

Crowdsourcing Cybercrimes in Pakistan through Online Resources



MCS

by

Naila Amir

A thesis submitted to the faculty of Information Security Department, Military College of Signals, National University of Sciences and Technology, Rawalpindi in partial fulfillment of the requirements for the degree of MS in Information Security

August 2019

Declaration

I hereby declare that no portion of work presented in this thesis has been submitted in support of another award or qualification either at this institution or elsewhere.

Dedication

“In the name of Allah, the most Beneficent, the most Merciful”

I dedicate this thesis to my parents and family who encouraged me each step of the way and I would like to pay special regards for their love, patience and kindness.

Acknowledgments

In the name of Allah, the most gracious and the most loving.

Firstly, praises to ALLAH Almighty, for His blessings during my research work and leading it to successful completion. I would like to convey my sincere gratitude to my supervisor, Dr. Rabia Latif, for her kind supervision and providing continuous support throughout this research. It was a great opportunity and honor to work under her guidance. I would like to extend thanks to my committee members; Associate Prof. Dr. Haider Abbas and Lecturer Narmeen Shafqat for their support and acceptance during research discussions I had with them and thesis preparation.

Also, my sincere thanks to my parents; my father (Mr. Muhammad Amir) and especially my mother (Mrs. Siraj-un-Nisa) who helped me stay motivated throughout research task. I would also like to thank my husband (Mr. Sajid Mushtaq) for his continuous support and collaboration all the way during this research.

Finally, many thanks to all the people who were of great support for me to complete research work directly/ indirectly.

Abstract

Crowdsourcing refers as a practice of engaging a group of people for a common goal powered by enhanced technologies and social media. As millions of events happens every second of day, it's not possible for journalists to cover all of them. In context of Journalism, crowdsourcing, nowadays, provide a platform for general public to act and communicate thus improving their coverage.

Cybercrime is defined as aggressive, intentional act performed through electronic communications. In recent years, there is intense increase in racism, sexism and other types of aggressive cyber threats. As Pakistan is in list of fastest growing Internet-using countries, ultimately increasing the need to tackle cybercrimes. Confronting cybercrimes therefore requires development of robust detection method in unsupervised manner.

The focus of this research is to propose a model for detection of cyber offences in Pakistan reported through news articles and blogs. The first step is to compile existing cybercrime detection methods and perform comparative analysis of methods along with their weaknesses. Comparative analysis will assist in proposing a model for cybercrimes detection through electronic media. Further, in order to analyze authenticity of these news attained via news articles/blogs, source verification process is performed on acquired data by selecting various authenticity parameters and assigning weightages to these parameters.

Table of Contents

Declaration.....	ii
Dedication.....	iii
Acknowledgments	iv
Abstract.....	v
Table of Contents.....	vi
List of Figures	ix
List of Tables.....	x
Chapter 1.....	1
Introduction.....	1
1.1 Overview.....	1
1.2 Motivation and Problem Statement.....	3
1.3 Objectives	4
1.4 Relevance to National and Army Needs.....	4
1.5 Thesis Contribution.....	5
1.6 Thesis Organization.....	5
Chapter 2.....	7
Literature Review.....	7
2.1 Introduction	7
2.2 History of Crowd Sourcing	7
2.3 Crowd Sourcing Cybersecurity.....	9
2.4 Benefits of Crowd Sourcing.....	10

2.5	Crowdsourcing Applications in Healthcare/BioMedicine	11
2.6	Crowdsourcing in Disaster Recovery and Management	12
2.7	Crowdsourcing in Geographical Applications.....	13
2.8	Crowdsourcing in RandD	14
2.9	Crowdsourcing in Logistics	15
2.10	Crowdsourcing for National Security:	16
2.11	Challenges of Crowd Sourcing:.....	21
2.12	Conclusion.....	23
Chapter 3		24
Comparative Analysis of Event Extraction Techniques		24
3.1	Introduction	24
3.2	Event Extraction	24
3.3	Comparative Analysis of Event Extraction Techniques.....	27
3.4	Conclusion.....	34
Chapter 4		35
Design Methodology and Implementation		35
4.1	Introduction	35
4.2	Architecture of Proposed Model.....	35
4.3	Query Based Search Model.....	36
4.3.1	Categorization for Keyword Based Search	36
4.3.2	Information Extraction	37
4.3.3	Database Generation.....	47
4.4	Source Verification	50
4.4.1	Social Stats.....	52
4.4.2	Domain Moz Rank.....	52
4.4.3	Domain SEO Score.....	53

4.4.4	Domain Rank.....	54
4.4.5	Site Authentication	54
4.5	Conclusion.....	54
Chapter 5	55
Results and Analysis of Proposed Model	55
5.1	Introduction	55
5.2	Overview of Proposed Model.....	55
5.3	Defined Authentication levels.....	55
5.4	Feature Scaling.....	56
5.5	Data Analysis	59
5.6	Limitations.....	63
5.7	Conclusion.....	64
Chapter 6	65
Conclusion and Future Considerations	65
6.1	Introduction.....	65
6.2	Overview.....	65
6.3	Future Considerations.....	66
Source Code	67
Methodology 1:	67
Methodology 2:	69
References	74

List of Figures

Figure 1.1: CrowdSourcing Model	2
Figure 2.1: Types of CrowdSourcing	8
Figure 2.2: CrowdSourcing Transportation	9
Figure 3.1: Overview Diagram of Information Extraction	24
Figure 3.2: Information Extraction Approaches	26
Figure 4.1: General Model for Extraction and Verifying the Targeted Data	34
Figure 4.2: Major Categorization for Keywords	36
Figure 4.3: Technique Adopted for Creating Customized Scappers	37
Figure 4.4: Psuedo Code for Url Cleaning	40
Figure 4.5: Html Dom Analysis	41
Figure 4.6: Technique for Keyword Based Queries on Content Search Engines	42
Figure 4.7: Installation of Python Tools	43
Figure 4.8: Keyword Search and Results on BuzzSumo	44
Figure 4.9: Controlled Session of Chrome	45
Figure 4.10: Loading of BuzzSumo Website	45
Figure 4.11: Html Dom Analysis for BuzzSumo	46
Figure 4.12: Scrapped Data in Excel Format	46
Figure 4.13: Layout of Generated Database	48
Figure 4.14: Technique for Source Verification	51
Figure 4.15: MOZ Rank Checker	53
Figure 4.16: Website SEO Checker	53
Figure 5.1: Levels of Source Verification	56
Figure 5.2: Authenticity of Blogs Extracted in Database	60
Figure 5.3: Domain Graphs as per the Verification Levels	61
Figure 5.4: Blog Domains having Medium High Category	62
Figure 5.5: Blog Domains having Low Category	63

List of Tables

Table 1: Comparative Analysis of Proposed Model with Past work.....	34
Table 2: Authenticity Criterion.....	56
Table 3: Social Stat and Domain Ranking.....	58

Introduction

1.1 Overview

Things are about to change even faster as they are now, leading to human-centric future. Due to continuously growing connectivity, it has now become easier than ever for group of people to contribute towards a common goal. Crowdsourcing, often introduced as wisdom of crowds, is mixture of two words: “crowd” and “outsourcing”. It involves process of obtaining ideas/required services from group of educated online community. Nowadays, most of the companies are utilizing the concept of crowdsourcing as a tremendous business marketing tool as this allows them to attain services from relevant field experts through online social media. It results in acquiring best services at minimum cost, hence increasing company’s productivity with little or no labor expense.

Turning social media in social productivity is a common practice nowadays. Such initiatives are challenging business communities and organizations across the world to reinvent traditional and conventional methods. In field of Journalism, crowd journalism emerged as a starting point, providing platform to general public/citizens to share news and views. In such scenario, articles and blogs are gathered not from a single author but from multiple sources. It also helps to achieve target quickly and flexibly with help of crowd workers. Besides development in other platforms, crowdsourcing has also established alliance with cyber security.

The main idea behind crowdsourcing cybersecurity is to use crowd sourced platforms for solving public security issues and hence raising security awareness. For example, crowdsourcing platform such as online blogs are being widely utilized for communicating cybersecurity threats and vulnerabilities, such as web browser security and bugs. This technique of reporting cyber security related issues via crowdsourced platforms have been

massively used by a vast number of famous enterprises such as Facebook, Microsoft, Mozilla, Twitter etc. Well established companies like

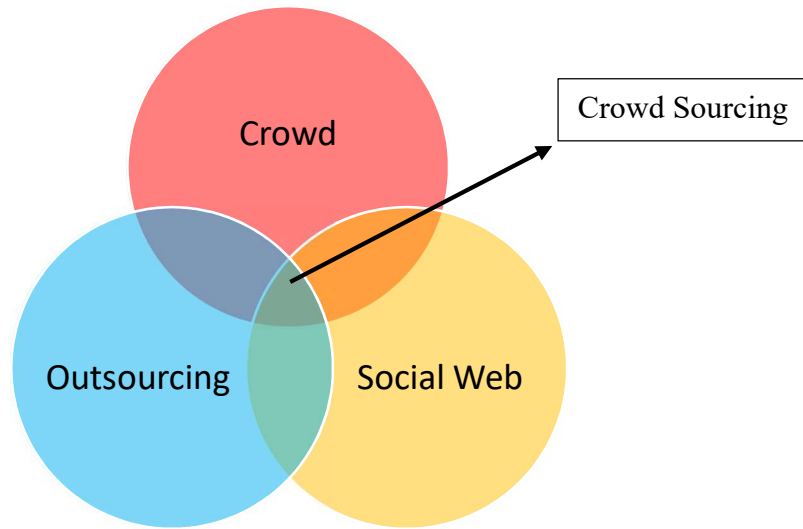


Figure 1.1: CrowdSourcing Model

Hackerone, Bugcrowd, Synack are managing bug bounty programs that offer rewards to white-hat ethical hackers for exposing vulnerabilities in computer networks. Synack co-founder says that this crowdsourcing model works more than as an internal security/audit team as several talented white hackers working for common target [1]. Famous example of crowdsourcing for security refers to twin bombing attacks in Boston where investigation data was collected via crowd sourced surveillance being provided by Marathon spectator's cellphone pictures, videos and feeds [2]. Within short span of time of bombing attack, Twitter, Facebook, YouTube and other social media got flooded with photos and videos, hence presenting huge amount of data for inquiry purposes.

Participation by general citizens in Boston Bombing initiated idea of illuminating country's security issues via crowd sourced approach. Usage of crowd sourcing platforms for solving public security concerns and raising security awareness is increasing day by day because of the fact that it attracts freelancers with various skills and expertise. Companies are dealing with their security issues by establishing information security strategies specifically formulated by power of crowd in general and concerned customer in particular [3]. Software houses/companies invites collaborators for exploiting vulnerabilities such as cross-site scripting, DDOS attacks, malwares, viruses and much more before launching a website or

software/applications. In return, after testing of software/website and evaluating impact of aforementioned issues, company delivers monetary rewards to such collaborators/white hat hackers.

