The Pennsylvania State University

The Graduate School

Department of Industrial and Manufacturing Engineering

LEGITIMATE OR NOT – LEARNING THE STATUS OF ONLINE PHARMACIES

A Thesis in

Industrial Engineering

by

Sowmyasri Muthupandi

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

August 2016

The thesis of Sowmyasri Muthupandi was reviewed and approved* by the following:

Soundar Kumara
Allen E Pearce And Allen M Pearce Professor of Industrial and Manufacturing Engineering
Thesis Co – Advisor

Hui Zhao
Associate Professor of Supply Chain Management
Thesis Co – Advisor

Janis Terpenny
Peter and Angela Dal Pezzo Professor
Head of the Department of Industrial and Manufacturing Engineering

*Signatures are on file in the Graduate School

ABSTRACT

Previous research has indicated the presence of illegitimate online pharmacies on the World Wide Web, which may be involved in sales of counterfeit or substandard products which have the potential for drug abuse. This thesis focuses on studying the relative usage of legitimate and illegitimate pharmacies and developing as automated classification system by mining web data. The list of safe and rogue pharmacies are identified from National Association of Board of Pharmacies (NABP) and web data is obtained using Similarweb and SEMrush. The referral data is used to develop the classification model. Along with an intuitive algorithm, Rating Method (RM), K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Logistic Regression (LR) and Support Vector Machines (SVM) models are developed and validates using leave-one-out cross validation on a dataset with 157 samples – 30 legitimate and 157 illegitimate pharmacies. RM had better accuracy (95.42%), kappa (0.8635), and specificity (96.67%) compared to other models. KNN had the highest sensitivity (99.2%) and second highest accuracy (92.2%) and kappa (0.715). KNN and RM were implemented on a larger dataset with 50 legitimate and 1136 illegitimate pharmacies. It was observed that KNN performed better than RM on the larger dataset with accuracy, kappa, sensitivity and specificity values 98.73%, 0.8085, 100%, 68.75% respectively. This thesis takes into account the static data, a snapshot of the World Wide Web to classify online pharmacies. Future research can consider dynamic data, changes in referring websites of the online pharmacies to improving the classification. Also, products sold in the pharmacies and the anomalies can be used to increase the granularity of the classification model.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Soundar Kumara, Professor at The Department of Industrial and Manufacturing Engineering, for his valuable guidance, great insights and support during the course of my Master's Thesis.

I would like to express my gratitude to Dr. Hui Zhao, Associate Professor of Supply Chain Management in Smeal College of Business, for her constant supervision and for providing me motivation and resources to conduct research.

I also thank Dr. Janis Terpenny, Head of the Department at The Department of Industrial and Manufacturing Engineering, for being a part of my thesis committee.

I take this opportunity to thank Pennsylvania State University for providing me an opportunity to pursue my research interests. Finally, I express my profound gratitude to my parents Muthupandi Veerappan and Santhi Muthupandi and my brother Gokul Muthupandi for their continuous encouragement and unfailing support throughout my graduate studies.

Chapter 1

Introduction

Traditionally, 'street market' which provides heroine and crack cocaine, run by hierarchical organized crime and 'free market' which involves selling of cannabis and ecstasy amongst small group of friends are identified as the illicit drug markets. However, the illicit ecommerce market is growing rapidly with the advancement of technology (Schifano et al., 2003). Nearly 2.9 billion people, almost 37% of world population used Internet in 2013 (Mackey&Liang, 2013). Among the internet uses in US, according to Pew internet survey, 75% look up for health and medical related information. Additionally, 1/3 of them engage in self-diagnosing their health problems. At the same time, Internet sales of drugs have increased over the years. The online prescription drug sales in 2003 was estimated to be $3.2 billion which was 20-fold higher than sales in 1999 (Jena et al, 2011). In year 2007, 14.2% of internet users bought at least one medicine or vitamin online. Almost 85% of online buyers had a regular provider (Desai et al, 2015). The outlets on internet that vend medicines are called online pharmacies.

Online pharmacies maybe legitimate pharmacies, subscription pharmacies, lifestyle pharmacies or no-prescription pharmacies. Legitimate pharmacies vend drugs to the customers only if a valid medical prescription is provided. Lifestyle pharmacies trade a limited number of drugs based on 'online consultation'. Subscription pharmacies provides access to other online pharmacies that sell drugs without prescription for a subscription fee paid through credit card. Finally, as the name suggests, no-prescription pharmacies offer to mail controlled drugs to patients without prescription (Littlejohn et al, 2005).

Lifestyle pharmacies which sell prescription drugs, subscription pharmacies and no-prescription pharmacies can be termed as illegitimate, rogue or illegal pharmacies. An online pharmacy is deemed illicit if any of the following conditions are true (Legitscript, 2016)

1. The online pharmacy sells prescription medication without any valid prescription
2. The online pharmacy vends drugs that are deemed to be unapproved under applicable law
3. The website facilitates the trade of drugs by pharmacies or other entities that do not have licenses to sell medication under applicable laws

Past studies suggest that illegitimate pharmacies are widely accessed and is a potential cause for drug abuse. We focus on understanding the relative usage of the legitimate and illegitimate pharmacies using web analytics. Additionally, we aim to study how the illegitimate and legitimate pharmacies are different from each other and building a model which classifies the online pharmacies. The motivation for the given study, the methodology adopted and the significant findings are summarized in the following sections.

1.1 Motivation

The sale of medication online has grown in a very short time along with the popularity and turnover of online pharmacies. In 2009, these pharmacies catered to about 8% of the European market and 10% of the international market (Letkiewicz & Goriski, 2009). The legitimacy of the websites that sell drugs online are debatable and the medicines purchased from the illegal pharmacies might be expired, counterfeit or inappropriate (Romero et al. 2004; Lineberry and Bostwick 2004; Fox et al. 2005; Gondim and Falcão 2007; Mackey et al 2013).

Illegal online pharmacies dispense medication without any authorized communication between the patient and the physician. The US drug enforcement Agency and General Accounting Office (GAO) conducted secret-shopper studies and identified that 8 illegitimate pharmacy shipped hydrocodone to patients without any prescription (Jena et al, 2011). Eventually, the Federal Trade Commission exposed 800 illegitimate online pharmacies offering products of questionable quality (Montoya and Jano 2007).

According to a national survey in Europe, 11 out of every 100 prescription drug abuse complaint involves purchase of medication from online pharmacies (Inciardi et al, 2010; Gordon et al, 2006). A middle-aged American, who had a family history of heart disease, died of heart attack caused by Viagra which he bought through online consultation (Mäkinen et al. 2005). Puri and Damle (2007) informed a women's demise due to consumption of counterfeit medicine bought online (Puri & Damle 2007).

Lanier (2004) and Weiss et al (2007) reported events of purchase of prescription medication from online pharmacies which led to suicide (Lanier 2004; Weiss et al. 2007). In addition, local dealers are purchasing drugs online and are selling them through brick-and-mortar stores which also poses major threat of drug abuse (Jena et al, 2011). Illegitimate pharmacies have international operations and participate in trade of biologically active compounds. In future, illegitimate online pharmacies has the potential to become a vehicle of bioterrorism (Letkiewicz & Goriski, 2009).

Several verification system for online pharmacies exists. Verified Internet Pharmacy Practice Site (VIPPS) verification seal provided by National Association Board of Pharmacies (NABP) is recommended by US Food Drug and Administration and Legitscript. Other verification systems include Canadian International Pharmacy Association and Pharmacychecker. However, they are not recommended by any governmental organizations (Fittler et al, 2013).

Consumers can verify the legitimacy of the online pharmacies by looking up for the specific domain in NABP recommended and not-recommended list of pharmacies or Legitscript database. However, NABP has reviewed only a small number of sample sites and its VIPPS accreditation is subject to application and annual participation fees. Even though Legitscript has the largest database of internet pharmacies (Fittler et al, 2013), it is not exhaustive. Also, one can check the legitimacy of only five pharmacies from a unique IP address every 24 hours from their database. They doesn't allow access to entire database without subscription. Currently, the sources for consumer education on legitimacy of online pharmacies have their own limitations.

According to Legitscript, globally about 20 new illicit online pharmacies are created every day. It is hard to record and monitor all these online pharmacies. Ineffective consumer awareness systems and absence of an exhaustive database on online pharmacies demand development of automatic online pharmacy classification system which can be used at the consumer end to make informed decisions which can avoid drug abuse.

1.2 Problem definition

We aim to develop an online pharmacy classification system which can state if the pharmacy with the given domain name is legitimate or not by studying the patterns in web data of these pharmacies. The problem is defined as

Given a set of N online pharmacies, the objective is to find the subsets N_1 and N_2 where,

$$N_1 \cup N_2 = N$$

Such that we classify N_1 as belonging to legitimate online pharmacies and N_2 as belonging to illegitimate online pharmacies. The problem is to learn the classifier which accomplishes the above objective.

1.3 Brief description of methodology

Previous research on online pharmacies were explored to see if they studied the characteristics of online pharmacies, distinguishing features between legitimate and illegitimate online pharmacies and classification of online pharmacies. It has been identified that analysis of web data of online pharmacies and classification of them are still nascent. In this study, we adopt web mining to study the relative usage of online pharmacies to analyze the extent to which they are accessed, identify distinguishing features between legitimate and illegitimate online pharmacies and develop a framework for automatic classification of them.

The legitimate and illegitimate pharmacies for the given study are identified from NABP list of recommended and not recommended sites. For identified online pharmacies, usage data and referral data were collected from SEMrush and Similarweb. The various datasets identified, their sources and their description are given in Table 1-1.

Table 1-1. Online pharmacy datasets

Dataset name	**Definition**
Traffic sources data	The percentage of traffic contributed from different sources to the online pharmacy
Engagement data	The values of metrics that measures the involvement of users with online pharmacies
Country data	The percentage of traffic entering the online pharmacies from different countries
Social media data	The percentage of traffic entering the online pharmacies from different social media websites
Referral data	The various websites that direct traffic to online pharmacies and its characteristics
Search data	The percentage of traffic from paid and organic search

Exploratory data analysis was performed on the collected data through descriptive analytics and data visualization. Exploratory data analysis suggested that the major traffic sources for legitimate and illegitimate pharmacies are direct, search and referrals. The biggest consumer for both legitimate and illegitimate pharmacies is United States. A small proportion of illegitimate pharmacies attracted traffic from paid search and illegitimate pharmacies attracted higher traffic through display ads compared to legitimate pharmacies. Facebook, Reddit and YouTube are top among the social media website that direct traffic to illegitimate pharmacies. Finally, the overall engagement of legitimate pharmacies is better that illegitimate pharmacies. In addition, exploratory data analysis showed that the number of backlinks from different websites connecting the online pharmacies will be effective for classification than engagement data.

The dataset used to develop the binary classifier included 153 samples, 30 legitimate pharmacies and 123 illegitimate pharmacies, and 9685 referral websites connected to it, that is 9685 attributes. Linear classifiers – Linear Discriminant Analysis, Logistic Regression and Support Vector Machines, were adopted for developing the online pharmacy classification model. Principal Component Analysis was used for dimensionality reduction and features which explain 99% variance are selected. Leave-one-out cross validation was used for validating the model as the sample size is small.

For the given problem, the error of classifying a rogue pharmacy as safe is costlier than the error of classifying a safe pharmacy as rogue. Considering rogue pharmacy to be the positive class, it is important to maximize the sensitivity of the model. Support Vector Machine facilitates cost sensitive learning which allows assigning weights to the type I and type II error in the cost function. The performance of Support Vector Machines across different cost values was also studied.

Along with the linear classification models, an intuitive rating method is proposed. It works on the hypothesis – if a referral website directs traffic to safe pharmacies then a there is a high probability that a new online pharmacy to which it directs traffic is safe. The rating method was trained and tested on the same data set using leave one out cross validation.

The best performing models are shortlisted from the sample set and they are used on a bigger dataset with 1187 pharmacies – 1136 illegitimate and 50 legitimate, to check the generalizability of the results. The performance of the models on the bigger dataset is evaluated using hold out cross validation. Two third of the sample is used for training and one third of the sample is used for testing.

1.4 Results

Seven linear classification models and KNN were developed and tested for classifying online pharmacies using the referral dataset I with 153 samples. The linear classification models includes LDA (Linear Discriminant Analysis), LR (Logistic Regression), SVM c=1 (Support Vector Machines with cost 1), SVM c=0.5, SVM c=0.3, SVM c=0.1 and RM (rating method). KNN with values of K from 1 to 9 were implemented and tested. RM performed better than all the other models. It had 95.42% accuracy, 95.12% sensitivity and 96.67% specificity. 1-NN had accuracy, sensitivity and specificity values of 92.16%, 99.19% and 63.33%. LDA performed better than LR. SVM c=1, 2-NN and LDA had the same accuracy (86.93%) but 2-NN has better sensitivity. The sensitivity increase with decrease in the cost parameter of SVM. However, the accuracy and specificity also decreased. Similarly, with increase in K accuracy and specificity of the model decrease and sensitivity of the model increased.

From the results it is evident that RM and KNN performed better on the given dataset. Hence, RM and KNN were implemented on a larger dataset, referral dataset II, with 1186 samples – 50 legitimate and 1136 illegitimate pharmacies. 1-NN and 2-NN performed the better compared to RM for the bigger dataset. They had 98.73% accuracy, 100% sensitivity and 68.75% specificity. RM had accuracy, sensitivity and specificity of 96.19%, 98.942% and 31.25% respectively.

1.5 Organization of Thesis

The construction of this Thesis is as follows. Chapter 2 describes the previous work and research gap. Data collection and Data description is described chapter 3. Chapter 4 includes the analysis and chapter 5 explains the results. Conclusions and scope for future work are discussed in chapter 6.

Chapter 2

Background literature

Online purchases are growing rapidly with internet usage. Drugs are becoming a popular merchandise in online commerce. However, to promote and sell drugs online certain criterion must be met by the online vendors. When they don't comply with any of those criteria they are deemed as illegitimate. Illegitimate pharmacies might be involved in selling counterfeit substances and might be a potential source for drug abuse (Mackey et al 2013). Over the past decade, illegitimate online pharmacies and drug abuse has become a growing concern. Several organizations such as Legitscript, National Association Board of Pharmacies (NABP) and Center for Safe Internet Pharmacies (CSIP) try to curb the detrimental effects of illegitimate pharmacies through consumer education and collaboration with major players in the industry. Only a NABP and Legitscript has a list of online pharmacies with are legitimate or illegitimate. However, it is not an exhaustive list as numerous online pharmacies are added every day in the World Wide Web. It is crucial to develop an automated system which identifies these online pharmacies so that it can be used by search engines and other websites to check traffic from them into illegal online pharmacies. Such a framework, which exploits the web mining and linear classifiers, for identifying illegitimate online pharmacies is proposed in this thesis. In this chapter, past studies on online pharmacies are explored followed by the introduction of basic concepts of web mining.

2.1 Research on online pharmacies

Online pharmacies, internet pharmacies or mail-order pharmacies, are pharmacies in which customers can buy medication through internet and their orders are supplied through mail. Online pharmacies that sell prescription drugs can be classified into four types – legitimate pharmacies, subscription pharmacies, lifestyle pharmacies and no-prescription pharmacies. Legitimate pharmacies mail the drugs directly to the customers only if a valid medical prescription is provided. Lifestyle pharmacies provide customers a limited number of drugs directly to the patient. The website generates prescription based on 'online consultation'. On the other hand, subscription pharmacies promise access to other websites that sell drugs without prescription for a subscription fee paid through credit card. Finally, as the name suggests, no-prescription pharmacies offer to mail controlled drugs to patients without prescription (Littlejohn et al, 2005).

Though, online pharmacies benefit consumers by providing drugs at lower cost, serious issues are raised due to the wide presence of illicit online pharmacies which dispense drugs without any direct physician examination or a valid prescription. Illicit online pharmacies may be involved in sale of counterfeit or substandard products which might contain no active pharmaceutical content and/or toxic agents that have been associated with patient deaths (Mackey et al 2013). Jena & Goldman (2011) empirically proved that increase in internet usage is correlated to drug abuse. They showed that every 10% increase of internet usage is associated with 1% increase in prescription drug abuse in United States (Jena & Goldman, 2011). Study by Jena & Goldman provides evidence that illegal internet pharmacies are partly responsible for drug abuse.

Illegitimate pharmacies attract consumers through price and supply. They sell drugs at low price and also sell drugs on shortage and recall. This motivates customers to purchase drugs from illegitimate pharmacies. In this section we discuss research on online pharmacies based price of drugs, availability – shortage and recall and finally we discuss the studies on web mining and identifying characteristics of online pharmacies.

2.1.1 Online pharmacies and price of drugs

Illegitimate pharmacies promote no prescription sale of drugs at a very low price. Quon, et al (2005) aimed to compare the price of 44 retail brand-name drugs across Canadian internet pharmacies and US drug chain pharmacies. They showed that Americans can save a mean of approximately 24% per unit of drug, if they purchased them from Canadian pharmacies instead of US drug chain pharmacies. Also, 41 out of 44 drugs were less expensive in Canadian pharmacies than US pharmacies. It is important to note that Quon et al used the list of Canadian pharmacies from PharmacyChecker.com which does not have a reliable source for safe pharmacies. So we infer that this study shows illegal pharmacies sell products at a cheaper price.

Also, illegitimate pharmacies vend the top grossing drugs and promote them online. Since they offer them at a lower price, consumers might tend to purchase these drugs from rogue pharmacies as cost is one of the major criteria in making a purchase decision. Liang&Mackey (2011) studied the prevalence of eDTCA (online Direct to Consumer Advertising) of top pharmaceutical companies and top grossing drugs on social media sites. They found that illicit online pharmacies advertised for no prescription sales for 9 out of 10 of the top grossing drugs on social media. Also, they identified that 16 out of the top 20 globally marketed DTCA drugs had online marketing by illegal pharmacies. They indicated that 60% of the top 10 drugs and 50% of the top 20 drugs had presence of online marketing by illicit pharmacies on Facebook (Liang&Mackey, 2011).

2.1.2 Online pharmacies and shortage drugs

Along with price, availability of drugs also impact its illicit online presence. Liang&Mackey (2012) assessed the online marketing of FDA shortage drugs. They estimated that out of 72 FDA shortage drugs 94% were available online. 68% of these online drug sellers were present on the not recommended list provided by National Association Boards of Pharmacy (NABP). Also, among shortage drugs with presence of online sales 91% of the drugs were sold by at least one of the websites on the not recommended list and 36% were sold only by websites on the not recommended list (Liang&Mackey, 2012).

2.1.3 Online pharmacies and drugs on recall

Drug recall is the critical function of Drug Regulatory Authorities (DRA) around the world. Drugs maybe recalled permanently or temporarily from the market to reduce potential risk to human health. Despite clear safety risks such drugs sold and marketed online outside the oversight of DRA. Drug recalls can lead to drug shortages and limit the options of treatment to a disease which might motivate patients to buy recalled drugs online. Mackey et al, (2013) studied the accessibility of recalled drugs online. They found that 50% of the permanently recalled drugs are available for sales online. Also, they indicated that among the recalled drugs available online, 50% of the drugs had presence on Twitter and only 18.8% had content on Facebook. Additionally, these drugs were available on business-to-business websites on wholesale basis as active pharmaceutical ingredient. However, they didn’t capture the regions from which these illicit online websites attracted traffic. The genuineness of the claim that these websites sells the drug was also not tested (Mackey et al, 2013).

2.1.4 Web analytics and characteristics of online pharmacies

Even though the above mentioned works provided evidence of online availability of drugs through illicit pharmacies they have not captured the traffic estimates or the relative usage of these websites. Mackey&Liang, (2013) tried to evaluate the accessibility to an illicit online pharmacy through online advertisement and estimate the traffic captured by these advertisements. They created a fictitious advertisement stating no prescription sales of drugs on Facebook, Twitter, Google+ and Myspace. These advertisements were linked to a website to study the traffic associated with it. Over 2795 visits were registered in 10 months originating from a number of countries including high income, middle income countries and emerging markets (Mackey&Liang, 2013).

