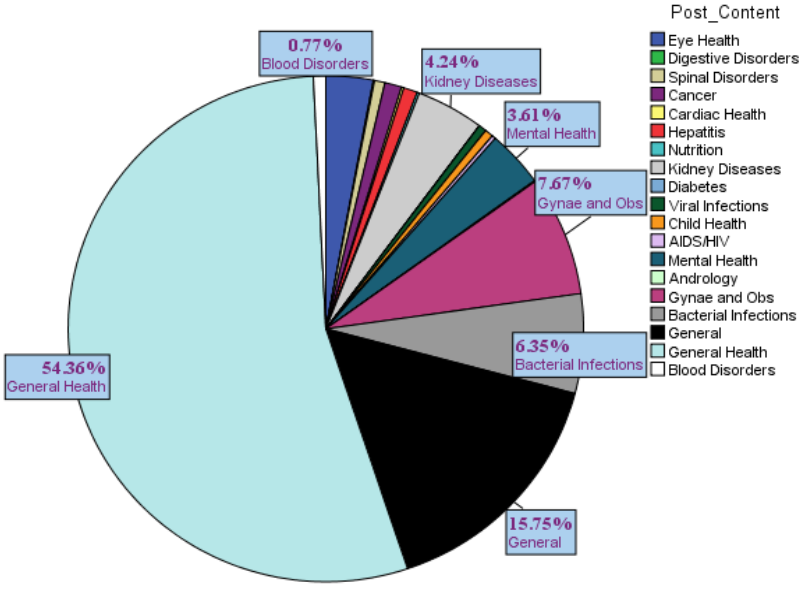


FIGURE 6.7: Post Content w.r.t. highest comments

Later on a predictive analysis for the automated dataset was also carried out with the help of classification tree. In this tree Post_Content was the dependent variable and other attributes included in the dataset were independent variables. The tree was designed according to the organization and city and some attributes were merged which had least effect on post content, CHAID was used as the growing method for this tree.

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Post_Content
	Independent Variables	Org_Type, No_of_Comments, Org, City
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
	Independent Variables Included	Org, City
Results	Number of Nodes	9
	Number of Terminal Nodes	6
	Depth	2

TABLE 6.5: Model Summary of Automated Dataset

In the tree some nodes were merged into single node having same information about the post content, the tree is shown below.