1.2 Motivation and Problem Statement

Astounding enhancements in Information technology has precipitously revolutionized the world into a global village. One works online, plays online, shops online and cutting short, everyone is living online nowadays. Apart from benefits being brought by these advancements, there is another side of such technological elevations. Dangers associated with these digital revolutions are darker side of modern technological advancements. It has introduced breakthrough to vulnerabilities, threats, frauds and criminals in cyberspace. Privacy of individuals and organizations are greatly infringed by ease of access. Therefore, it is of great importance to elaborate the impact of internet penetration on conservative societies keeping in view the rules and regulations being formulated by government agencies for governing cyberspace. Prompt development in information technologies greatly relies upon well regulated cyberspace [4].

Being among the list of developing countries, Pakistan is confronting multifaceted cyber threats. Cyber-attacks have acquired all government and private sectors in Pakistan in last few years varying from minor website deficits to continual and endless cyber threats. According to Internet World Statistics, Internet usage in Pakistan were 17.8% by year 2016 while its 22.2 % by 2018 and is increasing day by day.

Pakistan is no exception among many other countries around the globe facing serious cyber threats. The main reason behind such an immense increase is mostly due to lack of technological awareness among general public, non-existence of legislation to financial limitations and poor collaboration with international law enforcement agencies [5]. Although government of developing countries are trying their best to fight against cybercrimes in order to protect country's cyber infrastructure and privacy, yet due to ever-changing aspects and dimensions of cyber criminals, situation is getting challenging for government to control rate

of cybercrimes. Moreover, various cybercrimes are not being covered under law by legislation and hence providing an edge to dynamic criminals.

According to “Hamara Internet” campaign launched by DRF (Digital Rights Foundation), 79% of young women use digital technologies on daily basis making them more vulnerable to cyber bullying and online harassment. In Pakistan, where anti-cybercrime laws are not strictly enforced, there is a dire need of cybercrime event extraction by analysing enormous bulk of data over various social media platforms.

1.3 Objectives

The main objectives of this thesis are:

- Identifying various event extraction techniques on different crowdsourcing platforms
- Conduct comparative analysis of various algorithms applied in event extraction techniques
- Propose a query based search model for extracting “cyber threats in Pakistan” reported through online resources
- Designing and implementing Source verification technique by using quintuple elements

1.4 Relevance to National and Army Needs

Cyber space is often conferred as fifth war-fighting domain after Land, Sea, Air and Space. Moreover, due to rapid increase in adverse effects of Cybercrimes (like cyberbullying) and its implications for child and adolescent development, investigations are beginning to emerge in scientific literature. Although, Pakistan has instigated several measures to get a hold over e-offences but rapid and effective steps are yet to be established and implemented. Nowadays, possibility of sabotaging systems security has leveraged due to distribution of internet based services especially at government level (e-g. NADRA, PTA etc.). FIA’s National Response Centre for Cyber Crime (NR3C) offices are located only in major cities and is a major

impediment to reporting cybercrimes. In order to take concrete action against dynamic cyber threats, cyber-crime monitoring should be made effective. In current scenario, cyber threats event extraction via social media platforms will help in improving such ease of inaccessibility resulting in efficient and timely reporting of cyber-crimes with minimum involvement of manpower. Moreover, Cybercrime event extraction model can aid Armed Forces agencies to detect defaming remarks regarding army and hence prevent such actions in future.

1.5 Thesis Contribution

It is stated to the best of our knowledge that event extraction model proposed in this research has not been promulgated in any paper and is solely presented after this research. Further, methodology of source verification of data collected via crowdsourced platforms has not been presented in any other research work.

The main contributions of this work are as follows:

- Various event extraction techniques have been highlighted based on crowdsourcing platforms.
- Comparative analysis of highlighted techniques is done for getting insight of their several features.
- An event extraction model is proposed for extracting cybercrimes through crowdsourcing platforms such as news articles.
- In order to acquire certain level of source verification of prepared database, a methodology is introduced using quintuple elements.

1.6 Thesis Organization

The thesis is structured as follows:

- Chapter 2 presents the literature reviewed in thesis. This chapter will provide use of crowdsourcing technique in different fields of everyday life (such as health facilities).
- Chapter 3 covers comparative analysis of various event extraction techniques presented by several authors. These techniques used various crowdsourcing platforms (such as twitter, etc) for extracting events.
- Chapter 4 presents design methodology and implementation of technique introduced as a result of this research.
- Chapter 5 comprises of results & analysis of proposed event extraction model.
- Chapter 6 presents future considerations of proposed research.

Literature Review

2.1 Introduction

This chapter contains literature review of various crowdsourcing platforms and use of crowdsourcing in several perspectives. It also provides a gist of implementation of crowdsourcing techniques in field of cyber security.

2.2 History of Crowd Sourcing

The word crowd sourcing has been derived from two commonly used words: “Crowd” and “Outsourcing”. Crowdsourcing refers to getting a task done through a large group of people in the form of an open call. The word crowdsourcing was first coined in 2006 by Jeff How in a magazine named Wired. According to Jeff How, accomplishing a task with help of labor isn’t free but it always costs less than getting the same task done by a traditional employee. Such an act of outsourcing task is known as crowdsourcing [6]. Although the term crowdsourcing has emerged in 2006 but it derives its roots since 18th century. In 1714, the British Government came up with sailing problem known as “The Longitude Problem”. Due to this problem, several of men were killed so government offered £20,000 for the people who could find out the solution of subjected problem. Issue was solved by a common man named as John Harrison and was awarded with £20,000. This was the first act of crowd sourcing. Similarly, in 1936, Toyota company got its logo designed via public contest. In 1955, Joseph Cahill from Australia, offered £5,000 to design a building besides a harbor in Sydney [7]. Till now crowdsourcing has emerged in various shapes and forms.

Crowdsourcing word has been defined by Daren C. Brabham in his book Crowdsourcing (2013) as an online production model that takes advantage of online communities for problem-solving purpose. Hence crowdsourcing leverages, the collective intelligence of general public

to serve a common cause or premediated goal [8]. The greatest strength lying under concept of crowdsourcing is its two-fold nature; it cultivates engagement with general public and on the other hand it opens up various skills to access. By taking into account the needs and rights of crowd/community, crowdsourcing not only appears as an easy way to gather data, rather, it is also a rewarding research process. Crowdsourcing is also being used in project funding for developing innovative ideas together with the crowd via process of brain storming. Nowadays, crowdsourcing has emerged in various forms such as sharing intelligence, Micro working, Macro working, Crowd funding, Mobile crowdsourcing etc. Micro tasks such as survey participation requires mental re-collection for performing tasks quickly. Macro tasks such as software testing, product design requires deeper thinking and intensive efforts for task accomplishment [9].

Generally, crowdsourcing has been divided into four types; Crowd creation, Crowd coting, Crowd funding and Crowd wisdom.

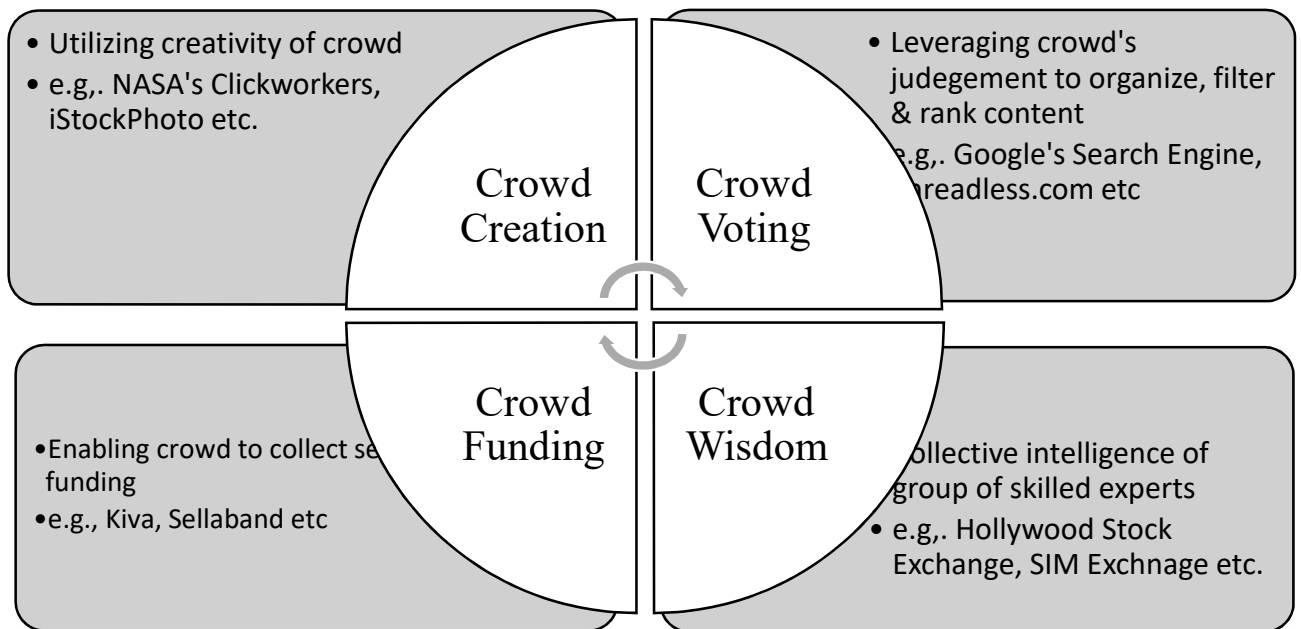


Figure 2.1: Types of CrowdSourcing

Crowd sourcing is an essential business idea; more likely to be referred as an e-Bay auction for talent. It's all about getting the right talent at the right time. For instance, it enables managers to complete business related tasks, that in traditional manner, an organization would either perform itself or outsourced it to third party vendor, by crowdsourcing it inviting a vast pool of talent to perform it while gaining insight into what actually customer wants. Uber is also an example of crowdsourcing as people who need to travel are paired-up with available drivers, which is a type of crowdsourced transportation.