However an important issue relates to identifying attributes that help in classifying legitimate and illegitimate pharmacies. It is crucial to study the characteristics of the illicit pharmacies as it provides the basic information for any system that aims to distinguish legitimate pharmacies from rogue pharmacies. Fittler et al (2013) aimed to identify the indicators of professional pharmacies by evaluating 136 pharmacies based on longevity, time of continuous operation, geographical location, display of contact information, medical information exchange, prescription requirement and pharmacy legitimacy verification. Among the 136 identified internet pharmacies 60 were defined as rogue pharmacies, 1 was yet to be verified, 23 were unapproved and 52 were not available according to Legitscript database. They observed the operations of 136 internet pharmacies and noted that only 56 pharmacies were continuously operational. 59 internet pharmacies displayed all the necessary contact information on the website. However, for most websites the declared physical location was not same as the area of domain registration. In addition only 9 (15.25%) pharmacies requested medical prescription before purchase. They identified that prescription requirement or availability of contact information does not correlate with rogue pharmacy status as indicated by LegitScript database. However, they figured that long term

continuous operation of the website has a strong correlation with illegal activities (Fittler et al, 2013).

2.1.6 Research gaps

Previous research has shown that there is wide presences of illicit pharmacies in web including social media. However, the extent to which these pharmacies are accessed and the means by which these websites attract traffic is still not explained by the research. Also, the relative usage of legitimate and illegitimate pharmacies are not studied. The perceptive of web mining, usage mining and structure mining, of online pharmacies has not been adopted. Additionally, an automated online pharmacy classification framework has not been proposed. This thesis aims to bridge these gaps mentioned above. It focuses on comparing the web data of legitimate pharmacies and illegitimate pharmacies and identify the distinguishing features of illegitimate online pharmacies. It focuses on studying the relative usage of online pharmacies through web usage mining. A model which classifies the online pharmacies as legitimate and illegitimate exploiting the website network is proposed and implemented. Before we undertake the analysis in the following sections we provide some basics regarding web mining.

2.2 Web mining – an introduction

Web mining is adopting data mining techniques to automatically discover and explore data from web services and documents (Etzioni, 1996). Web mining is classified as web content mining, web usage mining and web structure mining based in the part of web studied. Web content mining involves exploring web content data which might be textual, image, audio, video, metadata and hyperlinks (Kosala & Blockeel, 2000). Web structure mining models the link structures of the web using the topology of the hyperlinks with or without their information (Chakrabarti et al, 1999). Web usage mining studies the data generated by web surfers' sessions (Cooley et al, 1997).

In this study, we adopt web usage mining to study the relative usage of online pharmacies and web structure mining to develop a classification model. In this section we discuss the basics of web usage mining and web structure mining in this section. Also, we discuss the previous works on website classification.

2.2.1 Web usage mining

Web usage mining explores data from Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data as the results of interactions (Kosala & Blockeel, 2000). Web analytics is a function of web usage mining used by the websites to optimize their website. Web Analytics Association defines web analytics as "the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage" (Waisberg&Kaushik, 2005). It provides a summary of metrics and various traffic sources. The basic metrics provided by google analytics and their definitions are given in Table 2-1. The different traffic sources given by google analytics and their explanations are given in Table 2-2. In this study we adopt web analytics and explore the various metrics explained to study the relative usage of online pharmacies.

Table 2-1 Google analytics engagement metrics

Metric	**Definition**
Visits	The number of times some has visited or interacted with the given website
Bounce rate	The number of visits in which the user has viewed only one page
Page views number	The number of pages viewed in all visits
Pages visits	The average number of pages viewed in each visit
Average time on site	The average time spent by a user on the site
% new visits	The percentage of visits made by users who visited the website for the first time

Table 2-2 Google analytics traffic sources

Traffic source	Explanation
Direct traffic	It represents the traffic by visitors to the website by directly entering the url or entering from a bookmark. This gives a good estimate of the number of users who have deep knowledge about the website who remembers the website or has bookmarked it.
Referring urls/referrals	Referral websites are the urls linked to the given website. Referral traffic could be result of banner ads, campaigns or from blogs interested on the given website.
Search engine	The traffic entering from search engine. Search engine is the vehicle utilized by surfers to find their destinations. This includes both paid and search traffic.
Others	This includes traffic from ad campaigns and emails

2.2.2 Web structure mining

Web structure mining models the structure of the web and is used to categorize web pages and study the relationships and similarity between websites. It can be used to find the authorities sites also called as authorities and the sites which direct traffic to the authorities called as hubs (Kosala & Blockeel, 2000). Web structure mining models the web into a graph with websites as nodes and the links between the websites as links. It is widely used for categorization or classification and clustering of webpages. Many web page ranking algorithms exploit web structure mining.

2.2.3 Classification of websites

Researchers have tried to classify different websites into certain categories. There are two types of approaches adopted – content based classification or structure based classification. Hybrid methods which include both content and structure also exists.

Content based classification utilizes the web site content to classify the website whereas, structure based classification exploits the patterns in the structure of the website. Pierre (2001) proposed a content based classification approach. He analyzed the HTML tags of the websites and classified them based on the industry.

There is a close relationship between the websites' link structure and its functionality. Amitey, et al (2003) used websites structural information to classify them into 8 classes which include corporate sites, search engines. E-store and so forth. They argue that the functionality of the website is reflected is a set of structural and connectivity based features. They used decision tree and navies Bayes classification algorithms to classify the 202 websites. The accuracy if the classifiers was between 54.5% and 59%. The precision of certain classes exceeded 85%. Lindemann & Littig (2006) classified around 1400 websites into 5 categories namely academic, blog, corporate, personal and shop. They also exploited the structure of the website to classify them. They adopted Navies Bayes classification algorithm and achieved 82% precision and 80% recall from their model.

In this thesis we study usage data and structure data of the website. Usage data is not generally used to classify websites. We explore the feasibility of adopting usage data and structure data to classify online pharmacies. From review on classification of websites it can be observed that the models which adopted structure mining used Navies Bayes classifier which is a linear classification model. Hence we use various linear models for supervised learning for classification of online pharmacies. In addition Nearest Neighbor based classification is also trained and tested. The web analytics of online pharmacies are collected using the web analytics tools – Similarweb and SEMrush. The relative usage of legitimate and illegitimate pharmacies are studied through usage mining and classification model of online pharmacies is learnt using structure data. They are discussed in detail in the following chapters.

Chapter 3

Data

The characteristics of any website can be analyzed through web usage mining, web structure mining and web content mining. Web usage mining involves exploring and learning the data web surfer's behavior or sessions (Cooley et al, 1997). Web structure mining tries to model the linked structure of the web (Chakrabarti et al, 1999). Discovery of useful patterns from web documents, data or content is deemed as web content mining (Kosala & Blockeel, 2000). The research community's efforts to explore usage mining and structure mining of online pharmacies is still nascent. In this thesis we try to study the usage data and association of online pharmacies with other websites. We aim to identify distinguishing patterns for legitimate and illegitimate pharmacies and build a classifier.

To study and explore online pharmacies' web data, a representative sample with 30 legitimate and 157 illegitimate pharmacies are selected from National Association of Boards of Pharmacy. These pharmacies are selected based on their traffic and popularity. The usage data for these pharmacies are collected from Similarweb. The referring websites and keywords which contribute to traffic towards these pharmacies are identified using Semrush. The data collection process and data description for various datasets used in the study are discussed in detail in this chapter.

3.1 Data collection

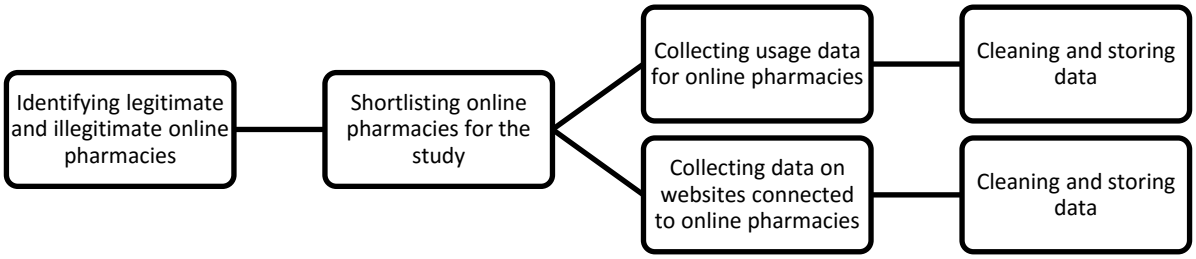

Figure 3-1 Steps involved in data collection

Figure 3-1 depicts the three phase data collection process adopted in this work. Phase one includes identification of the list of online pharmacies both legitimate and illegitimate. From the identified list a group of legitimate and illegitimate pharmacies are shortlisted during phase two. Phase three deals with collection of website traffic, referral websites and related organic keywords for the list of online pharmacies. Finally, the collected data is cleaned and stored in the format convenient for the study.

National Association of Boards of Pharmacy and Legitscript are the two organizations which has a list of legitimate and illegitimate online pharmacies. NABP offers the list on its website. On the other hand Legitscript doesn't disclose its data. Hence we chose to follow the list given by NABP. The list of legitimate and illegitimate online pharmacies is obtained from National Association of Boards of Pharmacy (NABP) website.

NABP is an international organization which includes all 50 United States, Australia, nine Canadian provinces, New Zealand and other countries which helps in developing and implementing uniform pharmacy regulations. The Verified Internet Pharmacy Practice Sites (VIPPS) program was initiated by NABP to accredit online pharmacies by verifying if they comply with the license requirements of different countries (NABP, 2016). Figure 3-2 shows the screenshot of recommended and not recommended sites list on NABP website.

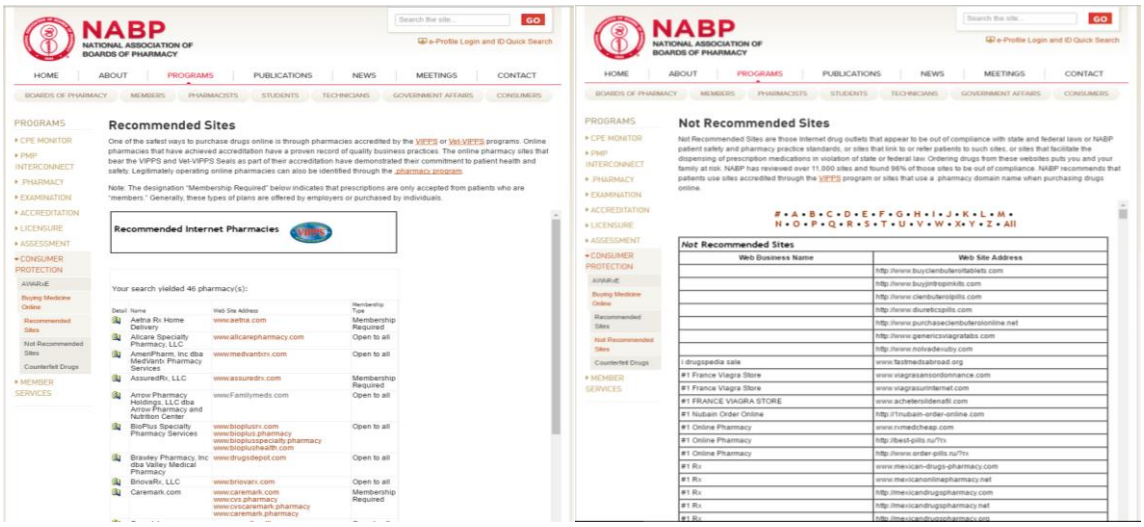

Figure 3-2. NABP list of recommended and not recommended sites web page

The NABP website provides around 35,000 illegitimate pharmacies and about 50 legitimate pharmacies. However, among the illegitimate pharmacies many are not operating right now. It is important to filter out these pharmacies and consider the ones that are alive. Similarweb provides web data for various websites.

Similarweb may not have data for a particular domain if the website doesn't exist, if the website is not a part of its database or the traffic of the website is very low to monitor. The online pharmacies for which the traffic data was available from Similarweb was taken into consideration for the given study. 139 illegitimate pharmacies and 30 legitimate pharmacies were shortlisted in phase two.

Usage data for online pharmacies are collected through Similarweb and SEMrush. Total traffic and engagement of the website according to Similarweb is taken for the study. Also the percentage of traffic contributed by each source – direct, search, referral, social media, display ads and email are collected. Different social media sites and the percentage of traffic they contribute to the online pharmacies are noted. Additionally, percentage of traffic from different countries are identified from SEMrush. The organic keywords and the referral domains data along with the number of backlinks for a particular website are obtained from SEMrush.

Data from Similarweb was collected through web scraping using R. The web crawler fetches the traffic overview for the given list of websites from Similarweb. Later, all the data is cleaned and compiled into different datasets. Figure 3-3 shows the engagement data as presented in Similarweb. Engagement data provides information about the user behavior on an average in the last six months. Country wise traffic data, represented in figure 3-4, provides the top 5 countries and the percentage of traffic from these countries from the given websites.

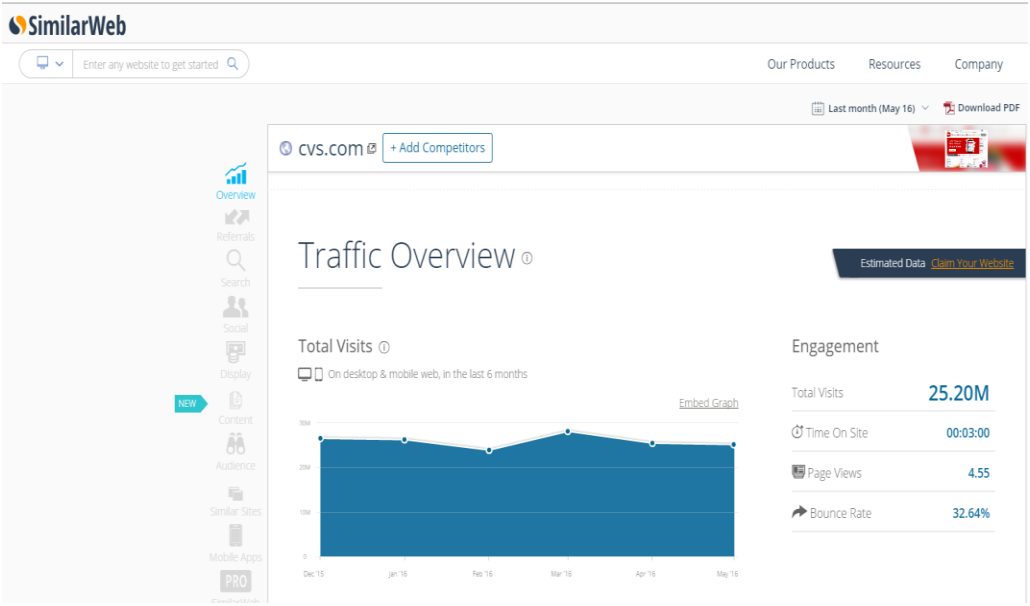

Figure 3-3 Similarweb engagement data

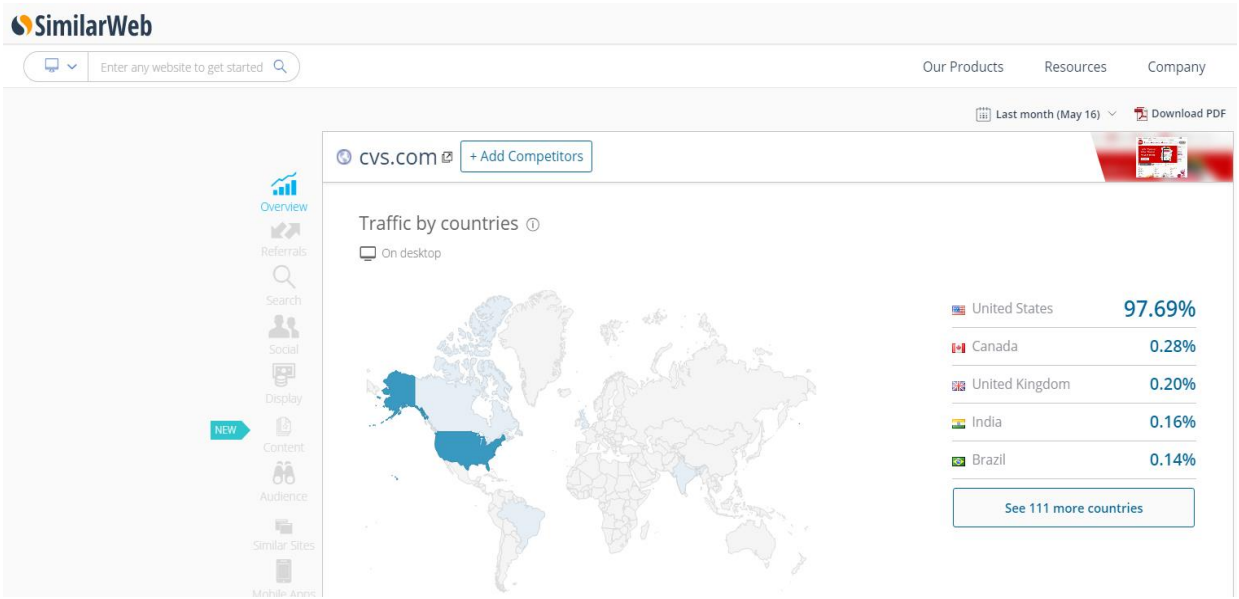

Figure 3-4. Similarweb country wise traffic data

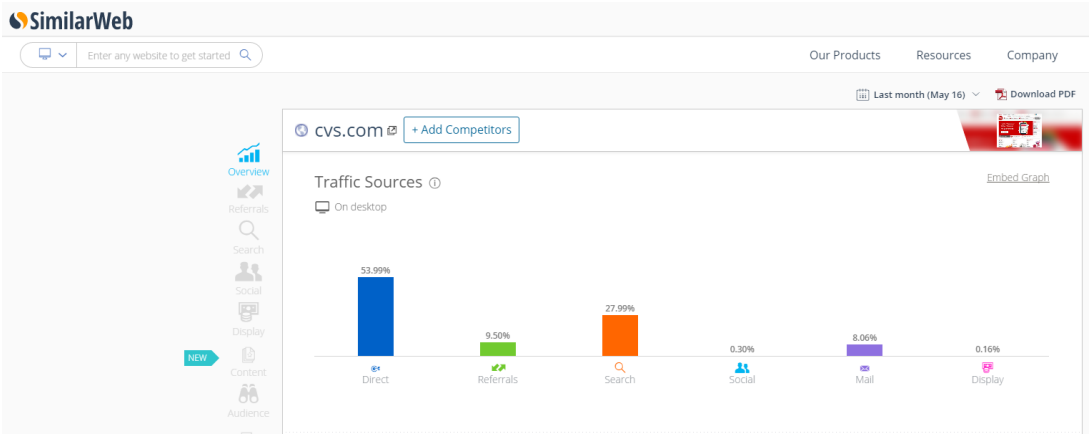

Figure 3-5. Similarweb traffic sources data

The percentage of traffic from each source from Similarweb is given in Figure 3-5. Search traffic is a cumulative of paid search and organic search. The percentage of organic search and paid search traffic is also explained by Similarweb. Figure 3-6 shows the Search traffic data. Additionally, the percentage of social media traffic entering a website from various social media is given. Similarweb provides percentage of social media traffic of the five top traffic contributors as represented in Figure 3-7.