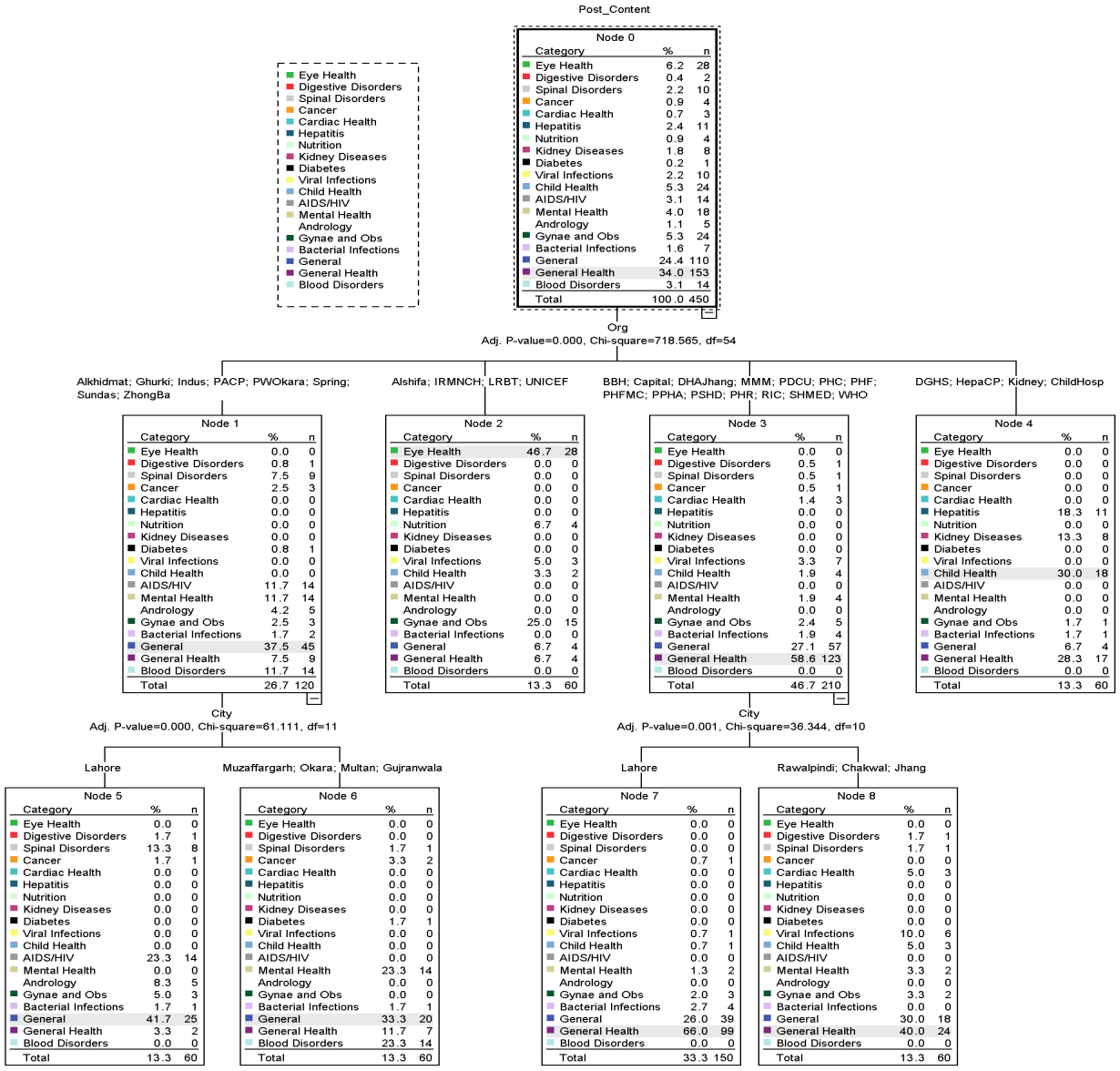


FIGURE 6.8: Classification Tree of Automated Dataset

This tree also produced a predictive value for dependent variable based on the independent variables and that predicted value was compared to the original value of post content to derive accuracy rate. The accuracy rate was near to 50%, this was because of the fact that through mining limited number of attributes were analyzed.

Moreover the results of both the datasets are different because both data sets had different number of organizations. Few organizations in both datasets were same and others were different according to their own health area. Further the changing privacy policy of Facebook has also effected the

attributes gathering and results of the research. When more attributes are gathered in lose privacy policy through mining it gives more accurate results as compared to the tight privacy policies. When policies are restricted the access to attributes gathering also gets restricted which results in less accuracy level of dataset.

7 Conclusion and Future Work

7.1 Introduction

This chapter summarizes the conclusion of the whole research process conducted. This chapter presents the concluding results after the final analysis of the study along with the factors effecting the research process. In section 7.2 it is discussed that what type of future work can be done on the basis of this work.

7.2 Conclusion

Pakistan has many healthcare issues such as most exceedingly terrible newborn child death rate in the world. Infant conceived in Pakistan faces a one in 22 chance of death, while a newborn in Japan had just a one in 1,111 risk of dying. Social media can be a beneficial tool by which human services associations may possibly circulate wide-ranging health information. Facebook might end up being an enhanced instrument for promoting health care and developing health-care professional's (HCPs) knowledge if used wisely. It can deliver substantial benefits in patients care, education, and health programs. There are various social networking sites to gather data for analysis and Facebook is one of them.

Facebook pages of healthcare organizations were utilized to mine data for the analysis of useful information. Facebook data of different health sector organizations of Punjab Pakistan was mined in two different datasets. Manual and automated datasets was preprocessed and statistical along with predictive analysis of both the datasets was accomplished. Classification tree was the machine learning technique utilized for predicting the upcoming health issues. Both datasets were explored in terms of accuracy level, manual data set gave an accuracy rate of approximately 70% while automated dataset's accuracy was nearly 50%.

The major factors observed that play vital role in this type of study were the choice of the organization selected for analysis and interaction patterns of general public on health related posts. The most impacting factor in this type of study is the varying cyber policies of social networking

sites, as the privacy policies of Facebook changes the mining techniques changes accordingly and accuracy effects because of this factor.

7.3 Future Work

After conducting this study it is concluded that some limitations of this exploration can be improved and worked upon in future. Below are few factors that can be improved in future research practices for better outcomes.

- The selection of healthcare organizations play an important role in research, number of organizations selected along with the main focus of the association should be kept in mind before study to get better results.
- A mechanism for evaluating the useful and useless public comments may give better results in future.

Bibliography

[AD14] L. Alfantoukh and A. Durresi. "Techniques for Collecting data in Social Networks." In: *2014 17th International Conference on Network-Based Information Systems*. 2014, pp. 336–341.

[DS13] D. Dumbrell and R. Steele. "Twitter and Health in the Australian Context: What Types of Information Are Health-Related Organizations Tweeting?" In: *2013 46th Hawaii International Conference on System Sciences*. 2013, pp. 2666–2675. DOI: 10.1109/HICSS.2013.578.

[Fac] Facebook. *company info*. URL: https://about.fb.com/company-info/.

[fac] facebook. *developers facebook*. URL: https://developers.facebook.com/docs/graph-api/.