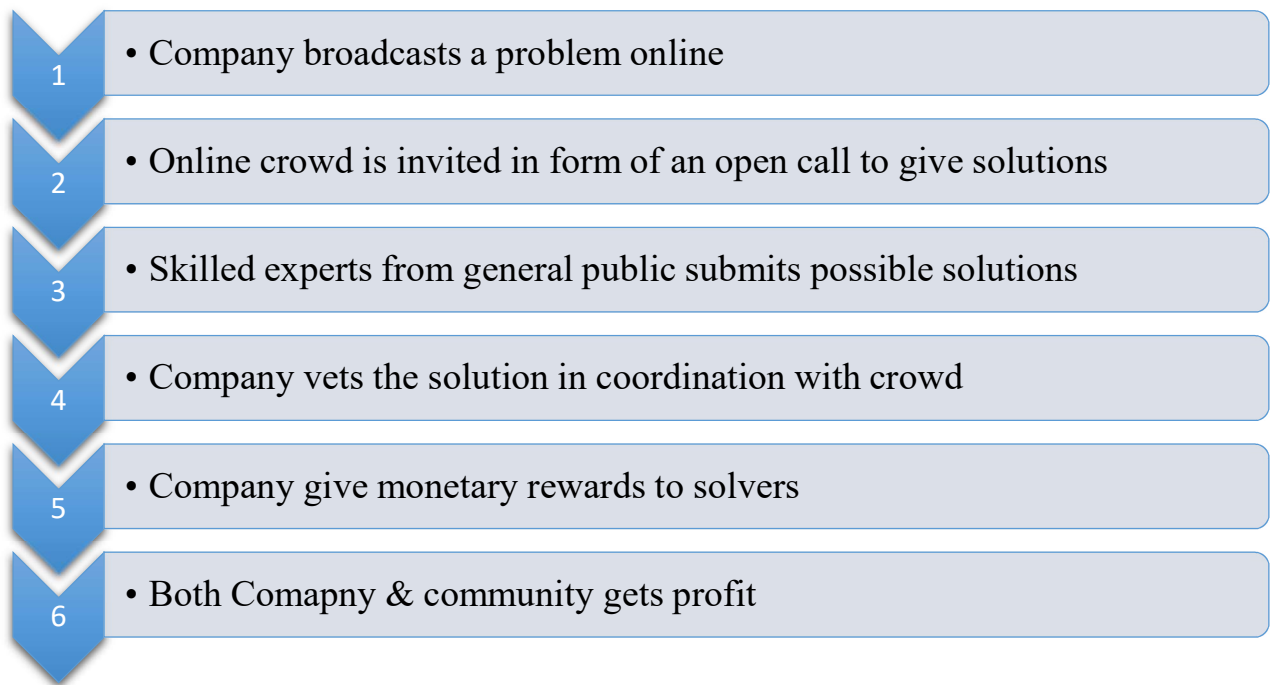


Figure 2.2: CrowdSourcing Transportation

2.3 Crowd Sourcing Cybersecurity

Modern world is technology-driven where information disseminates like wildfire within seconds. Johann Wolfgang termed the well-known phrase, “precaution is better than cure” and is more applicable in field of cyber security. Currently, internet is flooded with crowdsourcing platforms, providing a link between organizations and highly skilled individuals ready to lend their services either voluntarily or by demanding monetary rewards. Security is considered as an upcoming task in crowdsourcing [10]. Cyber threats are emerging in various forms affecting corporate, healthcare and several other social sectors of economy. Identifying loopholes before

malicious hackers can save us from disastrous data breaches and reputation damage. Crowdsourcing has aligned itself with cybersecurity and has generated fruitful results. The core concept behind aggregation of crowdsourcing and cybersecurity is to bring together the finest minds with the best advanced technology and practices to present an impartial finding towards vulnerabilities, and re-mediating these vulnerabilities timely and effectively.

Encountering surfacing cyber threats demands accumulating skilled experts to gather on a common platform and fight off against these threats with help of advanced technologies. Three active forms of cybersecurity are as under:

- **Collaborating for a specific cause:** In cybersecurity perspective, cyber security experts and skilled professionals collaborate to fight against upcoming threats.
- **Sharing Intelligence:** It involves combining individual skills and government/private entities.
- **Bug Bounty Program:** This form of crowdsourcing invites white hat hackers to perform audit such as vulnerability assessment and hack their security intentionally in control environment.

2.4 Benefits of Crowd Sourcing

Social media platforms are generally used for performing crowdsourcing. These platforms usually include open-sourced digital platforms; easy to access for every citizen. Implementing crowdsourcing via social media basically turns consumers into brand advocates by getting them involved. Modern internet advancement, development in collaboration tools, and web 2.0 technologies serves as key driver of this transition from traditional methods towards external stakeholder's engagement. Key elements of crowd sourcing can be partitioned into crowd, outsourcing and social media such internet platform. More precisely, crowdsourcing occurs at intersection of these three key elements.

Nowadays, with increase in tough modern corporate conditions, in order to keep up with the pace, businesses and organizations are compelled to step up at various levels. Widely adapted methods inculcates hiring best marketers, having manpower with diverse skills and deeper

public relations. From designing and vetting product ideas to work out for solving research and development problems, citizenship will surely help to perform better in marketplace. Internet facilities are being used to attain solicit feedback from active customers, as a result reducing time required for collection of data through formal focus groups.

With passage of time, Internet-based applications have facilitated the exchange of user-generated content, hence transforming users from consumers to content producers [11]. Wikipedia is one of the famous example of crowdsourcing. Instead of developing whole content of encyclopedia by themselves, Wikipedia allows audience to create content [12]. Waze is a successful application powered up by crowd. This app grant access to users to update online traffic conditions such as traffic jams and predicting best route to follow in similar situations. Due to enhanced and broad availability of technological proficiencies, digital strategies such as use of social media and crowdsourcing have great potential for use in timely recognition, care delivery and treatment and outcomes in healthcare emergency situations like cardiovascular scenarios. In similar way, it can be used in other clinical practices and medical interventions [13].

2.5 Crowdsourcing Applications in Healthcare/BioMedicine

Nowadays, vast array of medical research has been done through crowdsourcing platforms. It is not only limited to health research but it is also now being used in various health promotion/care activities. In field of medical research, Community based participatory research (CBPR) has made a prosperous track record such as preventing HIV, care and treatment. By involving public community at multiple stages, challenge contests are increasingly used for providing health equity and engaging community. Challenge contests gathers the wisdom of communities. Descriptions for specific health research challenges are declared in form of an open call via social media platforms. In response of call, contestants provide a descriptive solution and afterwards all acquired descriptive entries will be judged by expert jury members of committee. BioGames project was initiated with purpose of diagnosing red blood cells effected with malaria [33-35]. Mostly, process of quick diagnostics of Malaria is expensive and unreliable in developing countries. BioGames is an Android app and available online. In game, a brief tutorial was given to the user and then asked to play the game. The

game involves a syringe for purpose of killing malaria infected cells and collect healthy ones. The author believed that by involving huge crowd of people, hybrid algorithm involving concepts of crowdsourcing and machine learning will perform optimally better.

Crowdsourcing has also been utilized in assisting physicians for early diagnostics. Various platforms have been introduced for this purpose such as Amazon Mechanical Turk (AMK) for accessing different levels of cases. Another mobile application named as DocCHIRP (Crowdsourcing Health Information Retrieval Protocol for Doctors) helped solving clinician's problems. DocCHIRP application helped clinicians for managing patients routine care, as well as diagnosing mystery cases. Crowdsourcing methodology is also benefiting surveillance and participatory tasks in field of health care. Ushahidi is another open sourced mobile application with the ability to gather individual reports via email, SMS or other social media forums. It can classify, translate as well as geotag results. Keeping in view the literacy problems, Ushahidi includes feature of gathering voice messages/reports. Similarly, in order to track Asthma attacks, Asthmapolis (GPS-enabled inhaler), is presented for generating risk map of environmental triggers by compiling data attained from users using this inhaler application [36].

Biocuration is a term in field of biomedicine that refers to extracting key biological information from literature and other available resources and stored them in form of structured database. These databases can then be examined systematically. Several approaches have been proposed combining text mining and crowdsourcing for curating drug indications. Crowdsourcing approaches in bio medicine have surpassed variety of research activities such as dealing with complex biological structural problems, acquiring crowd judgements for curation of databases etc.

2.6 Crowdsourcing in Disaster Recovery and Management

Crowdsourcing methodology has also been widely deployed for purpose of information gathering and handling during disasters and mass emergencies. Crowdsourcing, along with advancements in social media, has fueled a new landscape of emergency and disaster response systems. During critical emergency situations, these systems allow affected citizens to generate

real-time information. These initiatives got generated by several domains involving no-profit/voluntarily organizations, government agencies and technical associations. Digital Operations Centre is one such kind of social media platform, launched by American Red cross in 2012, aimed at providing relief to humanity.

Similarly, Government Crisis Coordination Centre has been established in Australia with purpose of providing ease of information access, regarding developing situations during natural disasters, to crisis coordinators. One of such incidents include recovery efforts implemented via crowdsourcing projects after catastrophic Haiti earthquake. Information attained through massive crowd helped victim of calamity to find shelters, medicines and other medical necessities [14]. ESA (Emergency Situation Awareness) is an Australian organization focused on providing potential information regarding real time data of incidents or warning/alerts collected via community responses. NetQuakes is a projected based on communication based crowdsourcing initiated by U.S Geological survey. NetQuakes aims at gathering uniform spacing measurements via seismograph capable to operate on WiFi. Therefore, vastly deployed information and communication technologies paved a path for digital volunteer networks to contribute towards disaster recovery during crises [15].

2.7 Crowdsourcing in Geographical Applications

Due to the pervasive availability of mobile phones along with their ability to provide live position using several techniques such as Global Positioning system (GPS) or WLAN (Wireless Local Area Network), there arise a new origin of location based crowdsourcing. Citizens have now also become a quick and influential source of geographical information. Such volunteer action of providing geo-graphic information is often referred as “Geo-crowdsourcing” [16]. Massive proliferation of smart phones along with internet connection have given rise to various geo-crowdsourcing mobile applications. These mobile apps allow people to act as living sensor network creating, assembling and disseminating geographic data. Crowdsourcing is interesting from scientific point of view. It can be utilized to acquire crowd to run specially designed measurement applications for performing distributed network analysis. Noisetube is an example of such app. Noisetube allows citizens to use their GPS-equipped mobile phones as noise sensors enabling citizens to share their personal exposure to

noise in daily routine. By collection of geo-localized measurements from large group of people, one can produce collective noise map [17].

Waze is crowdsourced based mobile application which is being widely utilized for traffic and navigation tasks. The user is requested to enter hi/her desired destination and is navigated via application. During same, current road traffic data situations will be shared with others such as traffic jamming situations. It allows other users to edit information (in a similar way like Wikipedia) and update places such as stop and search points and accidents. Moreover, Waze was also being awarded with best mobile application [37]. Trapster is a well-known mobile application similar to Waze. Trapster is freely available application compatible with Android and other Windows phone gadgets. It provides current driving conditions by collecting data through GPS even if its' in passive state. Drivers use this application for warning traffic scenarios like wrong-way drivers or speed cameras.

Ushahidi is also famous example of mobile crowdsourcing using LBS data. Usahidi is an online platform in fact a website itself where information is retrieved from regions suffered from catastrophic crisis and after gathering sufficient data, same is represented in the form of visualized maps. With the help of these visualized maps, coordinators of disaster and recovery management take on overview of current situation and decides where instant support is required.

2.8 Crowdsourcing in R & D

Crowdsolving allow users to execute certain tasks in fields such as R & D department or data analysis. Colgate-Palmolive crowdsourced idea of developing mechanisms of injecting fluoride in tooth pastes. Similarly, NASA calls for an open suggestion of algorithms for detecting distortions in galaxy images caused because of dark matter. Crowdttesting utilizes human cloud for conducting analysis on user perceived quality or software evaluation like Google or BBC often perform usability, availability or load tests. Crowdflower is a platform for obtaining users/ crowd opinion regarding certain product quality [18].