Figure 3-6. Similarweb search data

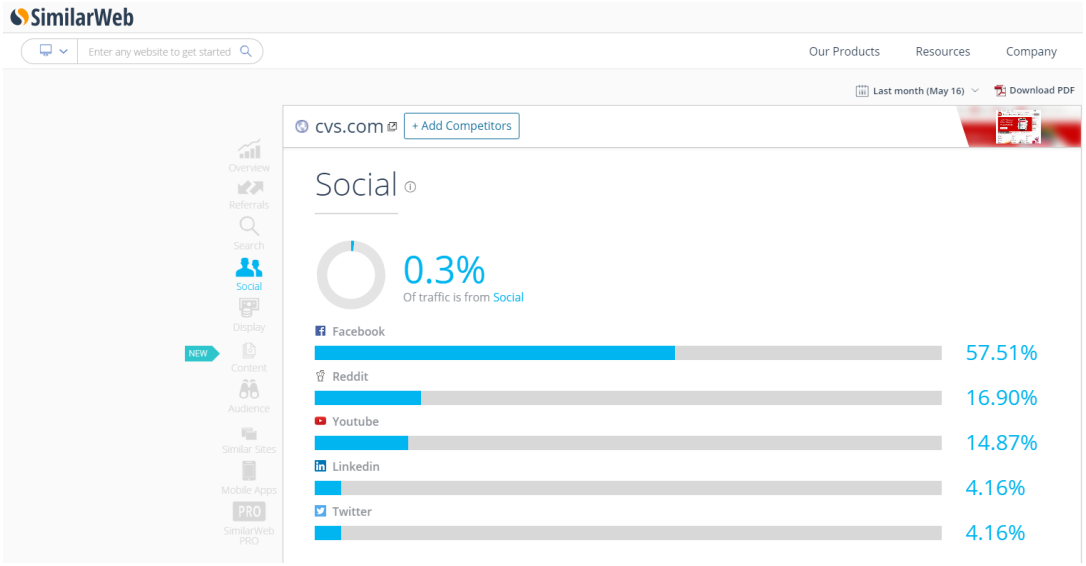

Figure 3-7. Similarweb social media data

SEMrush is used to collect traffic contribution from organic and paid keywords and data about the referring domains of online pharmacies. SEMrush source prevents web crawling and hence all the mentioned data was collected manually by typing in the domain name and downloading it into a csv file. Figure 3-8 shows the keywords data available in SEMrush. It provides the search traffic for a given domain from google. It also provides the percentage of traffic entering the website by a particular keyword. SEMrush provides the data on various referring domains to a given website as shown in Figure 3-9.

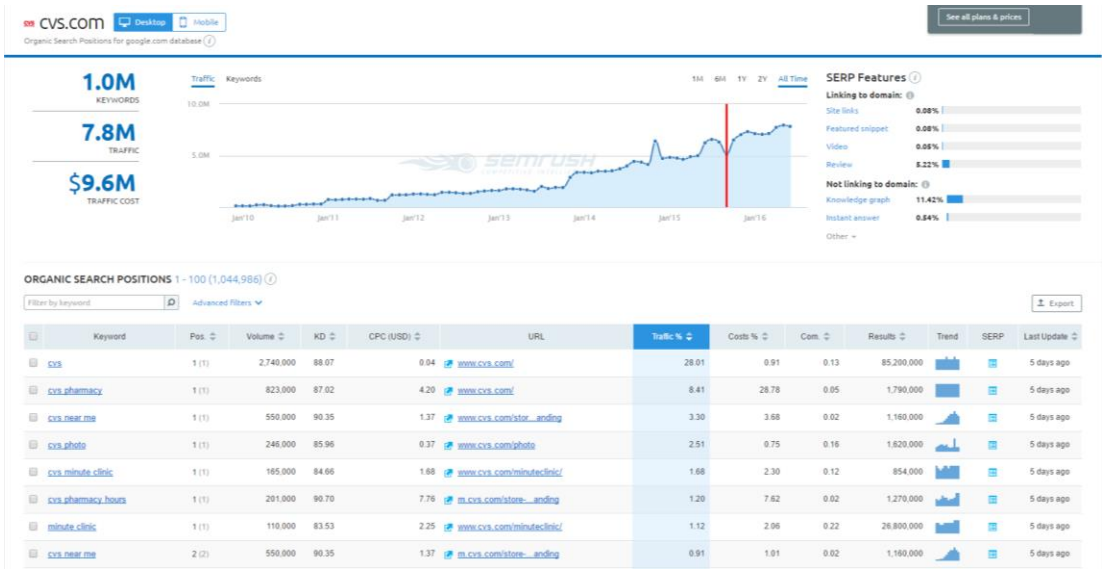

Figure 3-8. SEMrush keywords data

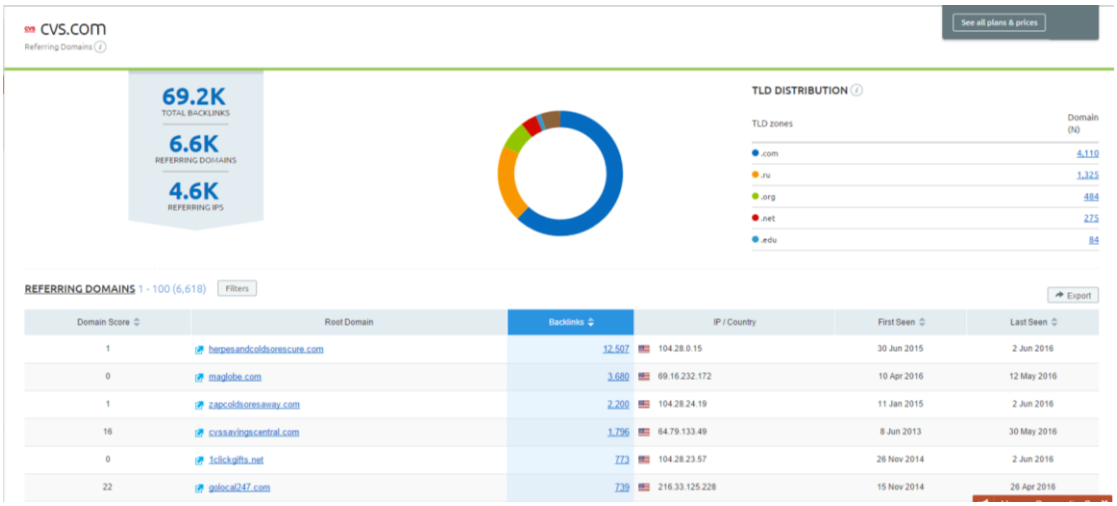

Figure 3-9. SEMrush referral data

3.2 Data description

The usage data and structure data for the list of online pharmacies shortlisted from NABP is collected using Similarweb and SEMrush. The data is cleaned and stored as different datasets. The various datasets consolidated and the number of samples are given in Table 3-1. The description of the dataset and the definition of its attributes are explained in this section.

Table 3-1. Datasets and sample size

Dataset name	Number of safe pharmacies	Number of rogue pharmacies	Total number of samples
Traffic sources data	30	127	157
Engagement data	30	127	157
Country data	30	139	169
Social media data	24	41	65
Referral data I	30	123	153
Referral data II	50	1136	1186
Search data	30	60	90

3.2.1 Engagement

Engagement data explains the extent of involvement of users with the website. This is measured in terms of the total number of views and time spent on the website. The data description of engagement data is given in Table 3-2.

Table 3-2. Data description of engagement data

Attribute	Description	Variable type
Pharmacy	Url of the online pharmacy	Categorical
Views	Total number of visits to the given online pharmacy from september 2015 to february 2016	Numeric
Time_on_site	The average time spent on the given online pharmacy by visitors from september 2015 to february 2016 (in minutes)	Numeric
Page_view	The average number of pages viewed during each visit to the given online pharmacy from september 2015 to february 2016	Numeric
Bounce rate	Percentage of visitors to the given online pharmacy, who navigate away from the website after viewing only one page from september 2015 to february 2016	Numeric
Class	Explains if the pharmacy is legitimate (safe) or illegitimate (rogue)	Categorical – takes values s (safe pharmacy) or r (rogue pharmacy)

3.2.2 Country

Country data explains the percentage of total traffic to the online pharmacies from different countries are given. Table 3-3 provides the data description for the country dataset. It is important to note that in the given dataset the same pharmacy and country can occur more than once.

Table 3-3 Data description of country data

Attribute	Description	Variable type
Pharmacy	Url of the online pharmacy	Categorical
Country	Name of the country	Categorical
Traffic	Proportion of traffic contributed by the given country to the online pharmacy from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Class	Explains if the pharmacy is legitimate (safe) or illegitimate (rogue)	Categorical – takes values s (safe pharmacy) or r (rogue pharmacy)

3.2.3 Traffic Sources

Traffic sources data include the proportion of traffic entering the website from defined sources. The data description of traffic sources data is given in Table 3-4.

Table 3-4. Data description of traffic sources data

Attribute	Description	Variable type
Pharmacy	Url of the online pharmacy	Categorical
Direct	The proportion of traffic entering the website by directly typing in the url on browser from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Search	The proportion of traffic entering the website through search engine from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Referral	The proportion of traffic entering the website through hyperlinks on other websites from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Social media	The proportion of traffic entering the website through hyperlinks on online social media from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Email	The proportion of traffic entering the website through hyperlinks from emails from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Display ads	The proportion of traffic entering the website through online advertisements from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Class	Explains if the pharmacy is legitimate (safe) or illegitimate (rogue)	Categorical – takes values s (safe pharmacy) or r (rogue pharmacy)

3.2.4 Social Media

The proportion of social media traffic entering the online pharmacy from various social media domains are consolidated into social media dataset. The features along with their description are shown in Table 3-5. It is important to note that in the given dataset the neither the pharmacy nor the social media site are unique entries.

Table 3-5. Data description of social media data

Attribute	**Description**	**Variable type**
Pharmacy	Url of the online pharmacy	Categorical
Social_media	Name of online social media	Categorical
Traffic	Proportion of social media traffic contributed by the given social media to the online pharmacy from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Class	Explains if the pharmacy is legitimate (safe) or illegitimate (rogue)	Categorical – takes values s (safe pharmacy) or r (rogue pharmacy)

3.2.5 Search

Search traffic maybe due to organic search result or paid search result. Organic search results are the web pages provided by the search engine based on the relevance to the user's query. It is also called the natural search result. Paid search results are like advertisements. The websites pay search engines to promote their webpages for a particular keyword. The search data provides the percentage of search traffic that constitutes organic and paid search for the online pharmacies. Table 3-6 provides the data description for the search data.

Table 3-6. Data description of search data

Attribute	Description	Variable type
Pharmacy	Url of the online pharmacy	Categorical
Organic	Proportion of search traffic contributed by organic search results to the online pharmacy from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Paid	Proportion of search traffic contributed by paid search results to the online pharmacy from september 2015 to february 2016 to the given online pharmacy (in percentage)	Numeric
Class	Explains if the pharmacy is legitimate (safe) or illegitimate (rogue)	Categorical – takes values s (safe pharmacy) or r (rogue pharmacy)

3.2.6 Referral

Referring websites of a given domain which direct traffic to the domain via hyperlinks. These hyperlinks are also called backlinks. Referral data provides the different referring websites to online pharmacies, their IP address and countries of origin. The data description of the referral data is given in Table 3-7. The Pharmacy column of the dataset is not unique. However, the combination of pharmacy and Referring domain will be unique entries.

Table 3-7. Data description of referral data

Attribute	Description	Variable type
Pharmacy	Url of the online pharmacy	Categorical
Referring_domain	Url of the website which has a link to online pharmacy	Categorical
Backlinks	The number of links from the referring_domain to the online pharmacy	Numeric
Ip address	Ip address of the referring_website	Categorical
Country	Country of origin of referring_website	Factor
Class	Explains if the pharmacy is legitimate (safe) or illegitimate (rogue)	Categorical – takes values s (safe pharmacy) or r (rogue pharmacy)

Chapter 4

Analysis and Methodology

Literature review indicates that illegitimate online pharmacies are present and accessed in the World Wide Web. However, the relative usage of the legitimate and illegitimate online pharmacies is not discussed in the past. Also, the web data of these online pharmacies are not explored to discover patterns and insights to develop an automated classification system which will reduce human efforts to classify the online pharmacies. This section aims to 1) explore the web data of online pharmacies to study the major traffic sources and relative usage, 2) identify datasets which enables to develop an automated classification system and 3) Engineer an automated classification using machine learning techniques.

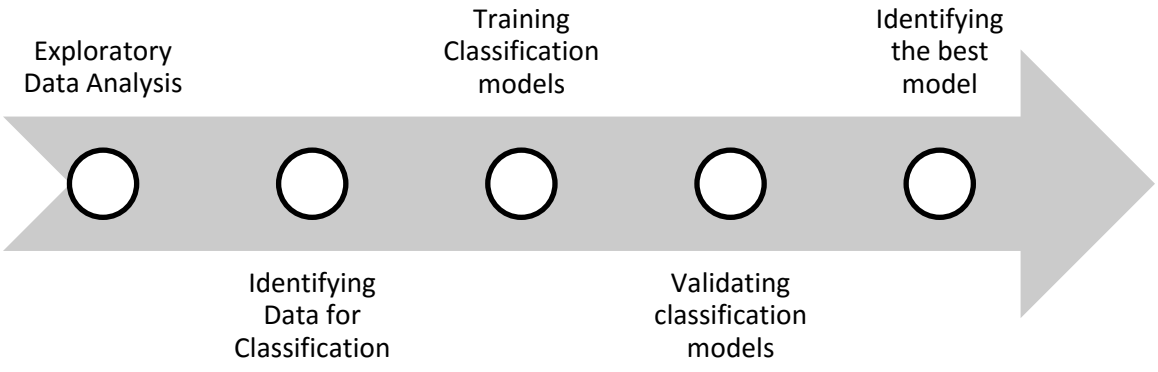

Figure 4-1. Methodology and Analysis

Studying the traffic data of online pharmacies and identifying the important sources is crucial in formulating policies to prevent consumers from visiting these websites. Also, this brings into light the traffic contribution from search engines and social media so that they can take steps to promote legitimate pharmacies. Studying engagement data – page view, time spent etc., will help to measure the relative usage of online pharmacies. Also, exploring the given datasets will help to identify the right set of features in developing the classification model. Currently NABP and Legitscript use manual methods to identify illegitimate pharmacies. An automated classification system of online pharmacies will improve the rate of discovery of online pharmacies. Also, this algorithm can be used by websites and search engines to approve of ads from online pharmacies.

Exploratory data analysis along with descriptive analytics and data visulazation is used to answer the following questions

1. What are the major traffic sources of legitimate and illegitimate pharmacies? What does it signify?
2. What is the number of views and average time spent by consumers on legitimate and illegitimate pharmacies? How are they different?
3. What are the major social media websites that refer traffic to online pharmacies?
4. Which countries contribute traffic to illegal pharmacies?
5. What attributes can be used to develop an online pharmacy classification system?

From the exploratory data analysis it was observed that, same set of referring websites point towards a particular type of pharmacy. Hence, we check the hypothesis: reliability score for an online pharmacy can be computed using the backlinks from the referring websites. This hypothesis translates into a linear classification model. The matrix given below represents the given dataset

$$\begin{matrix} & R_1 & \dots & R_m \\ \begin{matrix} P_1 \\ \vdots \\ P_n \end{matrix} & \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nm} \end{bmatrix} \end{matrix}$$

Each row $P_i$represents the pharmacy and column R_j represents the referring websites. Each element of the matrix l_{ij}gives the number of backlinks from referring website R_j to online pharmacy P_i. We aim to identify a vector $W = (w_1, w_2, \dots \dots \dots, w_m)$ such that if

$$W^T P_i > T \quad then\ P_i\ is\ legit$$

The referral data I consists of 153 pharmacies – 30 legitimate and 123 illegitimate and 9685 referring website. Also, the referral data is very sparse with almost 99% entries equal to zero. Also, the number of attributes is very high. Linear classifiers are preferred for sparse datasets with large number of attributes (Yuan et al, 2012). As discussed in literature review, linear classifiers are of two types – generative and discriminative. We have developed both classifiers in the given study. Through visualization it was observed that pharmacies of same class formed clusters. Hence nearest neighbor based classification is also adopted. Along with existing machine learning techniques, we have developed a rating based linear classification algorithm which can be used for the given problem. The weight vector (score of each referring website) is computed as the probability of a backlink from it referring to a safe website. The algorithm is discussed later in this section. Initially, the machine learning algorithms are implemented on referral data I and based on the performance of the classification models, the best two models are identified and implemented on a larger dataset referral data II with 1186 samples.

It is challenging for machine learning algorithms to identify patterns in sparse data with high dimensionality. Also, higher dimensionality leads to higher computational efforts (Yeasin et al, 2006). In this study we have compared the performance of different linear classification models on the given dataset with dimensionality reduction using PCA.

The models developed using referral dataset I are validated using leave one out cross validation. Leave one out cross validation is the K fold cross validation where K is equal to the number of samples. Leave one out cross validation is preferred when the number of samples is less. Hold out validation is used for models developed on referral dataset II. 67% of the samples are used as training set and the remaining 33% is used as the testing set. The models are validated and compared across each other using various performance measures.

The case being discussed is a binary classification problem. Binary classification is a task of classifying given samples into two groups based on classification rule. The performance measures of a binary classifier include accuracy, sensitivity, specificity and F-score. The confusion matrix for a binary classifier is given in Table 4-1 to understand the representations of True Positive, False Positive, True Negative and False Negative. The terms are described in Table 4-2. Table 4-3 gives the evaluation metrics, its formula and evaluation focus (Sokolova & Lapalme, 2009).

Table 4-1. Confusion Matrix

Data class	**Classified as positive**	**Classified as negative**
Positive	True positive (tp)	False negative (fn)
Negative	False positive (fp)	True negative (tn)

Table 4-2. Definition of elements in confusion matrix

True positive	Sample which belong to positive class is classified as positive
True negative	Sample which belong to negative class is classified as negative
False positive	Sample which belong to negative class is classified as positive (type I error)
False negative	Sample which belong to positive class is classified as negative (type II error)

Table 4-3. Performance metrics of binary classifier

Measure	Formula	Evaluation focus
Average accuracy	$(\frac{tp + tn}{tp + tn + fp + fn})$	Overall effectiveness of classifier
Precision	$\frac{tp}{tp + fp}$	Class agreement of data labels with positive labels given by classifier
Sensitivity/ Recall	$\frac{tp}{tp + fn}$	Effectiveness of classifier to identify positive class labels
Specificity	$\frac{tn}{tn + fp}$	Effectiveness of classifier to identify negative class labels
F-score	$\frac{(\beta^2 + 1)tp}{(\beta^2 + 1)tp + \beta^2 fn + fp}$	Relations between data's positive labels and those given by classifier

The classifier can make two types of errors – Type I error and Type II error. When a positive class is classified as negative, it is called Type I error. On the other hand, when a negative class is classified as positive it is called Type II error. Type I error and Type II error are also called false positive rate and false negative rate. Sensitivity describes the number of positives correctly identified and specificity describes the number of negatives correctly identified. The relationship between errors and sensitivity and specificity is given as

$$False\ positive\ rate\ (\alpha) = Type\ I\ error = 1 - specificity = \frac{fp}{fp + tn}$$

$$False\ negative\ rate\ (\beta) = Type\ II\ error = 1 - sensitivity = \frac{fn}{tp + fn}$$

A binary variables can be symmetric or asymmetric. If both the states of a binary variable (0 and 1) are equally important then it is a symmetric binary variable, otherwise it is called asymmetric binary variable. The given class variable has problem has two states – safe pharmacy (positive class) and rogue pharmacy (negative class). The outcomes of the binary variable are not equally important. The cost of misclassification of a rogue pharmacy as safe (Type II error) is

higher than misclassification of safe as rogue (Type I error). Hence the optimal model should have minimum type I error or maximum sensitivity. Also, the developed model should have reasonable accuracy. The agreement between observed and predicted classes are measured in terms of kappa (Khun, 2013). We consider accuracy, kappa, sensitivity and specificity to compare the models. In general all the machine learning model's cost function tend to improve accuracy. However, support vector machine's cost function can be adjusted by adding weights. In this study we have added different weights to increase the sensitivity by manipulating the cost function of the support vector machine and observe how the performance changes across various parameters. Also, the performance metrics of the KNN model differ with change in K values. The variation of the performance metrics across different values of K is also studied.

4.1 Exploratory Data Analysis

The different datasets discussed in Chapter 3 are explored using descriptive analytics and visualization to identify significant patterns to learn more about online pharmacies and to identify features to develop an automated classification model. This section explores the major traffic sources and engagement levels of legitimate and illegitimate pharmacies. Traffic sources are the ways through which the consumers access the online pharmacies. It is important to study them as it gives insights on customer loyalty, network of websites and social media related to online pharmacies, marketing and search engine optimization strategies of online pharmacies and market in different countries for these pharmacies. Analyzing engagement data shows the extent to which these websites are utilized and to some extent tells if the customer will come back or if he has made a transaction. In this section we will explore the various datasets to understand the relative usage of online pharmacies.