[Fb] Fb. *Fb policy*. URL: https://www.facebook.com/policy.php.

[GS] GS. *social-media-stats in Pakistan*. URL: https://gs.statcounter.com/social-media-stats/all/pakistan.

[HB14] C. Hutto and C. Bell. "Social Media Gerontology: Understanding Social Media Usage among a Unique and Expanding Community of Users." In: *2014 47th Hawaii International Conference on System Sciences*. 2014, pp. 1755–1764. DOI: 10.1109/HICSS.2014.223.

[hea] healthdata. *health data in Pakistan*. URL: http://www.healthdata.org/pakistan.

[IBM] IBM. *spss*. URL: https://www.spss-tutorials.com/spss-variable-types-and-formats/.

[ibm] ibm. *spss-statistics-software*. URL: https://www.ibm.com/analytics/spss-statistics-software.

[Lal+16] Muhammad Ikram Lali, Raza Mustafa, Kashif Saleem, M. Saqib Nawaz, Tehseen Zia, and Basit Shahzad. "Finding Healthcare Issues with Search Engine Queries and Social Network Data." In: *International journal on Semantic Web and information systems* 13 [Dec. 2016].

[mic] microsoft. *Power BI*. URL: https://powerbi.microsoft.com/en-us/.

[ngo] ngos. *health ngos pakistan*. URL: http://www.ngo.com.pk/health/health_ngos_pakistan.htm.

[NML17] M. Saqib Nawaz, Raza Mustafa, and Muhammad Ikram Lali. "Role of Online Data from Search Engine and Social Media in Healthcare Informatics." In: June 2017, pp. 272–293. ISBN: 9781522526070.

[pun] punjab. *punjab pakistan*. URL: https://www.punjab.gov.pk/provincialdepartments.

[res] research. *types-of-scientific-research*. URL: https://innspub.net/types-of-scientific-research/.

[SMV17] N. Straton, R. R. Mukkamala, and R. Vatrapu. "Big social data analytics for public health: Comparative methods study and performance indicators of health care content on Facebook." In: *2017 IEEE International Conference on Big Data (Big Data)*. 2017, pp. 2772–2777.

[sta] statista. *global social networks*. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[uni] International telecommunication union. *Measuring Information Society*. URL: https://www.itu.int/en/ITUD/Statistics/Documents/publications/mis2013/MIS2013_without_Annex_4.pdf.

[USA] USA. *judiciary Changing Cyber laws*. URL: https://www.judiciary.senate.gov/meetings/facebook-social-media-privacy-and-the-use-and-abuse-of-data.

[WVC18] A. Wongkoblap, M. A. Vadillo, and V. Curcin. "A Multilevel Predictive Model for Detecting Social Network Users with Depression." In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. 2018, pp. 130–135.

[Zul] Wajid; Taimur-ul-Hassan Zulqarnain. *Social and Traditional Media Usage: A Demographic Analysis of Pakistani Youth*. URL: https://aiou.edu.pk/SAB/gmj/GMJ%20Spring%202017/02.pdf.

[Bag+14] F. Bagayogo, L. Lapointe, J. Ramaprasad, and I. Vedel. "Co-creation of Knowledge in Healthcare: A Study of Social Media Usage." In: *2014 47th Hawaii International Conference on System Sciences*. 2014, pp. 626–635.

[Ban+18] S. Bansal, P. Kumar, S. Rawat, and T. Choudhury. "Analysis and Impact of Social Media and it's Privacy on Big Data." In: *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. 2018, pp. 248–253.

[Mel+17] N. Melethadathil, B. Nair, S. Diwakar, and S. Pazhanivelu. "Assessing short-term social media marketing outreach of a healthcare organization using machine learning." In: *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. 2017, pp. 387–392.

[Tuc+17] C. S. Tucker, M. Burrows, K. Lesniak, and S. Klein. "Cybersecurity Policies and Their Impact on Dynamic Data Driven Application Systems." In: *2017 IEEE 2nd International Workshops on Foundations and Applications of Self* Systems (FAS*W)*. 2017, pp. 363–365.

Appendix

Healthcare Organizations Identified in Province Punjab Pakistan

- http://www.pshealth.punjab.gov.pk
- http://dme.pshealth.punjab.gov.pk
- http://ppha.punjab.gov.pk
- http://phfmc.punjab.gov.pk/
- http://www.phc.org.pk
- http://www.irmnch.gop.pk
- http://health.punjab.gov.pk
- http://phf.punjab.gov.pk
- https://www.facebook.com/dhajhang/
- https://www.facebook.com/pcdchf/
- http://dghs.punjab.gov.pk
- http://www.pacp.gop.pk
- http://www.health.punjab.gov.pk/Prevention_and_Control_of_Hepatitis
- https://mss.punjab.gov.pk/
- http://bbh.org.pk
- http://www.ric.gop.pk/