Kickstarter is one of the widely used application of crowdsourcing that use services of location based services, also known as, LBS. Kickstarter is an internet platform famous for its crowdfunding applications. Crowdfunding, a form of crowdsourcing, inculcates funding projects where funds, in the form of monetary contributions, are collected from third parties for successful completion of project. Kickstarter publicize a project initiated by capital seeker whereby general public or crowd is boosted up for donation of money. For reaching out to maximum number of audience, Kickstarter has also launched a mobile application. Anyone with a registered account can contribute towards the project. The capital seeker get its project registered at Kickstarter forum and specifies the required monetary contribution required for project as well as sets time frame in which specified money needs to be arranged.

2.9 Crowdsourcing in Logistics

Crowdsourcing manifold numerous opportunities in fields of logistics. Process of logistics includes planning, implementing and controlling procedures. Logistics generally gets itself distributed into *classic logistics* and *information logistics*. Classic logistics refers to efficient flow of goods (information and services) along the entire path i.e. starting from the point of origin to the point of consumption that belongs to conformation of customer requirements. Similarly, information logistics belongs to controlling information handling process optimally. The term “crowd logistics” emerged covering both concepts of classic and information logistics. In crowd logistics, companies used to outsource their tasks to individuals referred as crowdsourcees.

Crowd logistics is generally handled in two different ways: *Tournament based crowdsourcing* or *Collaboration based crowdsourcing*. In tournament based crowdsourcing, a user designs and implements its solution to declared problem and gets it submitted to one who has called for solutions. For instance, the person/user who opted first for delivery of parcel from source to destination will be given the task. Whereas in collaboration based crowdsourcing, group of individuals’ workout together for solving a problem. When peer-peer crowdsourcing platforms, such as marketplace, are used for provision of services, individual task matching approach is utilized by both individuals/organizations for advertising jobs. The purpose is to re-distribute short terms jobs to self-employed people like freelancers. Checkrobin is a kind of

crowd-logistics application that is widely used platform for transport services. The main idea behind developing this application is that a number of cars cover their journey with lots of loading space being unused. This app offers ride-sharing service for private shipments. A customer will be cross-linked to a driver (crowdsourcee,) on the basis of perfect matchings, who will send or receive the shipment as they will cover the desired route in accordance with shipment destination.

Crowdsourcing is embedded in social media, allow governments, NGOs, businesses and organizations to retrieve knowledge and expertise from vast undefined pool of open public. As a result, crowd capital increases leading to competitive advantages.

2.10 Crowdsourcing for National Security

Cybercrimes are often referred as off spring of cyber space innovations. With the increase in continuous economic growth and development, the need of communication has also increased at rapid pace. Due to these advanced information system, service mechanisms have also got revolutionized. Therefore, traditional trends of manually provision of services have now been disappeared and replaced with online provision of services using information technology systems. The world has now become a global village with providing services such as E-learning, E-banking, E-business and E-government virtually, i-e., in short it can be defined as face to face without confronting face to face [38]. Apart from numerous blessings offered by information system technologies, it has also introduced several dangers associated with digital technology. Among these associated dangers, the most critical and biggest issue that the developing countries is confronting is cyber security against continuous evolving cybercrimes.

Cybercrimes have joined the community since long. According to [39], due to the increase in global revolution by information system technologies, cybercrime has become an obvious form of international crime. Cybercrimes, in contrast to traditional terrestrial crimes, are easy to learn and commit using few number of resources, and potentially cause huge damage. Another issue of cybercrime is that it can be committed virtually, without being available physically in attack destination/jurisdiction and hence often goes unpunished because of not being clearly illegal [40]. Hence, unfolding nature of cybercrimes offer new challenges to law enforcement

agencies as well as to international lawmakers [41]. Developing countries are not too familiar with these advanced technologies and this slacken progress has launched several cybercrimes. Since many years, in order to combat these plague like spreading cyber threats, both public and private sectors have been working individually for coming up with latest tools and technologies. But due to continuously evolving nature of cybercrimes, these methodologies have failed to keep up with the pace of cybercrimes tactics. Therefore, Cyber space is often dubbed as fifth war-fighting realm, the other being Water, Land, Air and Space.

Crime remains ambiguous and illusionary and since its advent, it always strives to disguise itself in face of development. Cybercrimes needs to be monitored, controlled and prevented via well-established rules of cyber legislation. Absence of proper legislation to financial restraint is one of the major barrier in way of establishing users trust and confidence. Law enforcement agencies have also gone crimped by non-competence of cyber legislation rules to pace up with rapidly changing technological curves. Due to absence of cyber law, criminals are continuously playing with data privacy of common people, in some cases it can cause severe damage to one's social and family life, whereby individuals are becoming an easy prey towards these diverse crimes [42]. With the increase in Internet usage globally, hacking, spamming, spoofing, phishing, money laundering, child pornography, cyberbullying, corporate espionage are few names among vast range of emerging cybercrimes. Multinational companies have also been relying upon Internet to reach out their audience for advertising and selling goods/products. In 2010, computer systems that have been utilized by Siemens company for monitoring of infrastructure got attacked by a virus named as Stuxnet. The virus caused great havoc in Iran as it had got control over main computer systems of nuclear facilities. After creating devastation in Iran's nuclear systems, Stuxnet moved towards other countries attacking similar computer systems. Within short span of time, Siemens released a tool for detection and removal of Stuxnet virus for enhancing security of its products.

Nowadays, keeping in view the importance of public engagement, their role in combating cyber threats cannot be overlooked or underestimated. Due to huge availability of data on social media platforms, utilizing data mining and advanced technological tools can solve security threats. With help of social media forums, crowdsourcing operations spread like a fire

among huge audience within a short span of time, also making it easy for them to collaborate and communicate. Moreover, due to ease in generating content on social media provided by Internet, users have now been transformed from consumers towards content manufacturer. Users also get attracted towards social media forums because of captivating multimedia nature involved in it [43]. Crowd can provide quick and expert solution for declared problem sometimes solutions turned out to be innovative. National security demands element of confidentiality in its matters whereas customized crowdsourcing platforms serve as reliable medium for collecting discreet intelligence and access to sensitive data can also be maintained in these forums. For example, in 2011, during riots in U.K., general public were called for uploading pictures/snapshots of those suspected of looting. Citizens were invited to upload pictures on specially designed smartphone application named as FaceWatch ID. Nearly 2800 pictures were acquired via application and authorities were directed to catch if they recognize someone among sorted photographs.

Application of crowdsourcing in law enforcement can also be scrutinized by establishment of Texas Virtual Border Watch in Mexico. It involves installation of surveillance cameras at Mexico Border [44]. These cameras were based on government websites and anyone with access to Internet can watch surveillance videos and can report prohibited activities like illegal immigration etc. British government is also known as one of the colonizer of implementing crowdsourcing for purpose of national security. GCHQ (Government Communications Headquarters) is one of the intelligence agency of Britishers. In 2011, GCHQ launched an interesting project with the purpose of recruiting intelligence officers for signals core. The project includes solving a code comprising of 160 numeric letters. Those participants who were able to solve the code were redirected to website alongwith a keyword where they were able to apply for subjected post. The main aim was to recruit employees with specialization in cracking codes [45].

In Pakistan's perspective, citizen's private sensitive data is now being digitized specifically via government organizations such as NADRA and PTA. These initiatives have been taken primarily to bring state on a level equal to international developed countries, but due to lack of technological awareness, several risks have also been initiated alongwith numerous

advantages, the more important one being the data privacy. As it was reported via Edward Snowden leaks that huge amount of sensitive data from major telecom companies in Pakistan have been accessed illegally by US NSA. Moreover, it was also spotted out that US CIA agency is archiving Pakistani citizen's biometric data continuously with the aim to incorporate the acquired data in its TIDE (Terrorist Identities Datamart Environment). TIDE agency works out in collaboration with DHS (Department of Homeland Security) for achieving national security. Edward Snowden also revealed that CIA and NSA are retrieving Pakistan national's sensitive data on regular basis via illicit means. Such a huge amount of data extraction cannot be performed intelligently without involvement of human intelligence. Among the most famous cyberattacks, Estonia's state attack was one of them. Similarly, like many of the developing countries, the nation's analytical framework was integrated with Internet. The Estonian's attack hits its organizations websites inculcating banks, parliament, ministries etc. The attacks crippled the nation entirely, damaging its critical infrastructure. All such scenarios lead to the conclusion that merely implementing identity and audit rules, will not be capable of providing required security. It requires that policy makers must take feedback from the common people [46].

Being in the category of developing countries, Pakistan has still a lot of room for improvement of cyber services. In 2014, cyber criminals started to hit Pakistani websites pertaining to armed forces and federal government agencies via DDoS (Distributed Denial of Service Attack). Due to lack of trained cyber security professionals and experts, FIA can't barely cope up with growing cyber threats. In the year of 2018, 4906 complaints got registered by NR3C. While among the cases registered in 2017, pending complaints were 1190. Hence, Cybercrime wing CCU had dealt with 6096 cases. Out of these 6096 complaints, only 273 among them got conversion into cases. 1557 enquiries were already got closed in last year and still 4266 lie still pending [47]. Several registered complaints comprise of harassment cases. In current situation, women have been the central goal for cyberbullying. Cyber bullying/cyber stalking or harassment has become a worst fear among teens these days [48]. Digital Rights Foundation, often termed as DRF, is well known NGO and has been working since 2012. DRF is launched by Nighat Dad with purpose to ensure safety of people on social media forums. In 2016, they also initiated a helpline concerning cyber harassment. Statistics depicts that women are

targeted more with cyber harassment as compared to men as in year of 2018, calls received (on cyber harassment helpline) by women comprises 59% of total [49].

To counteract these growing rate of cybercrimes, it was suggested by a Tanzanian Delegation to initiate Cyber Crime Unit (CCU) for tackling cyber crimes. Further, the delegation suggested to establish a concerned legislation and in order to facilitate its implementation, it will require to formulate CERT (Computer Emergency Response Team) [50]. Pakistan needs to establish national and sectorial CERTs and government should encourage collaboration between them [51].

The strategy of crowdsourcing cyber security could be the possible and reliable solution to resist cybercrimes [52]. We all know that two heads are better than one. In process of crowdsourcing, there are even thousands of millions of heads joined together for a specific cause and it could be achieving cyber security for national cause. There is a need to develop an innovative approach towards e-participation. According to this approach, government agencies would adopt methodology of passive crowdsourcing, hence, exploiting user generated content on social media like blogs, news and online platforms. In this way, citizen's needs, opinions, arguments regarding numerous government policies will reach out to government agencies without their stimulation [53]. Cyber-criminal activity came up with darker side of internet and it needs to be dealt via combined efforts. Nations should put their efforts to address upcoming cyber related issues by building capacity among them. The lack of concerns that have been put by government officials and agencies regarding citizens' safety and privacy is one of the major factor that have affected the citizens' trust level on utilizing existing services provided by government for tackling cyber issues. Most of the people especially women don't approach intelligence offices such as FIA because they don't believe that agencies will do something beneficial for them or the agencies would disclose citizen's identity leading them on the path of social embarrassment.