4.1.1 Traffic

The major traffic sources of any website are classified as Direct, Search, Referral, Social, Display and Email. The traffic obtained by user directly typing in the URL of the website is given by 'Direct' source. 'Search' refers to the traffic coming from the search engines like Google, Bing and Yahoo. The traffic coming in from the links on other websites are accounted for as 'Referral'. 'Social' indicates the traffic from social media websites like Facebook, Twitter and Pinterest. 'Display' indicates the amount of traffic coming in from banner advertising and 'Email' indicates the traffic coming in from links in email messages. The percentages of traffic coming into 157 online pharmacies – 127 illegitimate and 30 legitimate, from each source is analyzed in this section. Specifically, the five point summary and mean percentage of traffic coming in from each source to legitimate and illegitimate pharmacies are summarized in Tables 4-4, 4-5. Figure 4-2 shows the mean of percentage of traffic form each source into legitimate and illegitimate pharmacies.

Table 4-4. Descriptive statics of Traffic source for illegitimate pharmacies

	Direct	**Referral**	**Search**	**Social**	**Email**	**Display**
Min.	0	0	0	0	0	0
1st qu.	17.86	6.295	10.12	0	0	0
Median	30.47	15.03	38.16	0.15	0	0
Mean	34.32	21.74	39.27	0.8609	0.5695	2.453
3rd qu.	46.69	27.22	61.46	0.775	0.585	0
Max.	99.44	92.71	97.91	24.68	7.24	78.73

Table 4-5. Descriptive statics of Traffic source for legitimate pharmacies

	Direct	**Referral**	**Search**	**Social**	**Email**	**Display**
Min.	12.35	5.92	4.01	0	0	0
1st qu.	31.53	10.98	17.22	0.1625	0.14	0
Median	43.54	17.48	36.65	0.37	1.03	0
Mean	42.48	17.74	36.26	1.278	2.215	0.02833
3rd qu.	50.28	21.84	46.57	0.985	2.89	0.0275
Max.	81.27	42.27	74.41	20.4	9.79	0.23

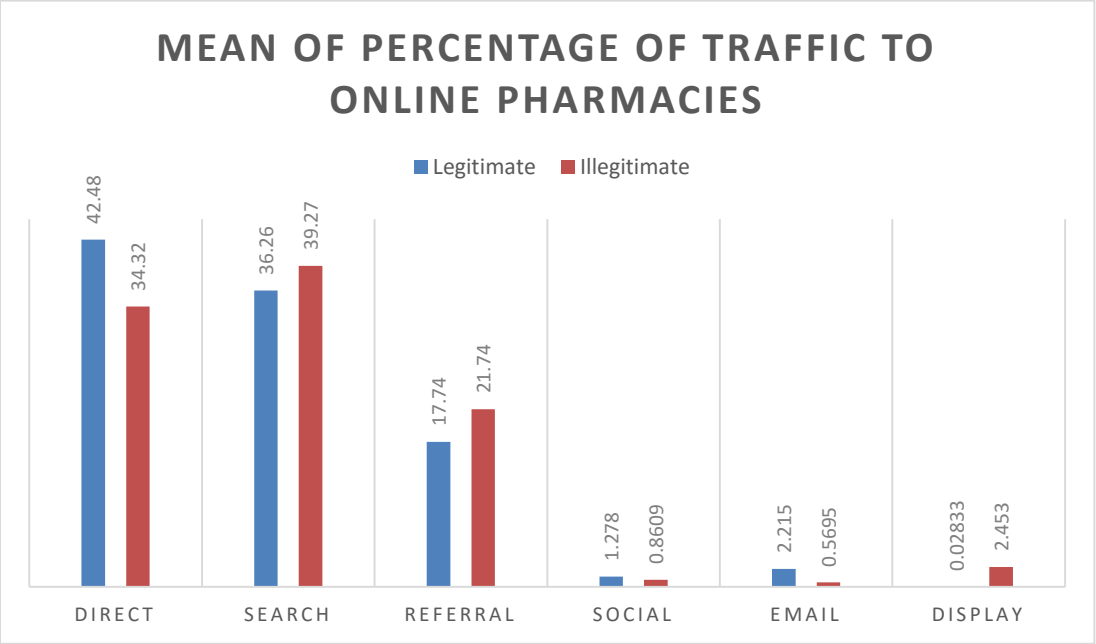

Figure 4-2. Mean of percentage of traffic to online pharmacies

From Figure 4-2 it can be observed that Direct, Search and Referral are the three major traffic sources for both legitimate and illegitimate pharmacies. Search traffic is the maximum to illegitimate pharmacies while direct traffic is maximum for legitimate pharmacies. Email traffic is minimum for illegitimate pharmacies and display advertisements traffic is minimum for legitimate pharmacies. Also it can be seen that relative usage of direct access and access through email and social media is higher for legitimate pharmacies. On the other hand, access through search, referral and display advertisements is higher for illegitimate pharmacies. The distribution of percentage of traffic from each of the sources to legitimate and illegitimate pharmacies are discussed in detail.

4.1.1.1 Direct

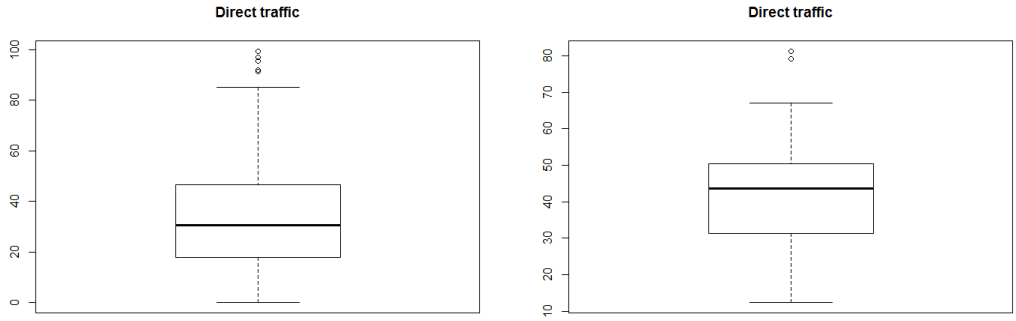

Figure 4-3. Distribution of traffic contribution from direct access (in %) for illegitimate and legitimate websites

Higher percentage of 'Direct' traffic source indicates that the website is a powerful brand. Direct traffic source reflects the set of users who visit the website knowing what they want. Figure 4-3 represents the traffic contribution (in %) from direct access for illegitimate and legitimate online pharmacies. Form the boxplots it is evident that legit pharmacies are more accessed directly compared to illicit pharmacies. Intuitively this is due to the fact that the legit pharmacies are more popular and would have more loyal customers. However it is also important to note that percentage of direct traffic is the second highest among the traffic sources for illicit pharmacies. This indicates that consumers have previous experience with these pharmacies and visit them again. A direct visit indicates successful transaction with these pharmacies. It can be understood that an illegitimate pharmacy with high direct traffic operate and sell goods which brings more customers to the site. However, the quality of the drugs sold by these pharmacies is still questionable. It is imperative to take actions to curb people directly accessing the illegitimate online pharmacies.

4.1.1.2 Search

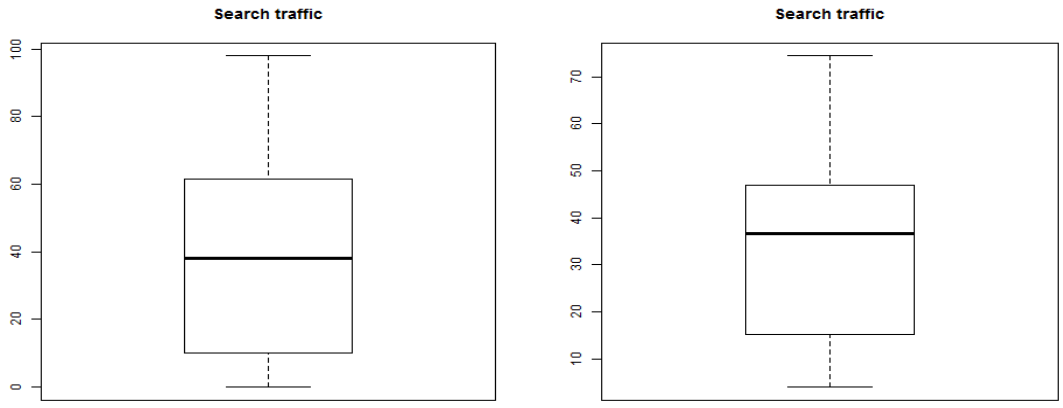

Figure 4-4. Distribution of traffic contribution from search (in %) for illegitimate and legitimate websites

Search is the major source of traffic for online pharmacies. Figure 4-4 gives the boxplot of percentage of traffic attracted by legit and illicit pharmacies through search. Both legitimate and illegitimate pharmacies attract a major proportion of traffic using search. There are two types of search – organic search and paid search. Organic or natural search results are webpage listing based on relevance to user query by the search engine. Search engine optimization is used to get higher ranking in organic search. Search Engine Optimization is used to obtain improved ranking of the domain for a particular keyword. Optimizing involves editing HTML, associated tags and the content of the webpage to increase the domains visibility in the search engine results. Almost the entire search traffic of the illegitimate pharmacies are due to organic search. It can be inferred that illicit pharmacies use search engine optimization to gather traffic into their domains. Paid search are advertisements for which the website owners have paid the search engine to display their website for a given keyword. The statics of percentage of ‘Search’ traffic through ‘Paid’ and ‘Organic’ mediums for 26 legitimate pharmacies and 59 illegitimate are given in Tables 4-6 and 4-7.

Table 4-7 indicates that almost no illicit online pharmacies use paid search as a medium to attract traffic. From dataset used in the study it can be seen that 39% of the legitimate pharmacies attracted traffic using paid search. On the other hand, only 14% of the illegitimate pharmacies attracted traffic using paid search. However, the maximum proportion of traffic attracted by paid search by illegitimate pharmacy is higher than legitimate pharmacy (60%). Even though, only a small proportion of illegitimate pharmacies are investing on online advertisements they are successful in attracting traffic.

Table 4-6. Summary statistics of percentage of Search Traffic from Organic Search for legitimate and illegitimate pharmacies

	Min.	**1st qu.**	**Median**	**Mean**	**3rd qu.**	**Max.**
Legitimate	61	93	100	93	100	100
Illegitimate	40	100	100	98	100	100

Table 4-7. Summary statistics of percentage of Search Traffic from Paid Search for legitimate and illegitimate pharmacies

	Min.	**1st qu.**	**Median**	**Mean**	**3rd qu.**	**Max.**
Legitimate	0	0	0	7	7	39
Illegitimate	0	0	0	2	0	60

4.1.1.3 Referral

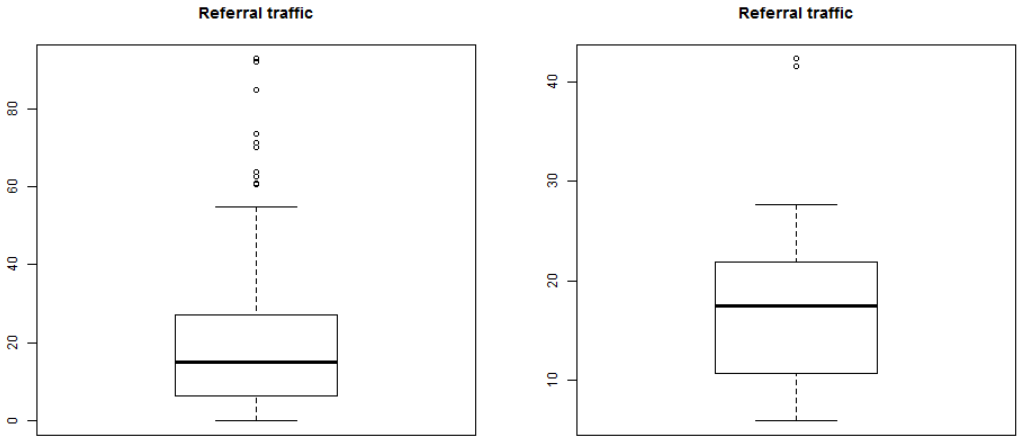

Figure 4-5. Distribution of traffic contribution from referrals (in %) for illegitimate and legitimate websites

Referrals are one of the three major source of traffic of both legitimate and illegitimate online pharmacies. The boxplot given in Figure 4-5 shows the percentage of traffic contributed by referrals to legitimate and illegitimate pharmacies. It can be observed from the boxplots that illegitimate pharmacies attract higher percent of traffic through referrals. The referring domains of online pharmacies along with the number of backlinks at each referring domain, IP address of referring domain and its country is given in the referral data. A backlink is a link on another website that points to online pharmacy and referring domain is the domain that hosts backlinks. Referral websites might send traffic to online pharmacies as a result of campaigns or banner ads. It could also be from blogs which is interested in the pharmacy. Depending on the traffic generated by the referring website, a marketing relationship could or should exist between the online pharmacies and the referring URL (Waisberg et al, 2009).

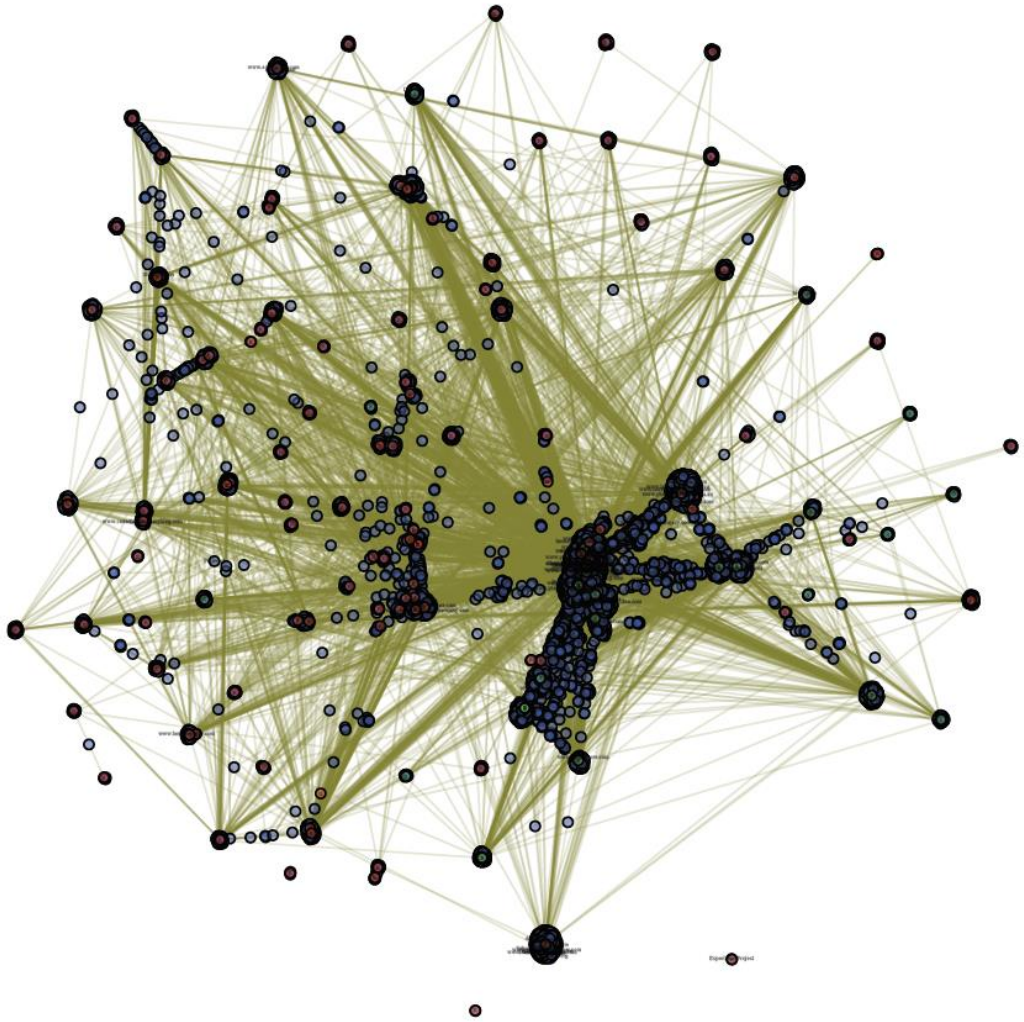

Figure 4-6. Referral network of online pharmacies

The network of referral data is given in Figure 4-6. Each node in the network denotes a website, it can be legitimate pharmacy, illegitimate pharmacy or a referring website. The rogue pharmacies are red, the safe pharmacies are green and the referring websites are blue. The link is established between if a backlink exists between. The weight of the links are given by the number of backlinks.

4.1.1.4 Social media

Previous research has indicated the presence of illegitimate pharmacies' contents on various social media sites. From Figure 4-2 it can be seen that the mean percentage of traffic entering the website through social media is less than 5% for both legitimate and illegitimate pharmacies. Figure 4-7 and Figure 4-8 show the boxplot of traffic from social media site to legitimate and illegitimate pharmacies. The mean and median percentage of traffic through social media is higher for legitimate pharmacies compared to illegitimate pharmacies. Hence it appears that legitimate pharmacies attract a higher percent of traffic than illegitimate traffic through social media.

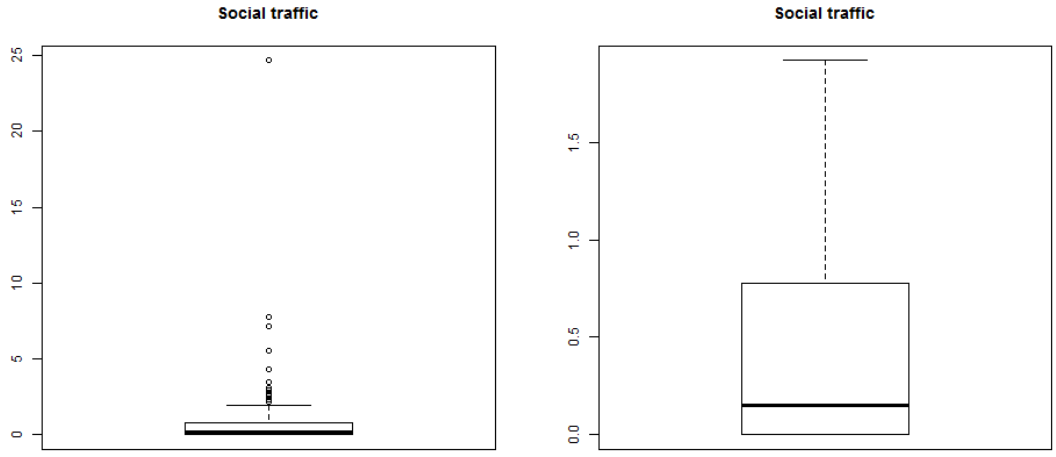

Figure 4-7. Distribution of traffic contribution from social media (in %) for illegitimate pharmacies

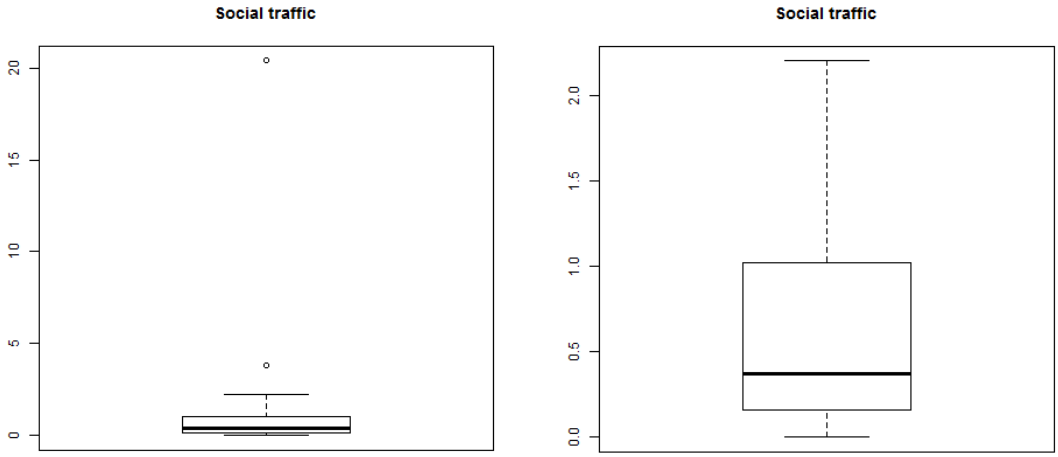

Figure 4-8. Distribution of traffic contribution from direct access (in %) for legitimate pharmacies

Further, the share of social media traffic from 26 social media sites to 24 legitimate and 41 illegitimate pharmacies is analyzed. 24 (92%) of the studied social media websites direct traffic to illegitimate pharmacies and 13 (50%) of them direct traffic to legitimate pharmacies. 11 (42%) of the websites direct traffic to both legitimate and illegitimate pharmacies. 13 (50%) of the sites direct traffic to only illegitimate pharmacies and 2 (7%) of the websites direct traffic to only legitimate pharmacies. Figure 4-9 depicts the social media and online pharmacies graph and Figure 4-10 presents the proportion of traffic from various social media websites to legitimate and illegitimate pharmacies. In Figure 4-9 the blue, green and red nodes represent social media websites, safe pharmacies and rogue pharmacies respectively. A link is established between the nodes if any traffic flow is present between the websites. The percentage of traffic between the websites are assigned as weights to the links.