- http://thqshorkot.com
- http://www.aimhsialkot.com
- http://www.tbsamlihospital.com
- http://www.nid.com.pk/
- https://wic.punjab.gov.pk
- https://dhqksb.punjab.gov.pk
- https://dhqnwl.punjab.gov.pk/
- http://www.lgh.org.pk/
- https://sialkot.punjab.gov.pk/GOVT_SARDAR_BEGUM_TEACHING_HOSPITAL
- http://fatima-group.com/mast/
- http://www.appna.org
- http://www.alkhidmat.org
- http://www.lrbt.org.pk
- http://www.mmm.org.pk
- http://www.opwacare.com/
- https://www.bsfoundation.com/
- http://hands.org.pk/live/?p=5865
- http://www.sundasfoundation.org/
- http://www.indushospital.org.pk
- http://www.pkli.org.pk
- https://www.ghurkitrust.org.pk/
- http://alliedenthospital.com/
- http://alkhidmatraazi.com/rawalpindi
- http://www.alshifaeye.org/
- https://www.kktpakistan.com/kkt-orthopedic-spine-center-rawalpindi/
- http://www.reliancehospital.org/

- http://www.alizaibfoundation.org
- https://www.hamzamedicare.com
- https://www.unicef.org/pakistan/
- http://mmhbhakkar.wixsite.com/musahospital
- http://mashospital.org
- http://www.saharaforlife.org
- http://thesultanfoundation.org.pk/project/sultan-foundation-hospital-vehari/
- http://www.chenonefoundation.com/
- https://shaukatkhanum.org.pk/
- http://malc.org.pk/
- https://www.who.int/countries/pak/en/
- http://punjab.prcs.org.pk/
- https://www.jsi.com/JSIInternet/IntlHealth/where/display.cfm?tid=1030&id=249
- http://mcwap.org/
- http://www.pws-inmol.org/
- http://www.psrd.org.pk/
- http://mcwap.org/
- http://sundas.org
- http://www.thephysioint.com
- https://bahriainternationalhospital.com
- http://www.zhongbahospital.com/
- http://www.bilalhospital.com
- http://www.fatimamemorial.org.pk/
- https://www.facebook.com/SHARIFSURGIMEDHOSPITAL/
- http://www.amanathospital.com/
- http://www.shalamarhospital.org.pk/

- http://signatureskincare.pk/
- http://www.capitalhealth.com.pk
- http://shahidhospital.pk/
- http://www.winner.com.pk

Healthcare Organizations Included in Manual Dataset

- https://www.facebook.com/alshifaeye2eye
- https://www.facebook.com/BHRawalpindi/
- https://www.facebook.com/dhajhang/
- https://www.facebook.com/PDCU.Punjab/
- https://www.facebook.com/KKTPakistan/
- https://www.facebook.com/Lrbt.Pakistan/
- https://www.facebook.com/PACPOfficial/
- https://www.facebook.com/punjabhealth/
- https://www.facebook.com/PSHDepartment/
- https://www.facebook.com/sundasfoundationgujranwala/
- https://www.facebook.com/ZBHospitalInLahore/

Healthcare Organizations Included in Automated Dataset

- https://www.facebook.com/alkhidmat.org/
- https://www.facebook.com/alshifaeye2eye
- https://www.facebook.com/BBH.Rawalpindi.Official/
- https://www.facebook.com/Capitalhealth/
- https://www.facebook.com/DGHSPb/
- https://www.facebook.com/dhajhang/
- https://www.facebook.com/GhurkiTrustTeachingHospital/

- https://www.facebook.com/HEPControlPUNJAB/
- https://www.facebook.com/indushospital/
- https://www.facebook.com/irmnch/
- https://www.facebook.com/pklinrc/
- https://www.facebook.com/Lrbt.Pakistan/
- https://www.facebook.com/muslimmedicalmission/
- https://www.facebook.com/PACPOfficial/
- https://www.facebook.com/pcdchf/
- https://www.facebook.com/PDCU.Punjab/
- https://www.facebook.com/PunjabHealthcareCommission/
- https://www.facebook.com/PunjabHealthFoundation/
- https://www.facebook.com/punjabhealth/
- https://www.facebook.com/PunjabPublicHealth
- https://www.facebook.com/PSHDepartment/
- https://www.facebook.com/PunjabHealthReforms/
- https://www.facebook.com/pwadhqokara/
- https://www.facebook.com/RICrwp/
- https://www.facebook.com/SpecializedHealthcare/
- https://www.facebook.com/springclinicmultan/
- https://www.facebook.com/sundasfoundationofficial/
- https://www.facebook.com/unicefpakistan/
- https://www.facebook.com/WHO/
- https://www.facebook.com/ZBHospitalInLahore/