In short, there is no magical or short cut solution towards cyber defence. Cyber defence mechanism involves quick and affective response to actions and assuring protection to nation's critical infrastructure, government entities and various networks. It focuses on prevention,

timey detection and responding towards cyber threats. Cyber security is distributed problem by inheritance. Therefore, the designed national security approach by government should not be focused entirely on implementing level of security on Internet or just to publish legislation [54]. Cyber security should mainly be implemented via crowdsourcing platforms as they are easy accessible as well as understandable to every common man. Crowdsourcing cybersecurity will also raise awareness among people. It is one of the best technique to attract skilled professionals and cyber security experts to speak up and serve the nation by utilizing their skills and gaining recognition and monetary rewards in return.

2.11 Challenges of Crowd Sourcing:

Crowdsourcing leads us to the road to open innovation that goes outside boundaries of organization to solve problems and hand off ideas to partners. However critical question arises over legal and ethical issues while using crowdsourcing platforms. Platforms used for crowdsourcing are generally open source digital platforms. Although positive impacts of social media are recognizable, but a new chapter has also been opened alongwith evolution of social media. Crowdsourcing face many challenges such as information security and project management. Despite of several benefits associated with connecting to general public, there lies a risk of connecting to social media networks. Further, due to little regulations regarding crowdsourcing usage, it leads to certain ethical or legal issues.

Due to availability of data on social media forums, hackers can collect the personal data and may use it for some illegal tasks. In order to avoid data exploitation, citizens should have limited access to personal information of other people. Protection of intellectual property rights is another major concern that arises while using crowdsourcing. For example, Products and designs being designed by crowd have high risk of infringement due to involvement of huge number of unknown participants from around the globe. The process of crowdsourcing relies on voluntarily participation from crowd and therefore, achieving target mass contributors cannot be guaranteed. Process of crowdsourcing starts with formulation of problem statement. It is one of the crucial part of crowdsourcing phenomenon. Having an ill-defined problem

statement will definitely lead towards less quantity of skilled contributors and results of crowd solving task will be not up-to desired expectations.

Scalability is another issue that we face during crowdsourcing ideas. Due to large coverage and ease of access, hundreds of thousands of users contribute towards crowdsourcing platforms such as social media networks. Therefore, scalability and availability of these social platforms are necessary for efficient and smooth running of crowdsourcing process. Security concept of social media forums and crowdsourcing are generalizable due to the fact that in social media, content is generated by a group of people while in crowdsourcing, content is generated by huge crowd. Therefore, security perspectives need to be more focused when it comes to using social media platforms in crowdsourcing. With rapid pace of development in existing social media platforms, user's participants get worried about security problems like confidentiality, privacy of personal information. To encounter this, strong enhanced security mechanisms are required to be implemented in order to avoid crowdsourcing attacks. Understanding complexity of human issues and criticality of manpower participation can minimize security risks.

Another challenge being faced is to recognize between real and fake users/participants. Due to the fact that users participate in crowdsourcing task through internet, therefore, one cannot depict the intention of user. Fake IDs and malicious intentions of participants affect the efficiency and security of task being executed. To combat the occurrence possibility of this threat, one should implement some sort of security mechanism for supervision of users before entering into the system. Creating government websites with formal license and asking individuals to enter personal details upon receiving security code are some of the methodologies that can restrict fake users to some extent. Implementing cryptographic algorithms during financial transactions of monetary rewards is another way to prevent hackers from malicious/false transactions. Inserting security plug-ins such as CAPTCHA technique and forcing users to change system defaults according to security instructions will also help in reducing security threats.

2.12 Conclusion:

Due to novel and innovative nature of crowdsourcing, classification and evaluation of frameworks are required to counter act the security risks. Therefore, several researchers present the idea of proper governance as key towards success of crowdsourcing mechanisms.

Comparative Analysis of Event Extraction Techniques

3.1 Introduction

This chapter describes the concept of event extraction followed by comparative analysis of various event extraction techniques. It also provides a gist of rise of cybercrimes globally and Pakistan in particular.

3.2 Event Extraction

With rapid pace in digital evolution, a huge amount of information is nowadays available on the web, also sometimes known as Knowledge Database. Extracting suitable information from bulk of data available on social media platforms can be utilized in several ways such as decision making processes etc. As the amount of data produced by citizens' increases, the more it requires to get intelligently processes for extracting desired information. Information available exists in various forms and sources, from news articles and blogs towards comments and conversation. Information extraction is the process of extracting contents of interest from tremendous amount of structured/ unstructured data. The main task lies in finding out relevant data and then storing it in pertinent from fro future use.

Nowadays, Social media platforms are serving for detecting trends of real-world events. Due to availability of enormous amount of data on social media, researchers collect bulk of data from these platforms in continuous time intervals and then perform various algorithms and techniques to acquire desired trends satisfying their information needs. From business perspective, event extractions from business news helps investors and marketers to perceive upcoming marketing trends and strategies. This will also help them to make valuable decisions.

The automatic extraction of information through social media has paved new pathway for organizing data and analyzing it, thus attaining clean interpretation of well-organized structured databases rather than having plenty of unstructured data. It has turned our personal desktops into structured knowledge databases. As a result, researchers and scientists are finding new ways for information retrieval by using machine learning and computational linguistics approach. Techniques keep on evolving on basis of variety of input resources being used, features of extraction task and the type of output produced [20]. An overview diagram of information extraction is shown in Fig. 3.1.

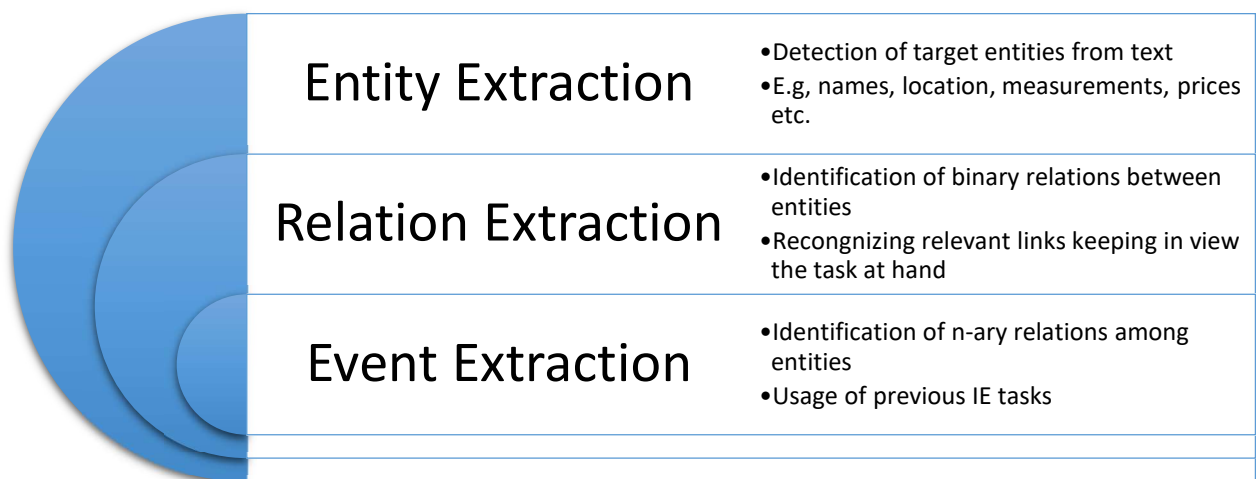


Figure 3.1: Overview Diagram of Information Extraction

Information extraction has a wide range of applications. It first involves the collection of information from various sources such as web pages, reports, scientific papers and legal documentation etc. Afterwards, various data mining techniques and algorithms are implemented on collected information to get the relevant data. The technology proves itself very successful in field of content management and knowledge discovery. Managing health care records, business intelligence (for gathering and analyzing structured information), financial investigation and scientific research are some of the applications of information extraction.

There are two primary approaches to design an Information Extraction system. The first one is *Knowledge Engineering Approach* while the other one being *Automatic Training Approach*. Key component of Knowledge Engineering Approach involves a person who is familiar with

information extraction system, who knows formalism for expressing rules for system, and who afterwards either upon himself or in consultation with some other experts, write rules for system. Hence, in Knowledge Engineering Approach, skills of knowledge engineer plays key role for contributing efficiency of overall system. Moreover, IE system is an iterative system as it involves writing rules again and again and then implementing rules followed by examining final results. If the results are not up-to expectation, then whole process is repeated to gain desired efficiency. Therefore, Knowledge Engineering approach involves a lot of labor.

Automatic Training Approach is quite different from formal one as it does not require system expertise for customization. In case of automatic training approach, as name suggests, it doesn't require a key person with specified detailed skills. It only requires one with enough information about domain, task of collecting corpus of texts, and extracting information. After training suitable text corpus, it was passed through a training algorithm and results can be implemented in analyzing novel texts. Data can also be trained during text processing phase by interacting with participant's users. Upon interaction, users are allowed to mark the system hypothesis as correct or wrong. In accordance with these user's interaction, system will get itself modified [22].

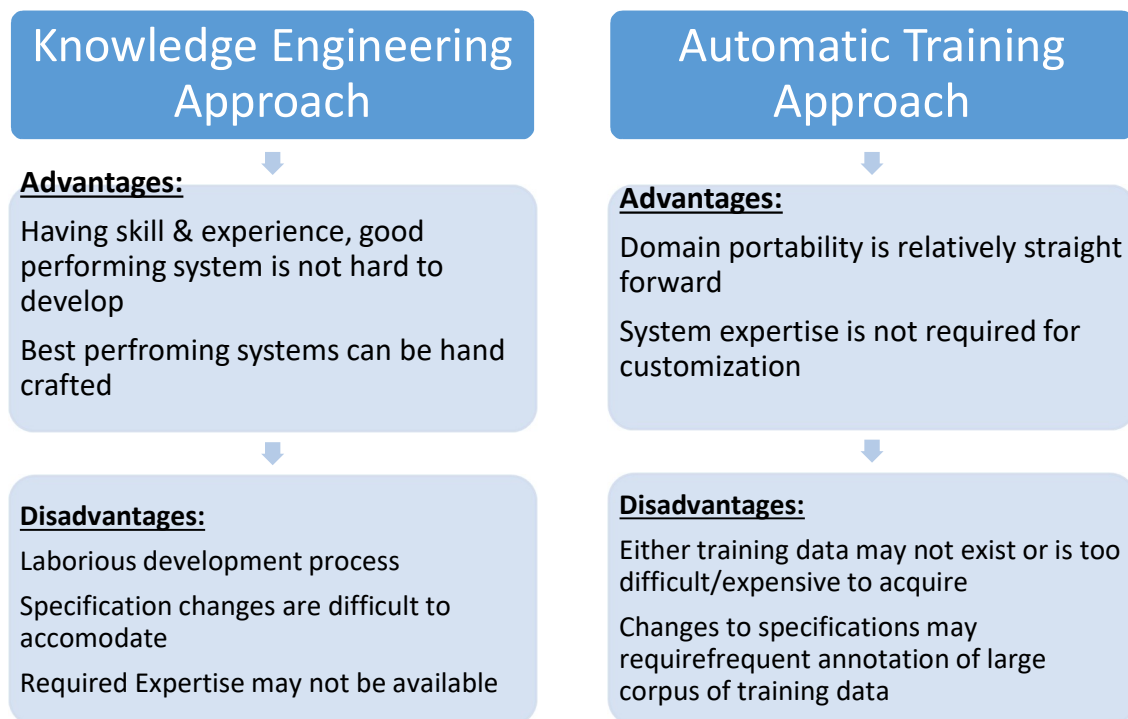


Figure 3.2: Information Extraction Approaches

3.3 Comparative Analysis of Event Extraction Techniques

From the past few decades, we witnessed a rapid increase in availability of textual information in digital form on Internet. Social media has now emerged as a universal platform that people are using more and more for communication purpose. Platforms such as Twitter or Facebook have now become a major source of information, especially, in situations such as social upheavals. A large amount of such information (e.g., e-papers, legal act reports, medical records etc) are available in raw form and therefore hard to search required information from junk of unstructured data. Event extraction is one common application of text mining, which is utilized in extracting specific knowledge regarding certain incidents referred in textual documents [19].