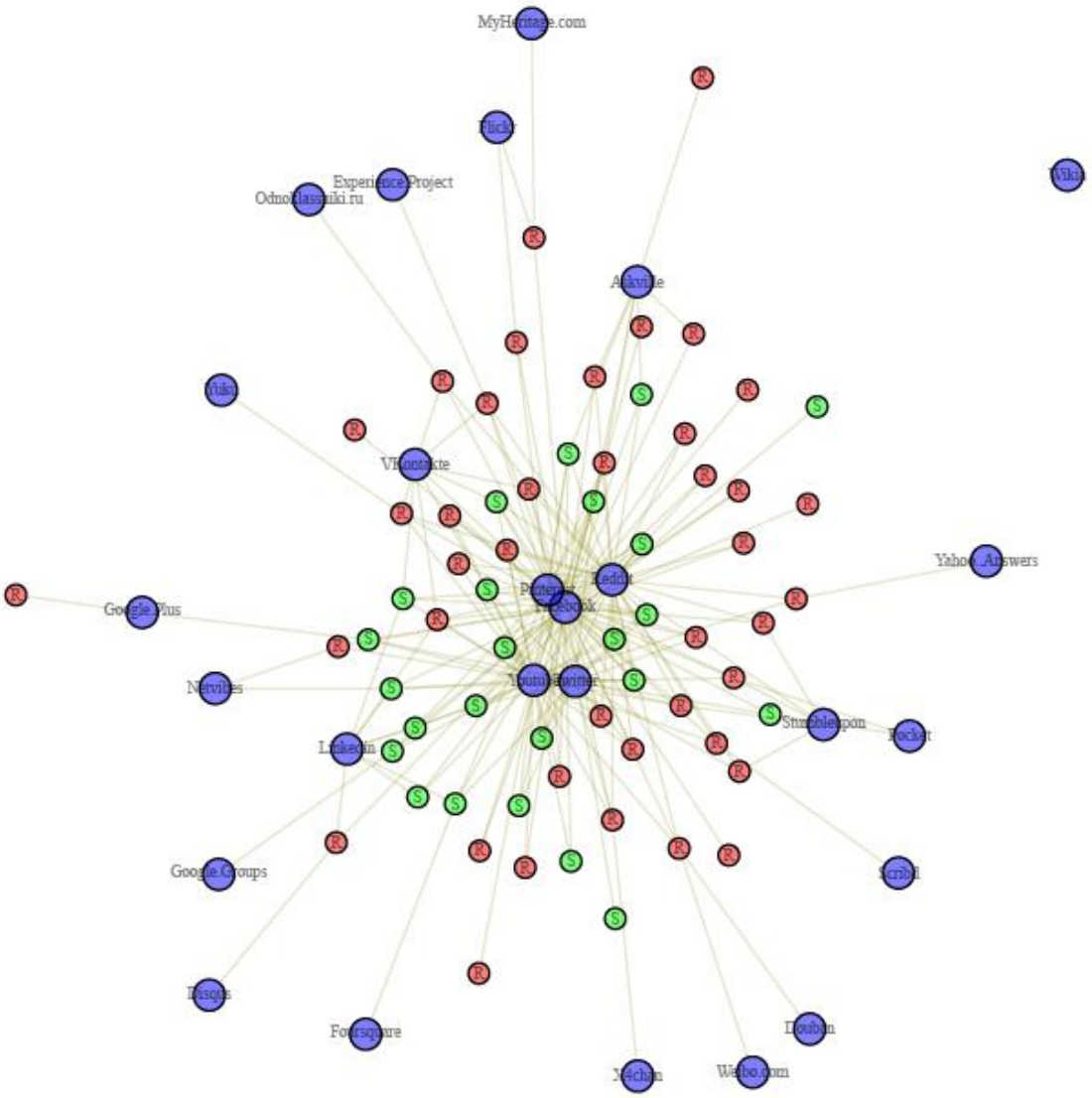

Figure 4-9. Social Media and Online pharmacies traffic network

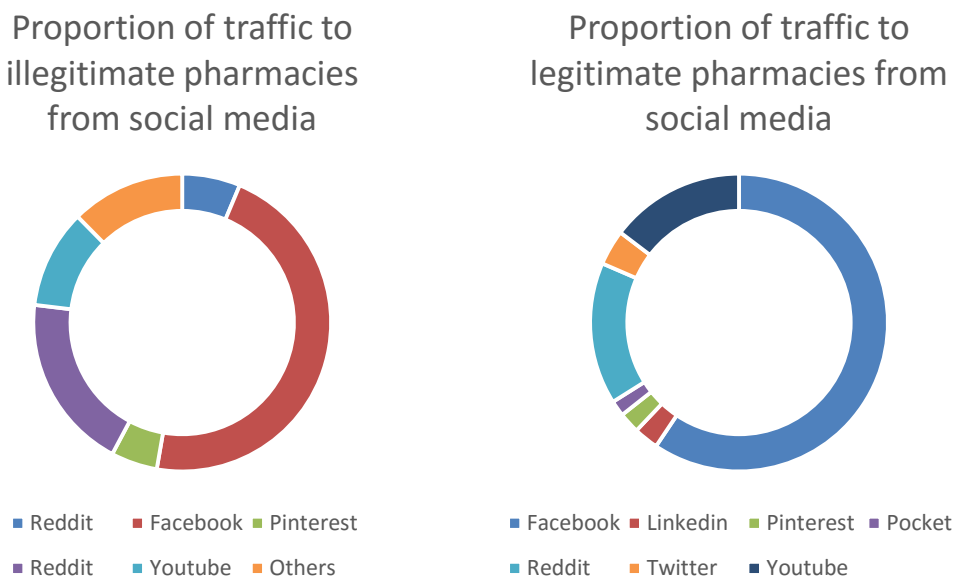

Figure 4-10. Social traffic to online pharmacies

4.1.1.5 Countrywise traffic

Country data set includes the traffic contribution from different countries to online pharmacies. The dataset includes 139 illegitimate online pharmacies and 30 legitimate pharmacies. It can be seen that online pharmacies attract traffic from 52 countries. Traffic from 27 country (52%) point only towards illegitimate pharmacies. On the other hand 3 country (6%) has traffic only towards legitimate pharmacies. Figure 4-11 shows the proportions of traffic from different countries to both legitimate and illegitimate online pharmacies. It can be observed that United States is the major consumer for online pharmacies as it has the highest traffic to both legitimate and illegitimate online pharmacies. United States has almost 60% of the traffic proportion to illegitimate online pharmacies and 97% of traffic proportion to legitimate online pharmacies.

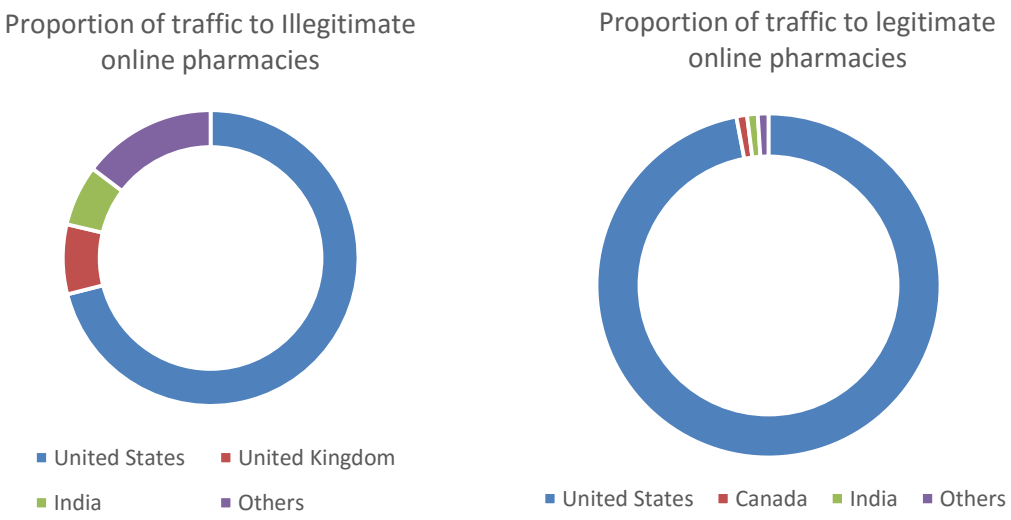

Figure 4-11. Country wise traffic distribution to online pharmacies

4.1.2 Engagement data

Engagement data explains the number of visits and the time spent by visitors on the website. We take into consideration average views to the websites, average time spent by the visitors in the website, average number of pages viewed by the visitors and average bounce rate. The descriptive statistics of engagement data for online pharmacies are summarized in Tables 4-8 and 4-9.

Table 4-8. Descriptive statics of engagement data for illegitimate pharmacies

	Views	**Time**	**Pageview**	**Bouncerate**
Min.	0	0	0	0
1st qu.	4000	1.642	2.73	34.73
Median	9000	3	3.74	49.52
Mean	22750	3.28	4.02	49.38
3rd qu.	20000	4.383	4.91	60.13
Max.	420000	17.42	13.46	100

Table 4-9. Descriptive statics of engagement data for legitimate pharmacies

	Views	**Time**	**Pageview**	**Bouncerate**
Min.	1.3	1.217	1.89	7.76
1st qu.	1510	2.633	3.785	15.96
Median	15000	4.367	7.02	32.14
Mean	107200	4.951	7.183	32.22
3rd qu.	117500	6.688	10.28	45.08
Max.	920000	10.55	13.54	71.34

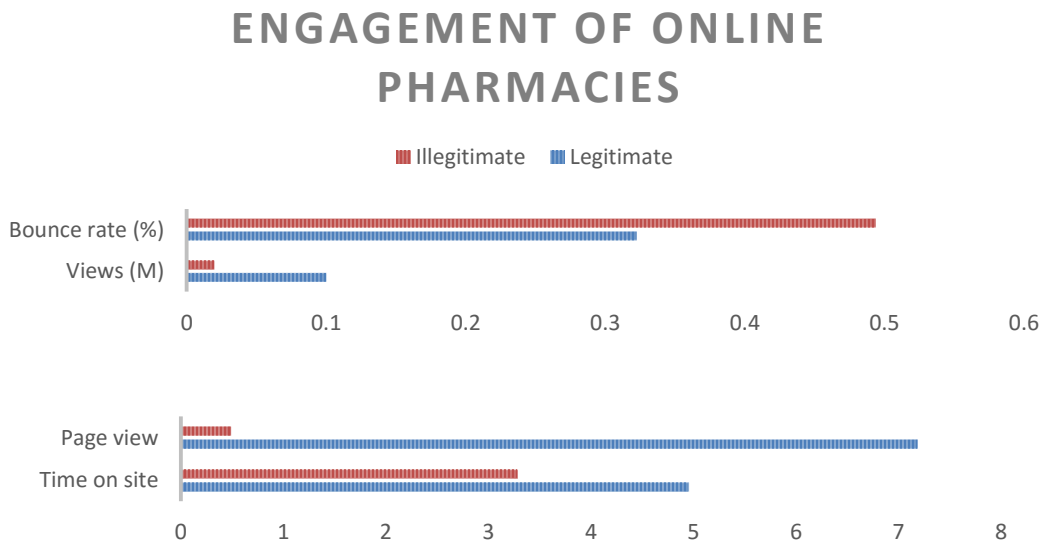

Figure 4-12. Engagement of online pharmacies

Figure 4-12 shows the average of different engagement metrics across legitimate and illegitimate pharmacies. It can be seen that the average views of legitimate pharmacies is higher than that of illegitimate pharmacies. Similarly, average number of pages viewed and time on site for legitimate pharmacies are higher than illegitimate pharmacies. The bounce rate of illegitimate pharmacies is higher that the legitimate pharmacies. Hence it can be summarized that visitors are more engaged in legitimate pharmacies than illegitimate pharmacies. The distributions of individual metrics for 30 legitimate and 127 illegitimate online pharmacies are discussed.

4.1.2.1 Views

Views indicate the traffic entering the websites. It can be used to gauge the popularity of the website. The average views of the legitimate pharmacies is higher than that of illegitimate pharmacies. The distribution of average views of legitimate and illegitimate online pharmacies are given in Figures 4-13 and 4-14.

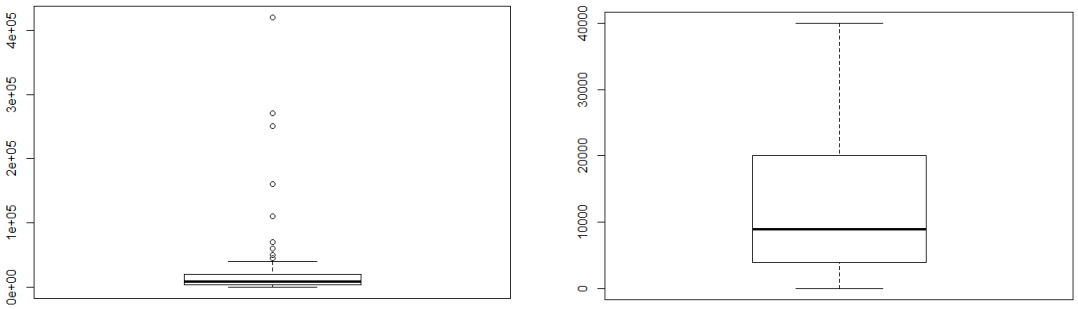

Figure 4-13. Views of illegitimate pharmacy with and without outliers

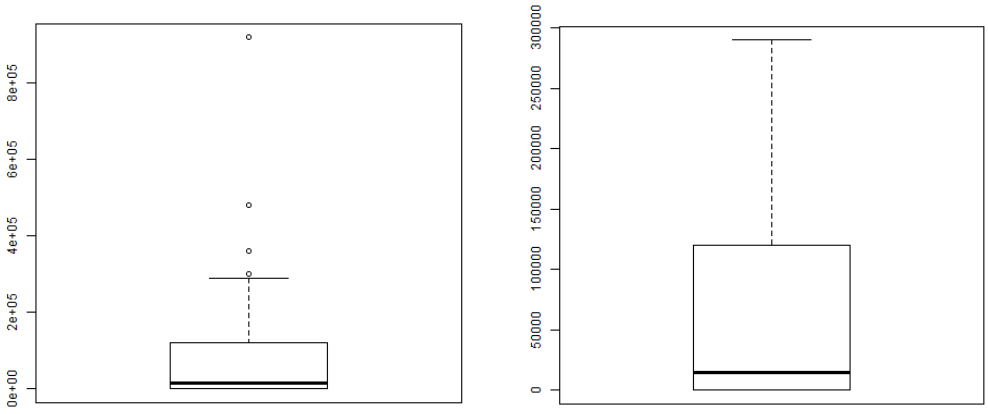

Figure 4-14. Views of legitimate pharmacy with and without outliers

4.1.2.2 Time on site

Time on site explains the average time spent on the website. Higher time on site implies it gathers higher attention from the user end and potential sales of drugs. On the other hand lower time implies the website is not very popular at the user end. Users might leave the site after browsing without making any transaction. Figure 4-15 shows the distribution of time on site (in minutes) for both legitimate and illegitimate pharmacies.

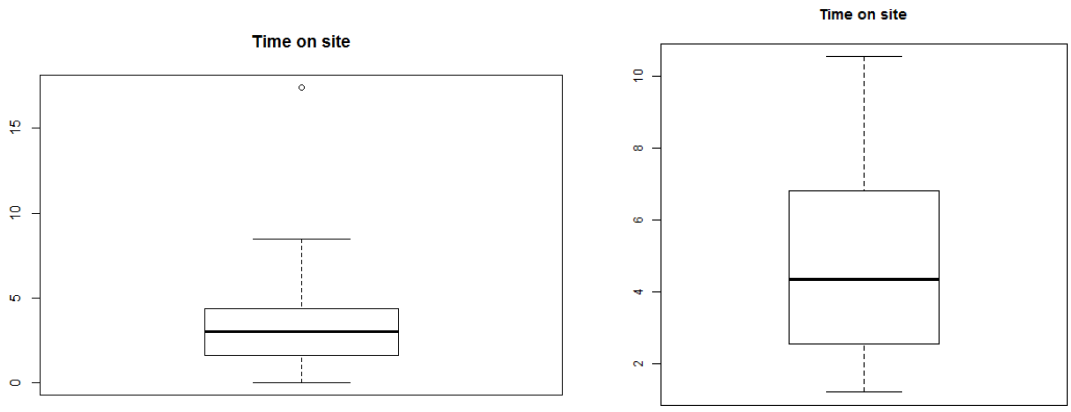

Figure 4-15. Distribution of average time spent on site for illegitimate and legitimate online pharmacies (in minutes)

4.1.2.3 Page view

Page view represents the average number of pages visited by a person in a given website. This is a useful metric to measure how the user explores the website. The page view of illegitimate pharmacies is lower that page view of legitimate pharmacies. A visitor might have to visit at least 2 pages to make a purchase from online websites and average page view less than two implies no transaction is made with the pharmacy. The distribution of page views for legitimate and illegitimate pharmacies are given in Figure 4-16.

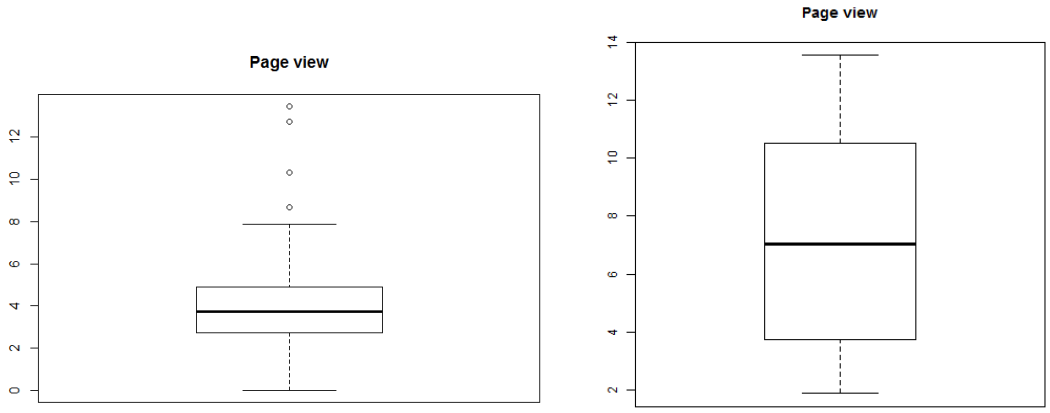

Figure 4-16. Distribution of average page view on illegitimate and legitimate online pharmacies

4.1.2.4 Bounce rate

The percentage of visitors to a particular website who navigate away from the site after viewing only one page is defined as bounce rate. Higher bounce rate implies less chances of transactions. The distributions of bounce rate for legitimate and illegitimate pharmacies are given in Figure 4-17.

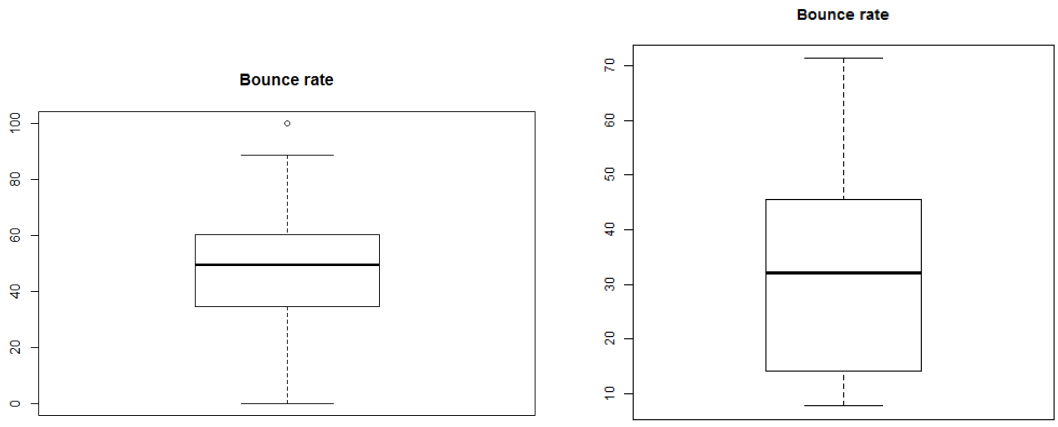

Figure 4-17. Distribution of bounce rate of illegitimate and legitimate pharmacies

4.2 Identifying data for classification

Dataset combined with traffic and engagement attributes and referral data set with the different referring websites and backlinks from them are further explored to check the feasibility of developing a classifier. The attributes of traffic and engagement data are Total number of views, Average time on site, Page view, bounce rate, direct traffic, search traffic, referral traffic, email traffic, social media traffic and ads traffic. The attributes in the referral data are the referring websites and the value of data points are the number of backlinks from each referral website to given pharmacy. In this section we visualize the datasets as scatter plots by projecting them onto their first two principal components and see if any feasible decision boundary exist.

4.2.1 Traffic and engagement data

The traffic and engagement data includes 157 samples and 10 attributes. PCA is performed on the traffic and engagement data. The scree plot of PCA of traffic and engagement data is given in Figure 4-18. The data is projected onto the first two components and plotted. The scatter plot is given in Figures 4-19 and 4-20. The two different classes rogue pharmacy (r) and safe pharmacy (s) are distinguished by blue and red colors respectively. It can be seen that the rogue pharmacies and safe pharmacies are uniformly spread. There are no pockets of one particular class. Hence, classification using engagement or traffic data may not be successful. This also indicates that engagement and traffic sources of safe and rogue pharmacies are not significantly different.

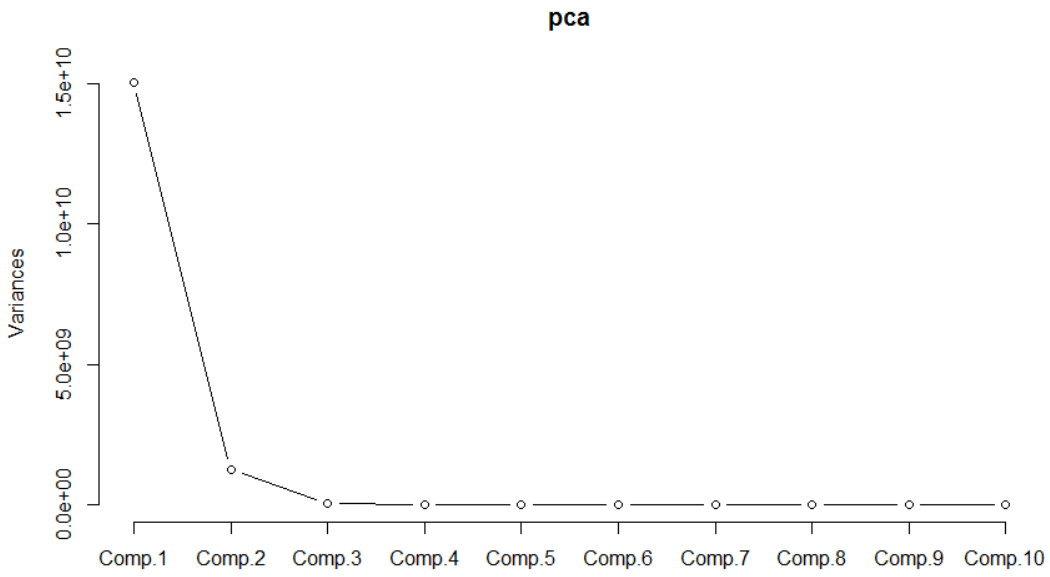

Figure 4-18. PCA of engagement and traffic data

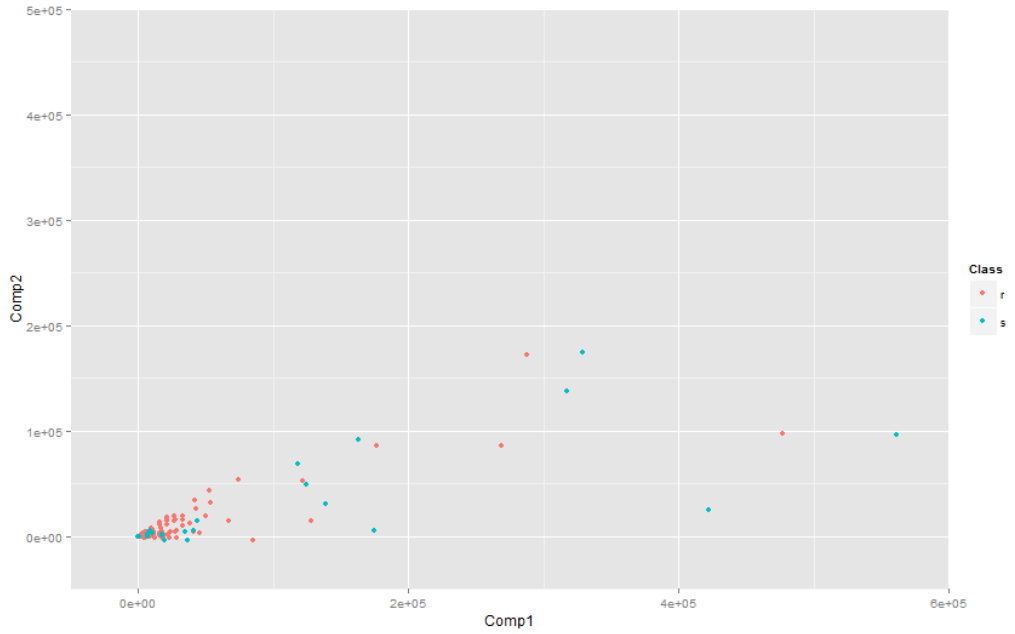

Figure 4-19. Scatter plot (I) of Engagement and traffic data

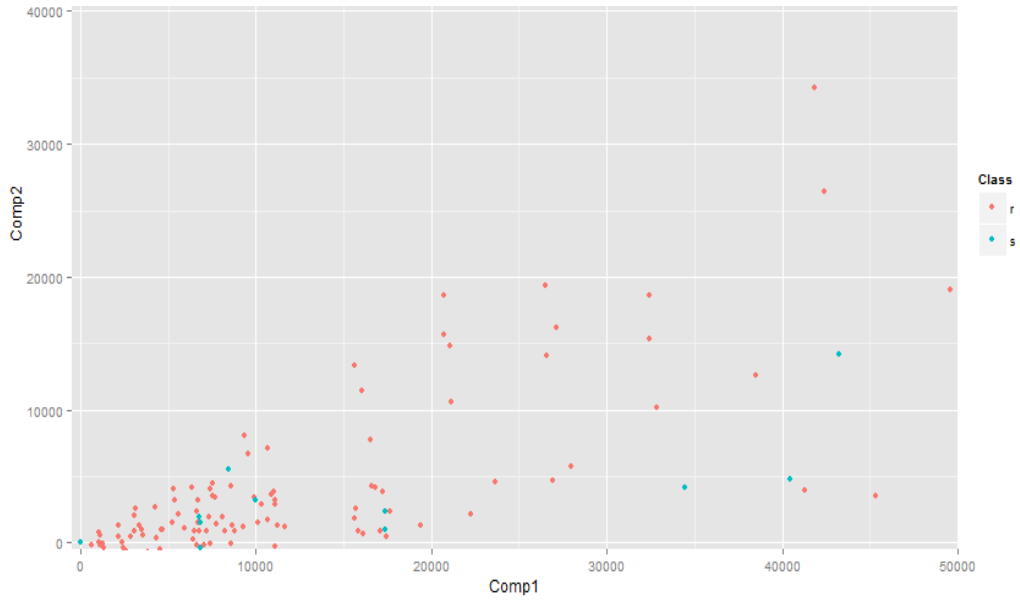

Figure 4-20. Scatter plot (II) of Engagement and traffic data II

4.2.2 Referral data

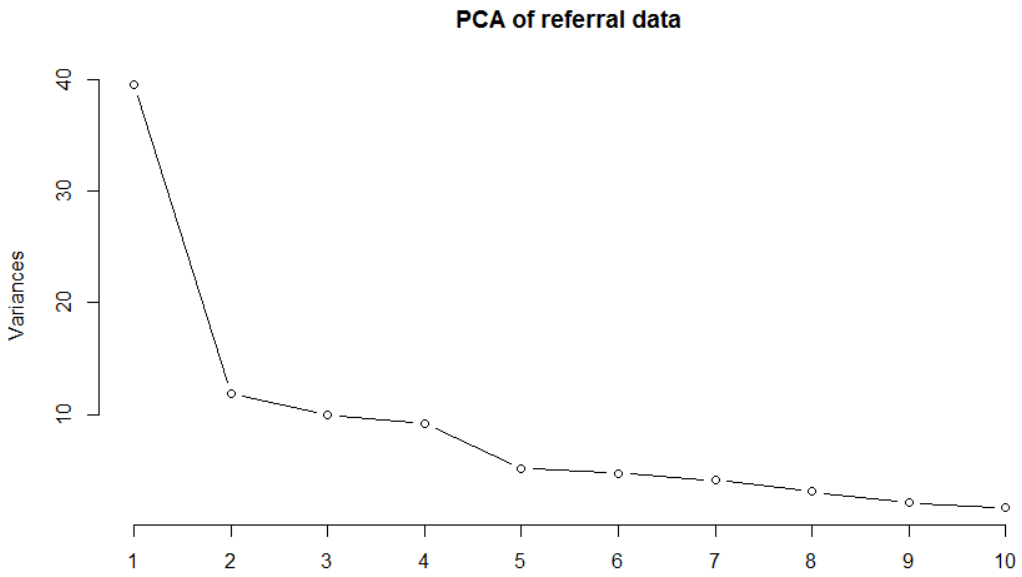

Figure 4-21. PCA of referral data

Referral data I includes the number of backlinks from different referring websites to 30 safe pharmacies and 123 rogue pharmacies. It consists of 9685 attributes, the referring websites that refer traffic to these pharmacies. Figure 4-21 shows the scree plot of the PCA of referral data and Figure 4-22 shows the scatter plot developed by projecting the data onto the principal components. The safe pharmacies (s) are depicted with red color and rogue pharmacies (r) are show in blue color. It can be seen that there is a distinct boundary between safe and rogue pharmacies.

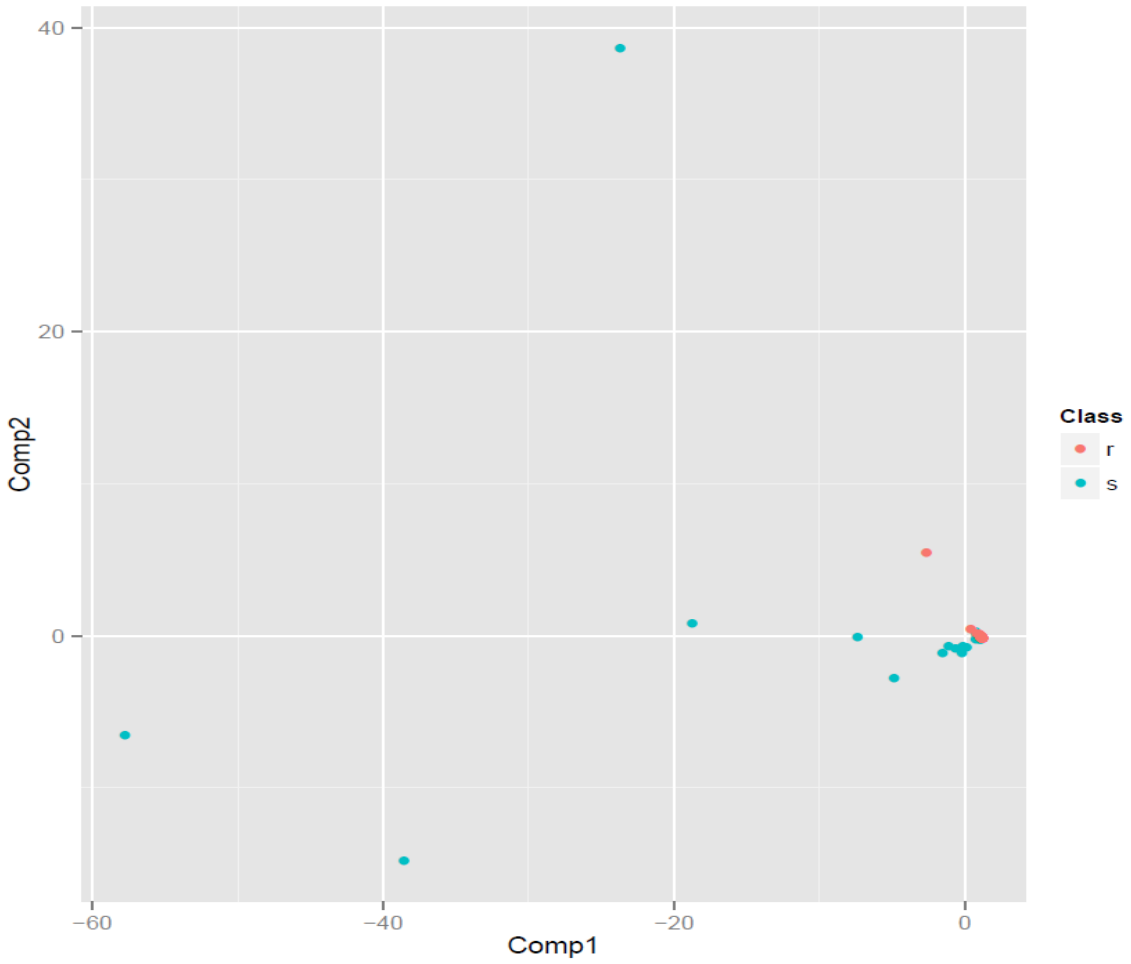

Figure 4-22. Scatter plot of referral data

Classification model can be implemented on the referral data. The obtained results are coherent with the literature. Structure data has been widely adopted and Usage data has not been used for classification of websites. The decision boundary appears to be linear. Additionally, in the past linear classifier (Navies Bayes) has been used on structure data to develop website classification models. Hence we adopt linear classifiers for classification of online pharmacies. From Figure 4-22 it can be observed that the rogue and safe pharmacies are present as clusters. Hence nearest neighbor based classification model KNN is also implemented and validated.

4.3 Classification of online pharmacies

Numerous online pharmacies appear on the World Wide Web every day. Currently, NABP and legitscript manually monitor and classify theses online pharmacies as legitimate or illegitimate. It is crucial to develop an automated system to classify the online pharmacies as it will save time and efforts. This section provides the framework and methodology suggested to classify online pharmacies. The methodology for classification is given in Figure 4-23.

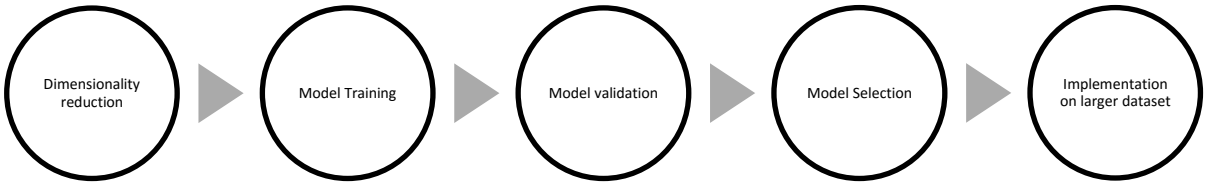

Figure 4-23. Steps involved in classification of online pharmacies

4.3.1 Dimensionality reduction

Higher dimensions might result misleading predictions due to the following reasons: 1) It is challenging for the machine learning algorithm to identify patterns from sparse data and 2) Higher dimensionality might add large amount of noise which lead to inaccurate results (Yeasin et al, 2006). Reduction of dimensionality is crucial in the given case for the machine learning algorithms to work effectively, as the number of attributes are 9685 and the dataset is very sparse. Principal Component Analysis (PCA) is a multivariate dimensionality reduction technique that extracts important information from a given dataset and represents it as a new set orthogonal vectors (Rencher, 2002). PCA is used for dimensionality reduction and features which represent 99% percent of variance are taken into consideration.

4.3.2 Model Training

Three different linear classifiers are trained and tested in this study. The linear classifiers include Linear Discriminant Analysis, Logistic Regression and Support Vector Machines. Linear Discriminant Analysis (LDA) is a generative linear classifier which identifies the linear combination of attributes that split up the samples into different classes. LDA estimates the joint probability distribution $p(x, y)$ for the predictor x given the class label y, and uses Bayes Theorem to predict $p(y|x)$ (Fisher 1936; James et al, 2013). On the other hand, Logistic regression and Support Vector Machines are discriminative classifiers which models the conditional distribution $p(y|x)$ using the predictors using training set (Cox, 1958; Jordan, 2002). Logistic regression employs the logistic function to estimate the probability. Support Vector Machines locates a hyper-plane which divides the data points with respect to their class by maximizing the distance between the data points of the same class and the hyper-plane (Vapnik 1995; Guo, Li & Chan, 2000). As mentioned earlier, it is important to consider the specificity during model selection. Support Vector Machines allows cost sensitive learning which can be used to manipulate the sensitivity of the model (Veropoulos et al, 1999). The cost parameters are set at different values to study the variation of specificity and accuracy of the model along with the cost and the more appropriate model is selected.

K Nearest Neighbor (KNN) is a supervised learning algorithm which classifies the samples in the test set based on the proximity of them to the samples of different classes in the training set (Fix and Hodges, 1951; Lihua, 2006). For instance, in 3-NN, that is KNN with K=3, for a given sample in test set the three nearest points from the training sets are identified as set S. The class of the majority of the instances in S is assigned as the class of test sample. The performance of the KNN model varies for different values of K. In this study, the statistics of the classification model with values of K 1 to 9 are studied.

Along with the mentioned machine learning techniques, an intuitive algorithm which computes the probability of a pharmacy being safe or rogue is developed and tested. According to the algorithm, the probability of a pharmacy begin safe is computed as the proportion of safe backlinks to the pharmacy. The algorithm is discussed in detail in the following paragraphs.