Event extraction is a specialized form of information extraction. At first, Information extraction was purely focused on perceiving messages from news wires. But due to the progressive advancements in natural language text types such news articles and web pages, more advanced techniques were required for extracting data with greater accuracies and on real time basis. With the advancements in text mining and big data, event extraction has gained much popularity. Therefore, many researchers have developed various techniques for extracting events from different social media platforms. Event extraction domain has rooted its domain knowledge from several domains like Datamining, Artificial intelligence (AI), Linguistics. Concept of event extraction falls back to 1980s, initiated by US Defense Advanced Research Projects Agency (DARPA), whose core objective was to automate the process of event identification related to terrorism from news wires.

This process was continued by ACE (Automatic Content Extraction), whose main purpose was to develop technology that infers entities mentioned, relation among them and events in which they participate, directly from human language data. Along with text data, it further includes audio and image data and language including English, Arabic and Chinese. The program initiated with a pilot study in 1999 [21].

In 2004, another program was initiated named as Informatics for Integrating Biology and Beside (i2b2). The program was introduced with aim to develop NLP (natural language

processing) techniques for extracting medication related information from patient records as it will help in accelerating adaptation of clinical findings into unique diagnostics and prognostics. Afterwards, Text Analysis Conference (TAC), was initiated with various sessions introducing term known as Event Track. The purpose of event track was to propose set of tasks for extracting events from text along with their participants [23]. With the passage of time, the process of event extraction spread over various domains such as politics, business, finance, political and many other fields.

Event can be regarded as a happening generally refers to incidents of considerable importance. These events characterize the state changes and hence widely used for decision making process or predicting upcoming trends. Currently, in research as well as in practice, numerous event extraction techniques have been proposed and implemented. Researchers have focused on several perspectives of event extraction methods. The main ones are event extraction methods with minimal processing cost, and these methods are to be executed with minimal involvement of man power or to be operated easily by non-specialists. Most commonly and constantly encountered problem while information extraction is that most of the data produced by humanity is in unstructured format that is human-understandable format. Hence processing such unstructured data intelligently requires advanced methods. Identifying key terms relevant to our domain or application and then storing it with existing data structures on basis of their mutual relationship are core concept of information extraction. Therefore, extraction tasks are generally distributed in three steps: First one is Named Entity recognition followed by relationship extraction and finally extracting targeted events.

Crowdsourcing has provided us with immense benefits and because of level of convenience it has provided to general public, it has been widely deployed in each perspective of daily life. Changsheng Wan et al., presented a crowdsourcing based parking reservation system in order to utilize parking spots fully. Keeping in view the user privacy concerns for drivers, a novel security protocol “SCPR” was introduced. SCPR (Secure Crowdsourcing based parking reservation system) main purpose is to authenticate drivers involved in parking reservation system and also hiding their real identities by designing key-agreement protocol based on cryptographic algorithms, thereby providing authenticity and efficiency at same time [24].

In [25], Matthew et al., proposed an open source prototype software known as RATCHET, whose goal is to leverage crowdsourcing task for development of attack trees. The software was used to build attack trees in accordance with organization's infrastructure and SVG (Scalable Vector Graphics- an HTML technology used for drawing in a web page) is used for displaying attack trees visually. They implemented the same for an IT (Information Technology) community focusing on building attack trees regarding computer security threats and vulnerabilities.

Nowadays, crowdsourcing is also emerging as an innovative approach for collecting people's affective response towards space and using such data for decision making process. The paper [26] describes how database can be collected by crowdsourcing people's affective response and these databases can be utilized for studying impact of environmental characteristics. The author presented a mobile application named as "EmoMap" Project. The project aimed at gathering people's affective response to space via mobile application and after this data, it is incorporated into geo spatial services such as smart location based services. It can also mark several places as places of interest or unsafe places of city for tourists.

Security of world's critical infrastructures are one of the most debated topic these days. In wireless communication systems, all data is communicated in clear way and is easy to manipulate by malicious hackers. Cryptographic solutions are generally impossible to deploy in wireless systems. In paper [27], a comprehensive model is proposed to counter this threat. The approach was illustrated by capturing large parts of ATC (air traffic communication) with help of crowdsourced sensor network named as OpenSky. The system was able to detect attacks such as jamming, spoofing and malicious insider threats. By implementing several verification methods, it can detect data maliciously injected into the wireless channels used for communication by ATCs.

In past few years, we have seen various successful research work in field of summarizing documents. Efforts have been made to put single document and multi-document summarizations in ease. The author in [28] presented a technique of preparing user-directed summary report by merging information extraction with summarization methods. Summary

reports generated via Information extraction supported summarization techniques are more accurate and domain specific.

The paper [29] proposed a cross-document event extraction and tracking technique aimed at presenting the events involving centroid entities with great accuracy. Important person entities which are often involved in events are referred to as centroid entities. The method links the events involving centroid entities in a proper timeline. The author was inspired by the fact that with passage of time, events keep on evolving, updated and corrected in several documents. The resulting information may overwrite the initial value, hence missing crucial facts. Cross-document Information extraction task will generate complete, salient and concise facts revolving around specified centroid entities.

Almost at each and every level of organization, knowledge-intensive works play an important role as workdays are usually portrayed as manipulation on digitized information. For a supportive system, it must take user's context into consideration. In this work [30], machine learning techniques are identified to be capable of timely task classification. User interaction with systems were recorded including low level events such as key strokes, mouse clicks and movement. Sheer amount of data is gathered on fine granular basis. Predefined set of static rules were used to map these recorded events with event blocks. On the basis of content present in event blocks, they are further assembled to form event block clusters. Like event blocks sharing significant amount of words are grouped within same cluster. These clusters are then verified and labelled by user and serve as ground truth for learning user task model.

Considering service to humanity, noise pollution is considered as a major drawback in urban environments as it put a huge effect on human behavior, health and productivity. Crowdsourcing has also put its benefits in this field. In paper [31], general public response was accumulated and being utilized for assessment of noise pollution. It turned citizens into live noise sensors just by turning on GPS-equipped mobile phones. Thus the response of users along with their geo-localized measurements were gathered producing a collective noise map. NoiseTube prototype was presented by author where an application is installed in mobile phones of general public turning them into noise sensors. This mobile app collects information

such as noise, GPS coordinates and send it to Noise Server. Noise server processes the data and produce noise map, thus presenting health risk levels at various geo graphic locations.

In recent years, crowdsourcing approaches have been implemented in field of bio medical investigations. Crowd involved is diverse in nature including domain experts and health information seekers. Information gathered in this field is usually classified in two domains: *Mining crowd data* such as through search logs and *Active crowdsourcing* via several technical forums like community challenges. In Mining crowd data, data is mostly owned by private organization such as TwitterAPI etc. These organizations make their data freely available for computation and knowledge discovery process. Ingram et al., [32] make use of an online tool “Google Trend” that gives statistics for search results on term usage in Google Search Engine. He searched for several terms regarding disordered sleep breathing in different seasons. Further, for comparative purpose, queries lying within time period of 7 years were collected from two representative countries. The potential of Google trend searches was empirically validated by using cosinor analysis that depicts that models performed much better when predictive variables from Google trend searches were included.

Information extraction is often termed as a process of extracting structured data from huge chunk of information buried in the form of unstructured database. Text documents contain numerous bulk of structured data. This data is of much importance as it could be utilized in various domains. For instance, government agencies may require archived data from newspapers for having a back track of news regarding some infectious disease outbreak. In [55], the author introduces QXtract system, an automatic query based technique for identifying documents that are useful for relation extraction from text documents. In start of QXtract system, the user is required to provide system with seed tuples. These seed tuples are then used for sampling process where sample of documents were retrieved from database. These documents contain targeted documents (useful for relation extraction process) and random documents (useless for relation extraction process). Further, these samples are then passed through information extraction process and as a result it produces a set of tuples as well as identifiers (either positive or negative depending upon their usefulness). The examples are then allowed to retrieve queries based on similarity with the ones labelled as positive. These queries

are used for extraction of so called promising documents which are further processed by information extraction system for retrieval of structured data from text documents.

Youtube has been declared as one of the biggest video forum as per official statistics. Due to availability of huge and continuously increasing video content on this platform, various search methods have been deployed for extraction of information from such tremendous haystack. The paper [56] showed that how to perform event detection on Youtube videos through concept of crowdsourcing. The paper's main contribution was to provide a generic framework designed for HTML5 videos. The whole process of event detection inculcates three types of events, i.e., visual events, occurrence events and the other one is interest-based events. Textual, visual and behavioral analysis was performed on Youtube videos while user was watching it. Once a video is played, video content is analyzed visually by detecting shots. The process operates on client side and user is offered to jump on desired shot upon clicking. Further, metadata corresponding to video is analyzed by using NLP techniques [57]. The list of named entities detected in video is displayed in front of user and user is offered to select targeted one. Finally, upon detection of visual events, process of interest based event detection starts by counting user's clicks in shots and attaching JavaScript listener to sound.

Digital services have been greatly familiarized among masses through social media as information propagation process is faster and quicker in social media platforms. Information dissemination through text documentation is one of the most effective methodology of communication. Event extraction from text corresponds to fetching events from text streams in social media. In [58], a two-step approach was presented for event extraction from digital text. First step involves identification of relevant messages containing the events by using binary classifier. Second step involves the procedure of event extraction by using CRF (Conditional Random Fields). The main focus of the proposed approach was on classifier building as to distinguish text streams/messages whether they contain an event or not.