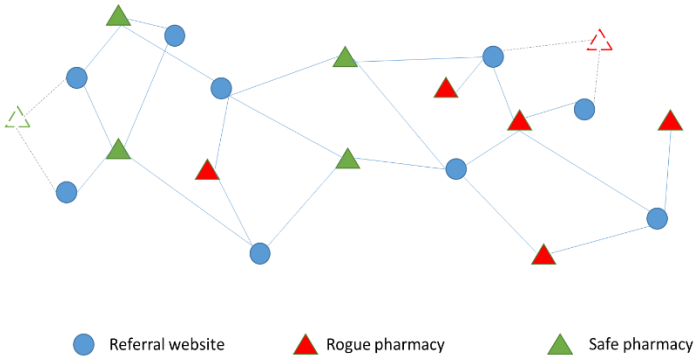

Figure 4-24. Representative graph of referral dataset

Let the data be represented as

$$\begin{matrix} & R_1 & \dots & R_m & Y \\ P_1 \\ \vdots \\ P_n \end{matrix} \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nm} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

The notations used in the algorithm are

i Index for pharmacies. $i = 1, 2, \dots \dots \dots, n$

j Index for referring websites. $i = 1, 2, \dots \dots \dots, m$

s Index for safe pharmacies.

l_{ij} Number of backlinks to P_i from R_j

y_i Class label forP_i. It can take values r or s

M_j Probability of a backlink for a referral website directing to a safe pharmacy

R_i Reliability score of pharmacy p

T Threshold

Steps in the algorithm include

Step 1: Compute the probability of a backlink for a referral website directing to a safe pharmacy

$$M_j = \frac{\sum_{k \forall s} l_{kj}}{\sum_{k \forall i} l_{kj}}$$

Step 2: Compute the probability of a backlink for a referral website directing to a safe pharmacy which is defined as the proportion of potential safe backlinks to the pharmacy. The number of potential safe backlinks to a pharmacy form a given referral website is given by the product of probability of a backlink being safe and the total number of backlinks from to the pharmacy from the referral website. The sum of potential safe backlinks from all the websites gives the total number of potential safe backlinks to the pharmacy. The proportion of potential safe backlinks provides the estimate of reliability of the pharmacy.

$$R_i = \frac{number\ of\ potential\ safe\ backlinks\ to\ pharmacy\ i}{total\ number\ of\ backlinks\ to\ pharmacy\ i}$$

$$total\ number\ of\ backlinks\ to\ pharmacy\ i = \sum_{k=1}^{m} l_{ik}$$

$$number\ of\ potential\ safe\ backlinks\ to\ pharmacy\ i = \sum_{k=1}^{m} M_k \times l_{ik}$$

$$R_i = \frac{\sum_{k=1}^{m} M_k \times l_{ik}}{\sum_{k=1}^{m} l_{ik}}$$

It is important to note that the sum of proportion of potential safe backlinks to the pharmacy and potential unsafe backlinks to the pharmacy is 1. Ideally if $R_i > 0.5$ then a pharmacy can be deemed as safe. However, when a given pharmacy doesn't have any referral website from the training set, its R_i will be indeterminate$\left(\frac{0}{0}\right)$. Hence, for such cases $R_i$is set to 0.5.

Step 3: Estimation of threshold T. Sensitivity of the model is a crucial factor to be considered. Classifying a rogue pharmacy as safe has higher cost compared to classifying a safe pharmacy as rogue. Taking this into consideration, from the training set threshold is estimated as

$$T = \min_{k \forall s} R_s$$

Step 4: Classification. The various parameters are estimated from the training set. For the given pharmacy R_i is calculated. If $R_i > T$, then i is classified as safe.

The different classification models developed on referral data I are given in Table 4-10. The models with better performance over others are implemented on a larger dataset to verify consistency and generalizability.

Table 4-10. Classification models

Model	**Dimensionality reduction**	**Classification algorithm**
LDA	PCA	Linear Discriminant Analysis
LR	PCA	Logistic Regression
SVM	PCA	Support Vector Machines with different cost values
KNN	No dimensionality reduction	K nearest neighbor with different cost values
RM	No dimensionality reduction	Rating Algorithm

4.3.3 Model Validation

Cross-validation is often used to test the performance of statistical model on previously unseen data. In K fold cross validation, data is divided into K disjoint sets. K models are trained using different combinations of K-1 groups and the model is tested on the remaining portion. The performance statistic of K fold cross validation is the mean of the statistic evaluated for each K model. The extreme case of K fold cross validation is leave one out cross validation. In leave one out cross validation K is equal to the number of samples in the dataset (Cawley & Talbot, 2003).

The leave one out cross validation estimator is an almost unbiased estimate of the generalization ability of classifier (Chapelle, 2002). However, for large datasets the variances of K fold cross validation and leave one out cross validation are likely to be similar. Leave one out cross validation is recommended in cases where training data is severely limited and is rarely adopted for large datasets as it is computationally expensive (Cawley & Talbot, 2003). In the given study, the number of samples is restricted to 153. Even a small change in the training data will result in substantial change in the developed model. Hence leave one out cross validation is used to validate the developed models.

4.3.4 Model selection

Model selection involves the identification of appropriate classifier among the developed ones. Different performance metrics are used to evaluate the effectiveness of the classifier. In the given study, we use accuracy, sensitivity, specificity and kappa to identify the best classifier. Accuracy is a measure which explains the overall effectiveness of the classifier. Sensitivity and Specificity represents the ability of the classifier to correctly identify positive and negative class labels (Sokolova & Lapalme, 2009). Kappa measures the agreement between observed and predicted classes (Khun, 2013).

Value path is an effective way to display tradeoff between solutions when there are more than two criteria (Schilling et al, 1983). Using Value Path it is easier to eliminate dominated models and identify the best suitable model. We adopt value path approach to visualize the performance of different models across various metrics. The models which performs better than the other developed models are identified through value path and are implemented on referral data II.

4.3.5 Implementation on larger dataset

Referral data II includes 1186 pharmacies – 50 legitimate and 1136 illegitimate pharmacies. Referral data is a representative sample of the existing online pharmacy market place. According to Legitscript, only 4% of the online pharmacies are legitimate (Legitscript, 2016). The appropriate classification models selected from the smaller dataset is implemented on the representative dataset to verify generalizability and consistency. The models are validated using hold out validation. Two thirds of the dataset is used from training and one third of the dataset is used for testing. The algorithm which performs better on the representative dataset is identified as the appropriate classification model.

The mentioned framework is implemented on the referral datasets – referral dataset I and referral dataset II collected from SEMrush and its results are discussed in the next chapter. Different linear classifier – Linear Discriminant Analysis, Logistics Regression and Support Vector Machines, along with the rating method and KNN are implemented. Cost sensitive learning of Support Vector Machines is adopted to alter the sensitivity of the model. Similarly, the performance of KNN across different values of K is also studied. R, a statistical programming language, is used to develop and validate the above discussed models.

Chapter 5

Results

Analysis of Web data of online pharmacies helped in understanding the major sources and relative usage of illegitimate and legitimate online pharmacies. It also helped in identifying the right set of attributes to be used in developing an automated online pharmacy classification system. Exploratory data analysis suggested that the referral dataset which includes the various referring domains and the number of backlinks from them can be used to develop an automated classification system.

A framework was proposed in Chapter 4 for the classification of online pharmacies. It involves dimensionality reduction using Principal Component Analysis to reduce the computational efforts followed by training and testing of different classifiers which include Linear Discriminant Analysis, Logistic Regression and Support Vector Machines and KNN. Along with the mentioned methods, an intuitive method – rating method is also proposed. All the developed models are validated using leave-one-out cross-validation. The results of these classification models are discussed in detail in the given chapter. The performance of these models across accuracy, sensitivity, specificity and kappa are compared using value path. The models which perform better are implemented on a larger dataset which is an ideal representation of the existing online pharmacies market place. The best classification model is identified based on the performance on representative dataset.

5.1 Dimensionality reduction

Principal component Analysis was performed on the referral dataset to reduce the number of features without loss of important information. Using PCA the attributes are transformed into 9685 orthogonal vectors. The proportion of variance and cumulative variance contributed by the principal components are shown in Figure 5-1 and Figure 5-2. The first 90 principal components explain the 99% variance of the dataset and are chosen as the new attributes for training the classifiers.

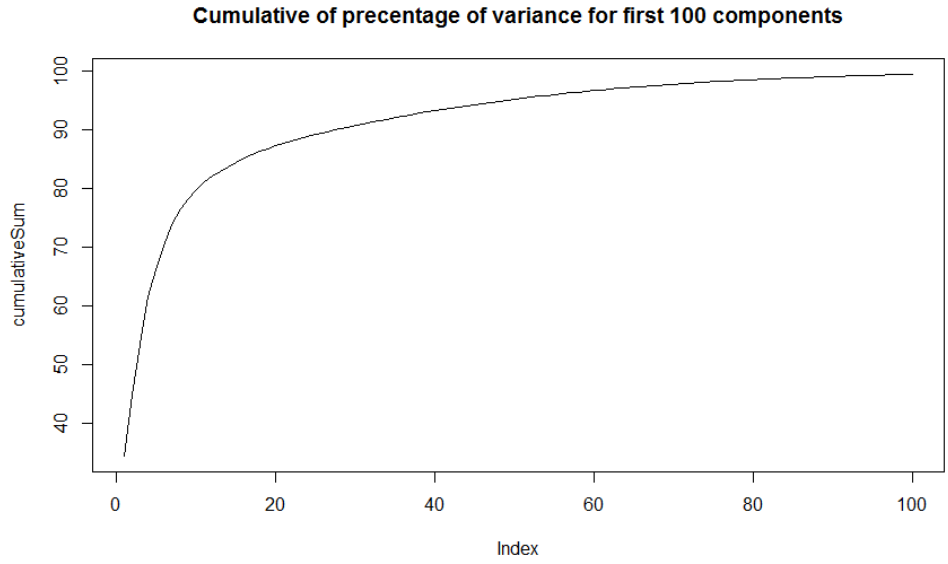

Figure 5-1. Cumulative of percentage of variance for first 100 components

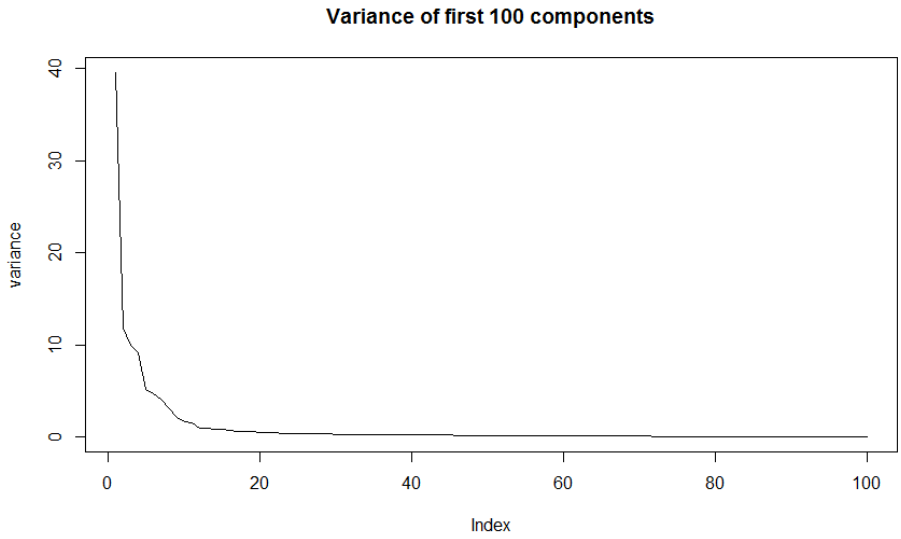

Figure 5-2. Variance for first 100 components

5.2 Model training and validation

Linear discriminant Analysis and Logistic Regression were performed on the dataset. Also, Support Vector Machines with three different cost values were implement. Rating method, developed for the given dataset is also trained and tested. All the mentioned models were validated using leave-one-out cross validation. The results of these models are summarized below.

5.2.1 Linear Discriminant Analysis (LDA)

Table 5-1. Confusion matrix of LDA

Data class	**Rogue**	**Safe**
Classified as rogue	110	7
Classified as safe	13	23

Table 5-2. Statistics of LDA

Statistic	Value
Accuracy	0.8693
Kappa	0.6145
Sensitivity	0.8943
Specificity	0.7667

5.2.2 Logistic Regression (LR)

Table 5-3. Confusion matrix of LR

Data class	Rogue	Safe
Classified as rogue	105	7
Classified as safe	18	23

Table 5-4. Statistics of LR

Statistic	Value
Accuracy	0.8366
Kappa	0.5448
Sensitivity	0.8537
Specificity	0.7667

5.2.3 Support Vector Machines (SVM)

The cost parameter of Support Vector Machines can be adjusted to alter the type I and type II error of the model. We have tried different cost values to achieve higher sensitivity. The performance of SVM for different cost values are summarized in Table 5-5.

Table 5-5. Statistics of SVM

	SVM (c=1)	**SVM (c=0.5)**	**SVM (c=0.3)**	**SVM (c=0.1)**
Accuracy	0.8693	0.8627	0.8562	0.8497
Kappa	0.5515	0.5225	0.4781	0.3516
Sensitivity	0.9431	0.9431	0.9512	0.9919
Specificity	0.5667	0.5333	0.4667	0.2667

Form Table 5-5 it can be seen that increase in sensitivity may reduce accuracy and specificity. SVM models have accuracy greater than 85%. SVM (c=0.1) has the highest specificity (99%) however, it has the lowest kappa and specificity. The graphs in Figure 5-3 shows the gain or loss in different performance measures which change in cost. It can be seen that as cost increases accuracy, kappa and specificity increases and sensitivity decreases. It is important to consider this trade of and select a model. SVM (c=0.3) has a reasonable performance across all the metrics.

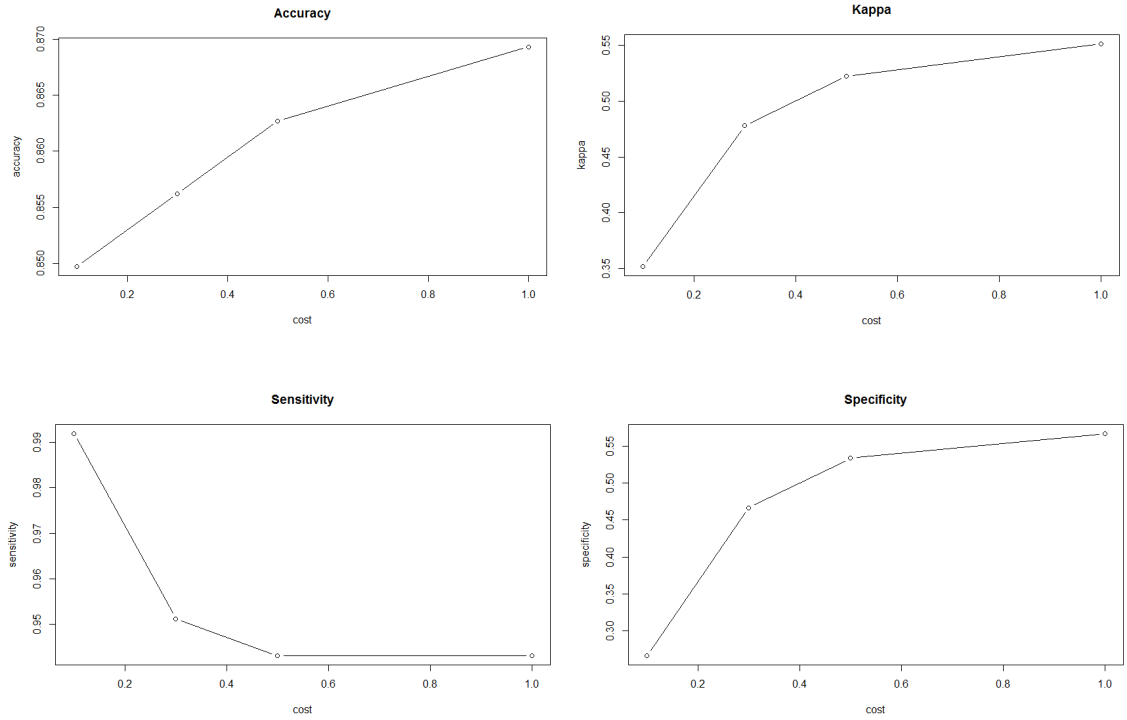

Figure 5-3. Change in performance measures of SVM with change in cost

5.2.4 K – Nearest Neighbor (KNN)

The performance of the KNN classification model varies with K values. The KNN models for values of K from 1 to 9 are implemented and tested on referral data I using hold one out cross validation. The results are summarized in Table 5-6.

Table 5-6. Statistics of KNN

	KNN (K=1)	**KNN (K=2)**	**KNN (K=3)**	**KNN (K=4)**	**KNN (K=5)**	**KNN (K=6)**	**KNN (K=7)**	**KNN (K=8)**	**KNN (K=9)**
Accuracy	0.922	0.8693	0.8627	0.8431	0.8301	0.817	0.8105	0.8039	0.8039
Kappa	0.715	0.4965	0.408	0.2867	0.1983	0.103	0.0525	0	0
Sensitivity	0.992	0.9756	1	1	1	1	1	1	1
Specificity	0.633	0.4333	0.3	0.2	0.1333	0.0667	0.0333	0	0

It can be seen that the sensitivity of the model increases with K while accuracy, specificity and kappa values decreased. The change in the statistics along with K is represented in Figure 5-4. 1-NN has performed the best across the different KNN models.

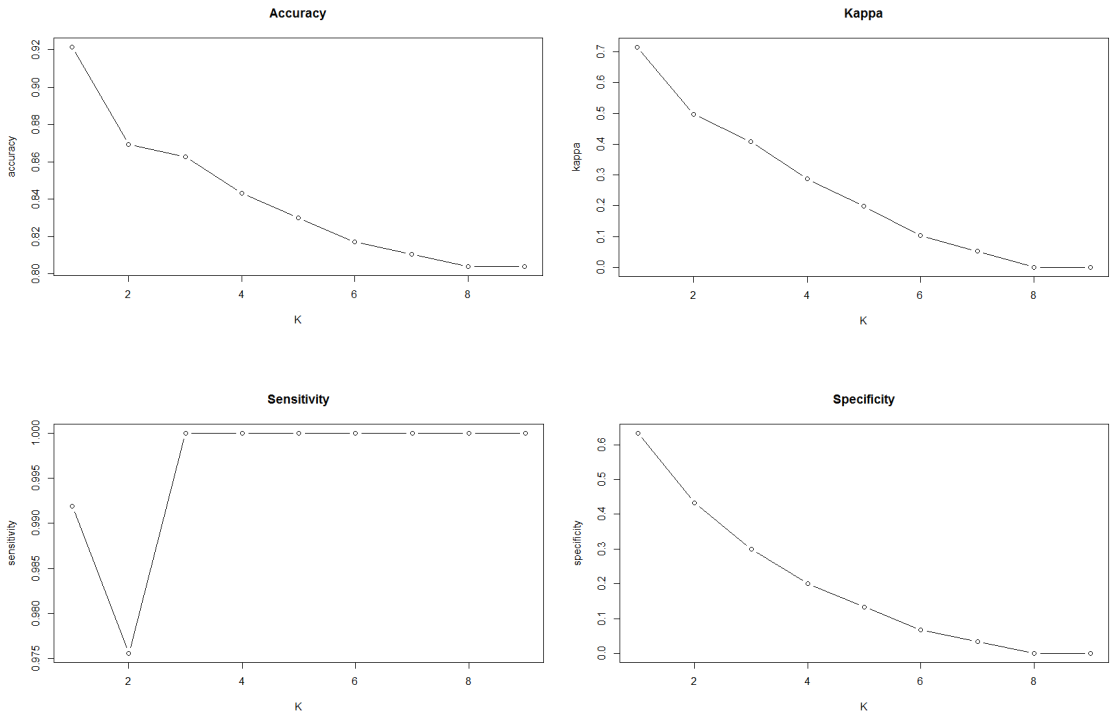

Figure 5-4. Change in performance measures of KNN with change in K

5.2.5 Rating method (RM)

Table 5-7. Confusion matrix of RM

Data class	**Rogue**	**Safe**
Classified as rogue	117	1
Classified as safe	6	29

Table 5-8. Statistics of RM

Statistic	Value
Accuracy	0.9542
Kappa	0.8635
Sensitivity	0.9512
Specificity	0.9667

5.4 Model selection

Table 5-15 describes the statics of different models that are developed. LDA, RM, SVM and KNN have accuracy greater than 85%. LDA and LR have the same specificity (76.67%) but LDA has higher accuracy and specificity than LR. KNN has the highest sensitivity (99.2%) followed by SVM and RM (95.12%). Kappa values of KNN and RM are greater than 0.7. Specificities of LDA, LR and RM are greater than 75%.