Social media platforms such as Twitter and Facebook are becoming one of the most popular means of communication especially in emergency situations like catastrophic disaster, shootings etc. Usually information is shared on these forums earlier than news media [59-60].

In some cases, posts regarding numerous events are not verified and sometimes riddled with rumors. Therefore, extracting information from social media platforms need to be handled with great care as spreading false or inaccurate information will cause great damage to society [61]. Users either back or deny the rumor by sharing their views and opinions or sometimes evidences. In paper [62], the author describes an annotation method for enabling annotations of targeted tweets i.e., tweets involved in rumorous conversation. This allows to track and analyse trajectory of these rumors.

In past few years, online social media such as blogs & other forums have come up as a promising source of security intelligence information. Nowadays, researchers are focusing on extracting useful information from technical blogs & tweets from security professionals [63]. In paper [64], Liao et al. used text mining tools for extraction of threat intelligence information i-e. key attack identifiers such as malware signatures, botnet IPs etc. The approach involves determining the set of context terms from technical articles & defining the relation among these extracted terms.

In process of event extraction, events can be characterized by a set of descriptive, collocated words. Sayyadi et al. developed an even detection algorithm which inculcates creating a keyword graph followed by community detection methods for description of events [65]. In paper [66], Twitter data has been utilized for estimating level on interest in existing vulnerabilities & predicting their chance of exploitation. Alan et al. proposed a weakly supervised seed-based approach for detection of security related tweets from Twitter [67]. Khandpur et al. presented an approach for extraction & encoding of cyber-attacks discussed in social media [68].

Table 1: Comparative Analysis of Proposed Model with Past work

	Method			Event	Source Verification	Goal	Data
	No Training	Keyword Expansion	Information Extraction	Event Characterization			
[69]	×	×	×	×	×	Cyber Attacks	WINE
[64]	✓	×	✓	×	×	IOC	Tech Blogs
[66]	×	×	✓	×	×	Vulnerability	Twitter
[67]	×	✓	✓	×	×	Cyber Attacks	Twitter
[68]	✓	✓	✓	✓	×	Cyber Attacks	Twitter
This Research	✓	✓	✓	✓	✓	Cyber Crimes	Blogs / Articles

3.4 Conclusion

This chapter focused on identifying existing techniques used for purpose of event extraction. Several researchers presented techniques for extracting desired information from text documents such as news, blogs, social media forums etc. Later on, comparative analysis is done for existing techniques with proposed model.

Design Methodology and Implementation

4.1 Introduction

This chapter presents design methodology of model proposed in this research. The model comprises of query-based approach for extraction of cyber threats in Pakistan. Further, a source verification process is designed for analyzing the authenticity of extracted cyber-related events.

4.2 Architecture of Proposed Model

In this research, two models have been designed and presented to prove the declared claim. Following are models that has been obtained for extracting information and then verifying their source:

- i. Query based search model for extracting “cyber threats in Pakistan” reported through online resources
- ii. Source verification model having quintuple elements

We have provided an easiest solution to generate a database of cybercrimes in Pakistan. The methodology that has been followed to present the security model is shown in Fig. 4.1.

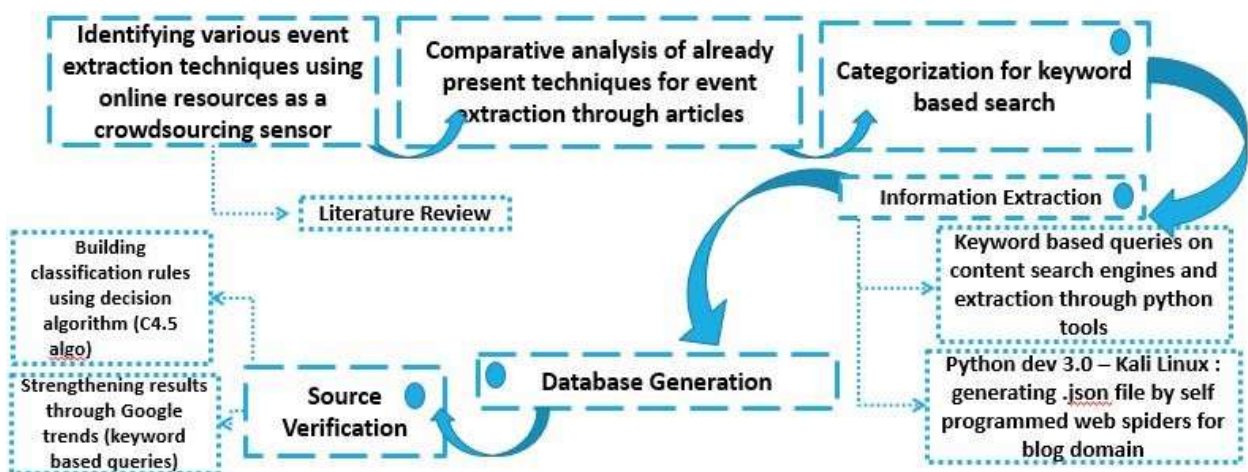


Figure 4.1: General Model for Extraction and Verifying the Targeted Data

The research phase was started with literature review in which various event extraction techniques were identified that used online resources as a crowd source sensor. This researched focused and targeted the online blogs for extracting information while assuming that online blogs had catered the required information of cybercrimes in Pakistan. Afterwards, a comparative analysis of already present techniques for event extraction through articles and blogs was done. This comparative analysis has already been described in previous chapters. This comparative analysis is then followed by the design of query based search model, in which the most important part is the categorization of keywords on which all the search is actually based. After categorization, information would be extracted and database would be generated. In the end, source verification would be done and the results would be analyzed using decision algorithm. However, the results would be strengthening using Google trends which will support our claims.

First of all, the first model is explained in detail which is based on the query based search for extracting the information through online resources.

4.3 Query Based Search Model

After going through different information extraction techniques, unique solution was proposed. The solution comprising of proposed model consists of following steps:

- i. Categorization for keyword based search
- ii. Information Extraction
- iii. Database Generation

These steps will be discussed in detail in this chapter. Each step includes separate work that needed separate attention. In query based search model, different categories have been defined and each category need different keywords on the basis of which blogs have been searched. Afterwards, using these categories, blogs have been searched and scrappers have been made to crawl the blogs to extract the required information. Different scrappers have been made and required information can be extracted. Thus, after extracting information, a new database has been generated in which required information is fit in for future use.

4.3.1 Categorization for Keyword Based Search

Some major categories have been identified on the basis of which keywords have been defined. These keywords have been used to search different blogs with targeted data. Here blogs have acted like crowdsourcing sensors and these sensors have been triggered using some initial surges which in our terms are the keywords. There are five major categories that are website attacks, account hacking, scamming, cyber harassment and data breach. These categories are then further detailed analyzed by creating appropriate seed queries which further created extended queries. Let block of seed queries of a particular category be "D". Each seed query of the particular block of independent category would be represented as D_i .

Thus

$$D = \text{Sum of all } (D_i)$$

$$D = D_1 + D_2 + D_3 + \dots + D_i$$

Shown below is the table consisting of major categories, seed queries and extended seed queries.

Major Categories	Keywords (Seed Query)	Keywords (Extended Query)
Website Attacks	Dos/Ddos attack, ransomware, sniffing, malware threats, website defacement etc.	india/pak website attack, special events (23 rd march, palwama, etc)attacks
Account Hacking	Account hijacking, account hacking, mitm attacks, security attacks,	Pm/president/social personalities etc account hacking, Security breach
Scamming	fraud, Phishing, social engineering	Atm skimming, financial fraud etc
Cyber Harassment	stalking, cyber bullying, pornography	Online harass, internet/cyber abuse, Porn videos/pics, child porn, sextortion etc.
Data Breach	Data compromise, data leak, sniffing,	Data leak in private/public sector, data compromise in (various places)

Figure 4.2: Major Categorization for Keywords

Categorization is solely depending on the events that are related to the particular categories. Fig. 4.2 presents major keywords that have been used for searching the required data. However, during the database generation a lot more inputs have been made to obtain the required data. It

was also observed that "cybercrime" and "cyber crime" are different keywords and produce different results when query is made on internet.

4.3.2 Information Extraction

After keyword categorization, techniques are adopted to extract the required information. For this purpose, scrappers were built to scrap the information from the online blogs. Two ways have been adopted to extract the information. The first one has some drawbacks and is slow that's why there was need to move forward to develop another way. Following are the ways that have been adopted to scrap the information online.

- i. Customized scrappers for each blog domain
- ii. Keyword based queries on content search engines then scrapping data through scrapper

First of all, the technique listed at number (i) is described in detail, here we have made customized scrappers for each blog domain.

4.3.2.1 Customized Scrappers for Each Blog Domain

In this technique, python language has been used. Programming is done using KALI LINUX Shell. The reason for developing separate scraper for each blog domain is that each blog domain has different tags and scrapping is done by fetching data using html tags. As html tags are different for each blog domain thus scraper that is used to fetch data is also different. Scraper will find the required html tags and then will fetch the values associated to that tags. If the tags are different or not listed than it would be not possible for the scraper to fetch the value as the match fail condition will occur. Fig. 4.3 describes the technique in the form of workflow that have been adopted.

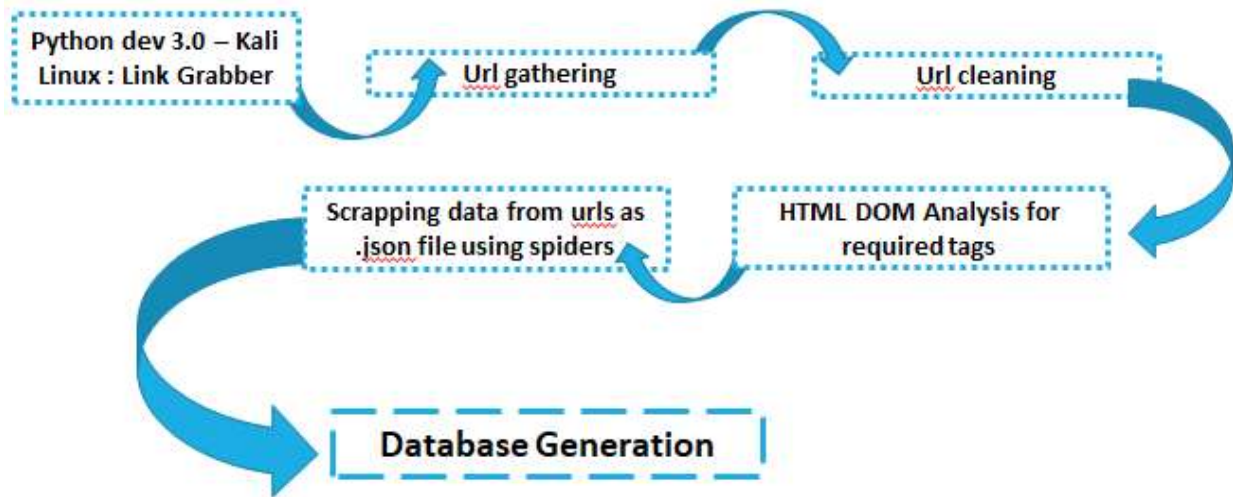


Figure 4.3: Technique Adopted for Creating Customized Scappers

Python dev. 3.0 in Kali linux have been used. There was a laptop with dual operating system, one being Windows8 and another was Kali linux. Kali linux have been used initially as it is popular for its attack nature. Additionally, there were tools available in kali linux for carrying out a kind of hack. As the focus was to fetch the data by parsing the html tags thus, kali linux have been opted for the subjected task, initially. At first, link grabber was used to extract the links available on the specific site. It's the already available tool of python. It grabs all kind of links including URLs image links, advertisements links etc.