Table 5-9. Performance metrics of classification models

	LDA	LR	SVM (c=0.3)	KNN (K=1)	RM
Accuracy	0.8693	0.8366	0.8562	0.922	0.9542
Kappa	0.6145	0.5448	0.4781	0.715	0.8635
Sensitivity	0.8943	0.8537	0.9512	0.992	0.9512
Specificity	0.7667	0.7667	0.4667	0.633	0.9667

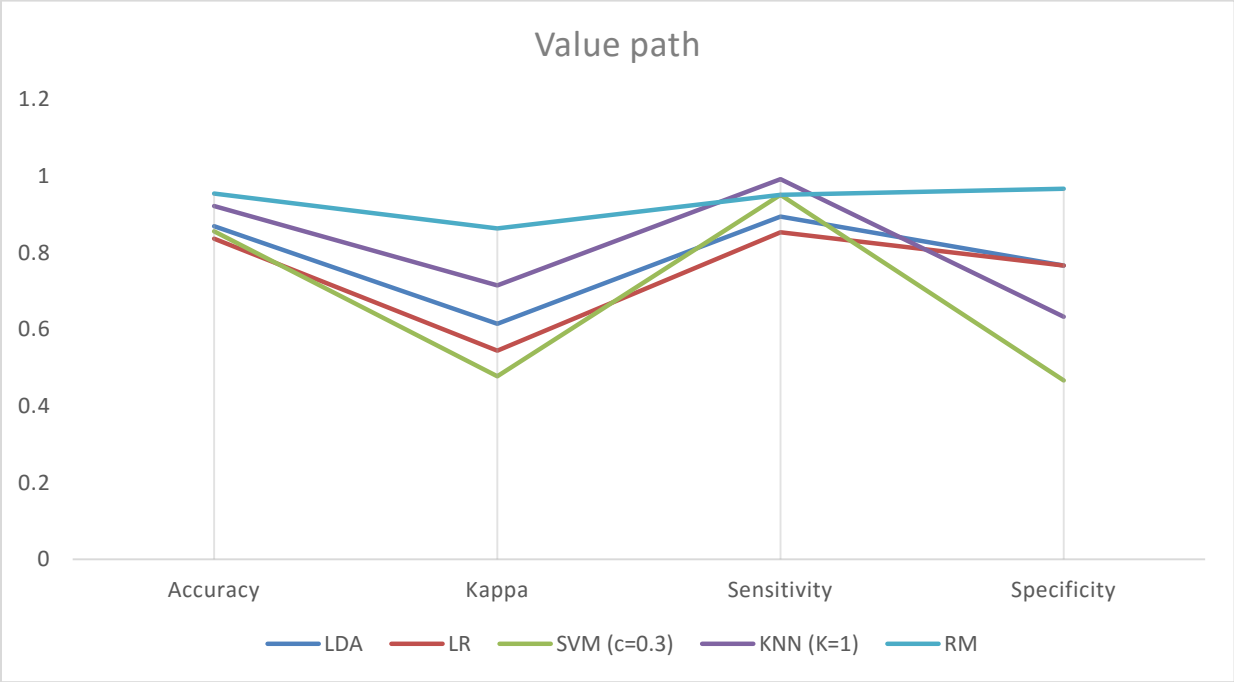

Figure 5-5. Value path of classification models

A value path of performance of models across accuracy, kappa, sensitivity and specificity was plotted. It can be seen that LDA dominates LR. RM has the highest accuracy, kappa and specificity and second highest is sensitivity. KNN highest sensitivity and second highest accuracy and kappa. Hence form the given dataset and analysis RM and KNN are recommended for test implementation in the representative dataset of online pharmacies.

5.5 Implementation on larger dataset

RM and KNN performed better across the different performance measures compared to LDA, LR and SVM. Hence, they are implemented on referral dataset II to check consistency. The models are trained using 67% of the data and tested on 33% of the data.

Table 5-10. Confusion matrix of RM (referral dataset II)

Data class	**Rogue**	**Safe**
Classified as rogue	374	11
Classified as safe	4	5

Table 5-11. Statistics of RM (referral dataset II)

Statistic	**Value**
Accuracy	0.9619
Kappa	0.3819
Sensitivity	0.9894
Specificity	0.3125

Tables 5-10 and 5-11 presents the confusion matrix and the statistics of RM on the referral dataset II. It can be seen that the performance of RM is poor on larger dataset. The kappa and specificity values are less than 0.4.

Table 5-12. Statistics of KNN (referral dataset II)

Name	**KNN (K=1)**	**KNN (K=2)**	**KNN (K=3)**	**KNN (K=4)**	**KNN (K=5)**	**KNN (K=6)**	**KNN (K=7)**	**KNN (K=8)**	**KNN (K=9)**
Accuracy	0.9873	0.9873	0.9848	0.9822	0.9721	0.9721	0.9645	0.9594	0.9594
Kappa	0.8085	0.8085	0.7618	0.7116	0.4659	0.4659	0.2151	0	0
Sensitivity	1	1	1	1	1	1	1	1	1
Specificity	0.6875	0.6875	0.625	0.5625	0.3125	0.3125	0.125	0	0

Table 5-12 represents the statistics of KNN on referral dataset II and Figure 5-6 represents the change in accuracy, sensitivity, specificity and kappa for different values of K. It can be observed that 100% sensitivity is achieved by all the models. 1-NN and 2-NN has the same values for all performance measures and dominates the other models.

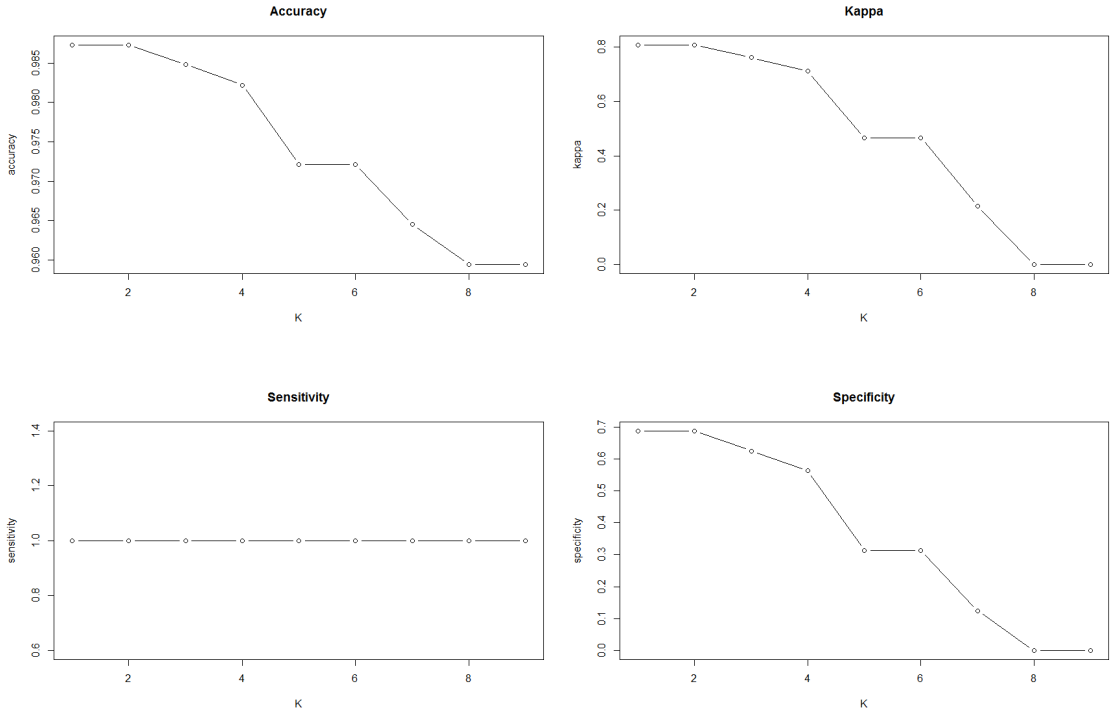

Figure 5-6. Change in performance measures of KNN with change in K for referral dataset II

Form the results it is evident that KNN classification model is appropriate for classifying online pharmacies based on the referral data. It performed well on both smaller and larger datasets and it can be implemented on large scale in the World Wide Web.

Usage and structure data for a list of legitimate and illegitimate pharmacies identified from NABP is gathered from SEMrush and Similarweb. Through web analytics the relative usage of illegitimate and legitimate pharmacies are studied. The network of referral websites of the online pharmacies along with their backlinks are used to develop an automated classification model. Linear classifiers are performed on the given data along an intuitive rating method and KNN. KNN has performed better than all the other classifiers both on small and large dataset. The conclusions and future directions for research are summarized in the next Chapter.

Chapter 6

Conclusions and Future work

Previous research shows that illicit online pharmacies are widely present and accessed. Also, they have tried to estimate the traffic attracted to the illicit websites from advertisements. However, the different ways in which the pharmacies are accessed (direct, search, referrals, social media, display ads and mail) and traffic estimates from each source have not been studied. Also, developing a classification system using web data has not been studied. In this study we have investigated these aspects by studying the traffic data for 30 legitimate pharmacies and 127 illegitimate pharmacies according to NABP. The traffic data for these websites are identified from Similarweb and SEMrush. Results from exploratory data analysis show that it is imperative to have a system that distinguishes legitimate pharmacies from rogue pharmacies. These might help the policy makers to formulate efficient policies and consumers to make informed decisions. In addition, social media sites and search engines should provide safe search options which will prevent such pharmacies coming up on their sites. Our key findings are

- United States is the biggest consumer for both legitimate and illegitimate online pharmacies.
- The overall engagement of legitimate pharmacies is better than illegitimate pharmacies.
- The major sources of traffic for both legitimate and illegitimate pharmacies are Search, Direct and Referrals.
- A small proportion of illegitimate pharmacies are attracting good paid traffic

- Illegitimate pharmacies attract higher traffic than legitimate pharmacies through display ads.
- Facebook, Reddit and YouTube are the top social media websites which attracts traffic into both legitimate and illegitimate pharmacies.
- Almost 30% of the traffic to illegitimate online pharmacies are Direct. This may be due to lack of awareness at the consumer end. On the other hand, consumers' intention to buy drugs without prescription might also be the reason. Both the cases might cause potential drug abuse.
- Classification system based on traffic and engagement data may not perform well.
- Developing online pharmacy classification based on referral data gives feasible results.
- Rating method and nearest neighbor based classification (KNN) performed better than other linear classifiers (Linear Discriminant Analysis, Logistic Regression and Support Vector Machines). Rating method performed well on the smaller dataset and KNN performed better than rating method in case of larger dataset.

The major limitations of this work include the size of the dataset for relative usage analysis. The number of websites taken into account is 30 legitimate and 127 illegitimate websites. These websites obtain the maximum traffic among online pharmacies, however considering other websites will improve the accuracy of the results. Rating algorithm performed well for classifying online pharmacies. However, the generalizability of it needs to be tested.

This thesis is among the first to attempt classifying online pharmacies. One of the possible future directions will be to explore what is sold at each of these pharmacies or anomalies in each pharmacy (FDA has the data on recalled drugs, expired drugs and banned drugs and pharmacies which sell them) and use these features to increase the granularity of classification we have developed.

Numerous websites are opened and closed daily and it is hard to keep track of them. Hence, considering dynamic data instead of static as in the present study is a better approach to classify online pharmacies. Also, Organic Drug based keywords from search queries can further be analyzed to identify the type of drugs consumers tend to purchase from the illegitimate pharmacies to discover buying patterns and consumer intent.

References

AMITAY, E., CARMEL, D., DARLOW, A., LEMPEL, R., AND SOFFER, A. 2003. The connectivity sonar: Detecting site functionality by structural patterns. In Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HYPERTEXT). ACM Press, New York, NY, 38–47.

CAWLEY, G. C., TALBOT, N. L. C. 2003. Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers, Pattern Recognition, 36, 2585 – 2592

CHAKRABARTI, S., DOM, B., GIBSON, D., KLEINBERG, J., KUMAR, S., RAGHAVAN, P., RAJAGOPALAN, S., AND TOMKINS, A. 1999. Mining the link structure of the World Wide Web. IEEE Computer, 32(8), 60–67

CHAPELLE, C., VAPNIK, V., BOUSQUET, O., MUKHERJEE, S. 2002. Choosing multiple parameters for support vector machines, Mach. Learning, 46 (1), 131–159.

COOLEY, R., MOBASHER, B., AND SRIVASTAVA, J. 1997. Web mining: Information and pattern discovery on the World Wide Web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)

Cox, D. R. 1958. The regression analysis of binary sequences (with discussion). J Roy Stat Soc B, 20, 215–242.

DESAI, K., CHEWNING, B., MOTT, D. 2015. Health care use amongst online buyers of medications and vitamins. Research in Social and Administrative Pharmacy, 11(6), 844–858

ETZIONI, O. 1996. The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11), 65–68,

FISHER, R. A. 1936. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7 (2), 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. hdl:2440/15227.

FITTLER, A., BŐSZE, G., & BOTZ, L. (2013). Evaluating Aspects of Online Medication Safety in Long-Term Follow-Up of 136 Internet Pharmacies: Illegal Rogue Online Pharmacies Flourish and Are Long-Lived. Journal of Medical Internet Research, 15(9), e199. http://doi.org/10.2196/jmir.2606

FIX, E., HODGES, J. L. 1951. Discriminatory analysis, nonparametric discrimination: Consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.

FOX, N., WARD, K., & O'ROURKE, A. 2005. The birth of the e-clinic. Continuity or transformation in the UK governance of pharmaceutical consumption?, Social Science and Medicine, 61, 1474–1484. doi:10.1016/j.socscimed.2005.03.011.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. 2013. An Introduction to Statistical Learning, 127-161, DOI 10.1007/978-1-4614-7138-7

GONDIM, A. P. S., & FALCÃO, C. B. 2007. Evaluation of Brazilian online pharmacies. Revista de Saude Publica, 41(2), 297–300. doi:10.1590/S0034-89102007000200019.

GORDON, S. M., FORMAN R. F., SIATKOWSKI C. 2006. Knowledge and use of the internet as a source of controlled substances. J Subst Abuse Treat, 30, 271-274.

GUO, G, LI, S.Z., CHAN, K. L. 2000. Face recognition by support vector machines. Automatic Face and Gesture Recognition. Proceedings. Fourth IEEE International Conference, 196-201

Lihua, Y., Qi, D., and Yanjun, G. 2006. Study on KNN Text Categorization Algorithm, Micro Computer Information, 21, 269-271

INCIARDI JA, SURRATT HL, CICERO TJ, ROSENBLUM A, AHWAH C, BAILEY JE. 2010. Prescription drugs purchased through the internet: who are the end users? Drug Alcohol Depend, 110(21), 9.

JENA, A. B. & GOLDMAN (2011), Growing Internet Use May Help Explain the Rise in Prescription Drug Abuse in the United States, Health Affairs, 30(6), 1192-1199

JENA, A. B., GOLDMAN, D. P., FOSTER, S. E., & CALIFANO, J. A. (2011). Prescription Medication Abuse and illegitimate Internet-Based Pharmacies. Ann Intern Med, 155, 848–850

JORDAN, A. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems 14, 841.

KOSALA, RAYMOND, AND BLOCKEEL, H. 2000. Web mining research: A survey. ACM Sigkdd Explorations Newsletter, 2(1), 1-15.

KUHN, M., JOHNSON K. 2013. Applied Predictive Modeling. 41-43

LANIER, W. L. 2004. Near-death experiences delivered to your home by your friends on the Internet. Mayo Clinic Proceedings, 79(8), 979–982.

LEGITSCRIPT. 2016. https://www.legitscript.com/

LEGITSCRIPT. 2016. The Internet Pharmacy Market in 2016. – Trends, Opportunities and Challenges. http://www.safemedsonline.org/wp-content/uploads/2016/01/The-Internet-Pharmacy-Market-in-2016.pdf

LETKIEWICZ, S., GÓRSKI, A. 2010. The Potential Dual Use of Online Pharmacies, Science and Engineering Ethics, 16(1), 59-75.

LIANG, B. A., & MACKEY, T. K. 2011. Prevalence and Global Health Implications of Social Media in Direct-to-Consumer Drug Advertising. Journal of Medical Internet Research, 13(3), e64. http://doi.org/10.2196/jmir.1775

LIANG, B. A., & MACKEY, T. K. 2012. Online Availability and Safety of Drugs in Shortage: A Descriptive Study of Internet Vendor Characteristics. Journal of Medical Internet Research, 14(1), e27. http://doi.org/10.2196/jmir.1999

LIANG, B. A., & MACKEY, T. K. 2012.Vaccine shortages and suspect online pharmacy sellers, Vaccine, 30, 105–108

LINDEMANN, C. AND LITTIG, L. 2006. Coarse-grained classification of Web sites by their structural properties. In Proceedings of the 8th ACM International Workshop on Web Information and Data Management (WIDM). ACM Press, New York, NY, 35–42.

LINEBERRY, T. W., & BOSTWICK, J. M. (2004). Taking the physician out of "Physician shopping": A case series of clinical problems associated with Internet purchases of medication. Mayo Clinic Proceedings, 79(8), 1031–1034.

LITTLEJOHN, C., BALDACCHINO, A., SCHIFANO, F. & DELUCA, P., (2005). Internet Pharmacies and Online Prescription Drug Sales: a cross-sectional study. Drugs: education, prevention and policy, 12, 1, 75–80

MACKEY, T. K., AUNG, P. & LIANG, B. A. (2013). Illicit Internet availability of drugs subject to recall and patient safety consequences. International Journal of Clinical Pharmacy.

MACKEY, T. K., & LIANG, B. A. (2013). Pharmaceutical digital marketing and governance: illicit actors and challenges to global patient safety and public health. Globalization and Health, 9, 45. http://doi.org/10.1186/1744-8603-9-45

MACKEY, T. K., & LIANG, B. A. (2013). Global Reach of Direct-to-Consumer Advertising Using Social Media for Illicit Online Drug Sales. Journal of Medical Internet Research, 15(5), e105. http://doi.org/10.2196/jmir.2610

MÄKINEN, M. M., RAUTAVA, P. T., & FORSSTRÖM, J. J. (2005). Do online pharmacies fit European internal markets? Health Policy (Amsterdam), 72, 245–252. doi:10.1016/j.healthpol. 2004.09.007.

MONTOYA, I. D., & JANO, E. (2007). Online pharmacies: Safety and regulatory considerations. International Journal of Health Services, 37(2), 279–289. doi:10.2190/1243-P8Q8-6827-H7TQ.

NABP (2016). http://www.nabp.net/about

PIERRE, J.M. 2001. On the automated classification of Web sites. Link¨oping Electron. Art. Comput. Inform. Sci. 6.

PURI, K., & DAMLE, P. (2007). Improving the online pharmacy experience. Available via Infosys Technologies Limited.

QUON, B. S., FIRSZT, R., EISENBERG, M. J. 2005. A Comparison of Brand-Name Drug Prices between Canadian-Based Internet Pharmacies and Major U.S. Drug Chain Pharmacies. Ann Intern Med. 143, 397-403. doi:10.7326/0003-4819-143-6-200509200-00004

RENCHER, A. C. 2002 Principal Component Analysis, in Methods of Multivariate Analysis, Second Edition, John Wiley & Sons, Inc., New York, NY, USA.

SCHIFANO, F., LEONI, M., MARTINOTTI, G., RAWAF, S. & ROVETTO, F. (2003). Importance of cyberspace for the assessment of the drug abuse market: preliminary results from the Psychonaut 2002 Project. Cyberpsychology & Behavior, 6, 405–410

SCHILLING, D. A., REVELLE, C., AND COHON J. 1983. An approach to the display and analysis of multi-objective problems. Socio-Economic Planning Sciences, 17(2), 57–63

SOKOLOVA, M. AND LAPALME, G. 2009 A systematic analysis of performance measures for classification tasks, Information processing and Management, 45, 427 – 437.

VAPNIK, V. 1995. The Nature of Statistical Learning Theory. New York: Springer-Verlag

VEROPOULOS K, CAMPBELL C, CRISTIANINI N (1999). Controlling the Sensitivity of Support Vector Machines. Proceedings of the International Joint Conference on Artificial Intelligence, 55–60.

WAISBERG, D. AND KAUSHIK, A. 2009. Web Analytics 2.0: empowering customer centricity. The original Search Engine Marketing Journal, 2(1), 5-11.

WEISS, L., GANY, F., ROSENFELD, P., CARRASQUILLO, O., SHARIF, I., BEHAR, E. 2007. Access to multilingual medication instructions at New York City pharmacies. Journal of Urban Health, 84(6), 742–754. doi:10.1007/s11524-007-9221-3.

YEASIN, M., BULLOT, B., SHARMA, R. (2006). Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans. Multimed., 8(3), 500–507.

YUAN, G. X., HO, C. H., LIN, C. J. (2012), Recent Advances of Large-Scale Linear Classification.