S_n = Specific URL from where data needs to be fetched.

Link Grabber = (Grab all link exist on S_n)

G_n = Result containing all kinds of links including S_n

As now link grabber has extracted all kinds of links, large size file has been obtained containing all kinds of links. Here, only those links needs to be gathered in which link is directing towards another site. In other words, discarding the links of images, advertisements etc.

Now, there is a file containing number of links, however, targeted links have not been extracted till now. To extract the targeted links, URL cleaning process has been performed. Fig. 4.4 shows the technique of URL cleaning.

```

Query (keyword);
View (https://....);
{url, text_url, title_url} = Linkgrabber (site_url); //python coding
// url cleaning
If title_url ÷ { set of keywords} || text_url ÷ { set of keywords}
then
    save(url)
else
    return;

```

Figure 4.4: Psuedo Code for Url Cleaning

Fig. 4.4 showed that after query of particular keyword, a site will be opened; this site's URL will be given to link grabber for purpose of extracting further links. The link grabber, as a result, produces links along with text_url, title_url for each link. Title_url in each case would represent the title of the page to which that link is referring to. Thus, enough knowledge of the page has been gained to decide whether the page is of interest or not. An algorithm has been designed where the title URLs are matched with already specified keywords. If it matches, URL will be retained, otherwise, it will be discarded. After this step, a compressed size file having only targeted URLs have been obtained.

Next step inculcates the HTML DOM analysis for each domain. Html DOM analysis was done by clicking the Inspect element section of site and then searching for the required tags. The tags were present in divisions while some of them were referred as an anchor tags. Anchor tags are mostly used for authors. After collecting a large number of domains, then for each domain, it was required to find tags. The process seems very tedious. Thus, scrappers were built for two blog domains. The scrapers were for "Pak observer" and "the nation" blogs. Html DOM analysis includes the inspection of site for required tags as shown in Fig. 4.5.

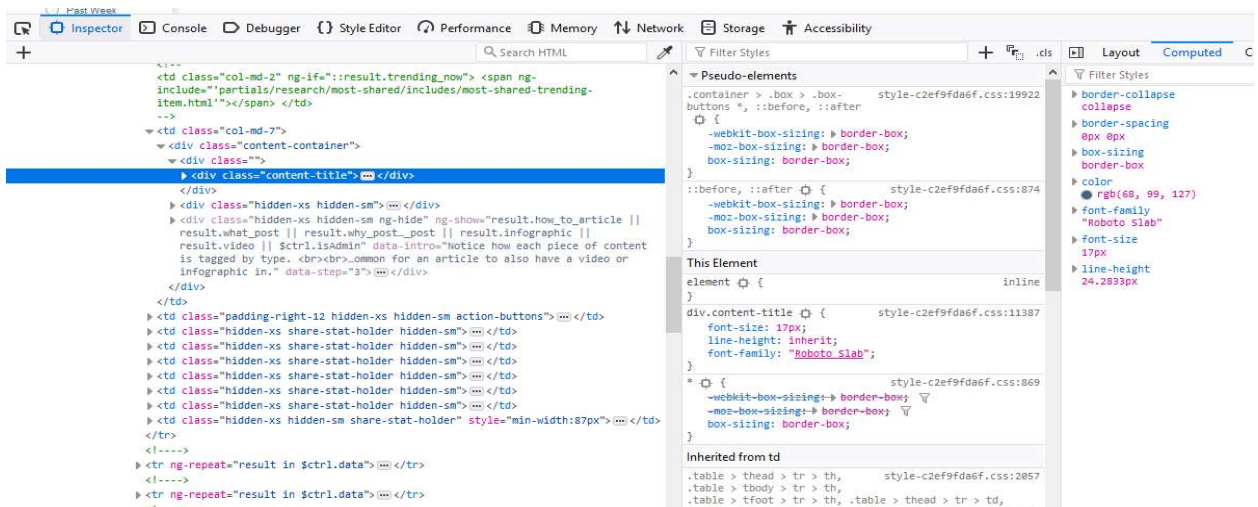


Figure 4.5: Html Dom Analysis

In the aforementioned technique, scrapy tool was used that comes along with python. It scraps the data using html tags. We have provided some tags and on the basis of the tags it produced the results. However, there exists some drawbacks of this technique.

- i. Each blog domain has different tags, so you need a different scrapper for each blog domain
- ii. This technique is too slow and tedious
- iii. Only get data of that specific URL (blog) after single execution of code, thus for 100 entries; code needs to be executed for 100 times using 100 different URLs
- iv. It required a lot of manual work

Thus, another technique has been developed in which a different approach has been used.

4.3.2.2 Keyword based queries on content search engines then scrapping data through scrapper

In this technique, python language has been used. Programming is done using Windows 8 command prompt. A new technique has been designed to make work speedy. Some content search engines were found out as these search engines automatically fetch the required data corresponding to the keyword that has been queried. But the result obtained was not free of cost, it has to be bought. Here, programming skills were utilized and data has been fetched using programming codes. The workflow of the technique adopted has been shown in Fig. 4.6.

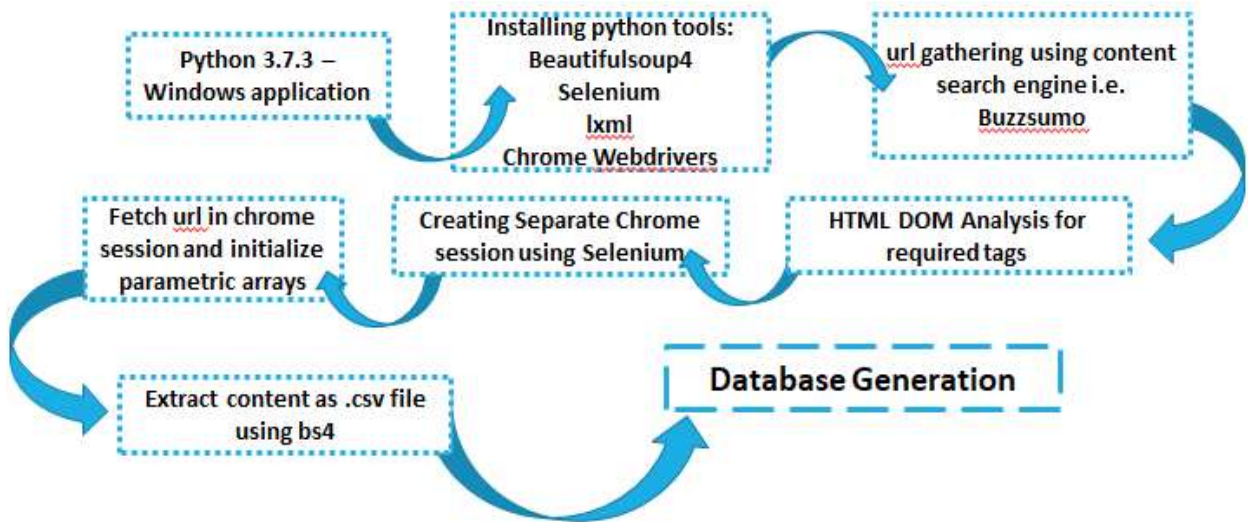


Figure 4.6: Technique for Keyword Based Queries on Content Search Engines

Python 3.7.3 has been used which is a windows application. Requisite python tools have been downloaded by using command prompt. Four python tools were downloaded for programming scraper.

- i. **Beautifulsoup4:** Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with a parser, hence, providing idiomatic ways for tasks such as navigating, searching, and modifying the parse tree. It speeds up the process saving programmer's hours or days of work.
- ii. **Selenium:** Selenium Python bindings provide a simple API for purpose of writing functional/acceptance tests using Selenium Web Driver. By using Selenium Python API, one can access all functionalities of Selenium Web Driver in an intuitive way. Selenium Python bindings provide a convenient API to access Selenium Web Drivers like Firefox, Ie, Chrome, Remote etc.
- iii. **Lxml:** The lxml XML toolkit is a Pythonic binding for the C libraries libxml2 and libxslt. It is unique in a way that it combines the speed and XML feature completeness of these libraries with the simplicity of a native Python API, mostly compatible but superior to the well-known ElementTree API.
- iv. **Chrome Web drivers:** Chrome web drivers controls the Chrome Driver and allow to drive the browser. In short, it creates a new instance of the chrome driver by starting the service and then creating new instance of chrome driver. Execute Chrome Devtools

Protocol command and get returned result. Fig.4.7 illustrates the installation of tools through command prompt.

```
Administrator: Command Prompt
Used up to 3 times (corresponding to WARNING,
ERROR, and CRITICAL logging levels).
--log <path> Path to a verbose appending log.
--proxy <proxy> Specify a proxy in the form
[user:passwd@]proxy.server:port.
--retries <retries> Maximum number of retries each connection should
attempt (default 5 times).
--timeout <sec> Set the socket timeout (default 15 seconds).
--exists-action <action> Default action when a path already exists:
(s)witch, (i)gnore, (w)ipe, (b)ackup, (a)rtort.
--trusted-host <hostname> Mark this host as trusted, even though it does
not have valid or any HTTPS.
--cert <path> Path to alternate CA bundle.
--client-cert <path> Path to SSL client certificate, a single file
containing the private key and the certificate
in PEM format.
--cache-dir <dir> Store the cache data in <dir>.
--no-cache-dir Disable the cache.
--disable-pip-version-check Don't periodically check PyPI to determine
whether a new version of pip is available for
download. Implied with --no-index.
--no-color Suppress colored output

C:\WINDOWS\system32>pip install selenium
Collecting selenium
  Downloading https://files.pythonhosted.org/packages/80/d6/4294f8b4bce4de8abf13e17198289f9d8613b8a44e5d66a7f5ca98459853/selenium-3.141.0-py2.py3-none-any.whl (984kB)
  100% |#####| 911K 573K/s
Collecting urllib3 (from selenium)
  Downloading https://files.pythonhosted.org/packages/62/0b/ee1d7de624db8ba789801226aebfab96a2c71cd5fa7629d6ad3f61b79e/urllib3-1.24.1-py2.py3-none-any.whl (118K)
  100% |#####| 122K 429K/s
Installing collected packages: urllib3, selenium
Successfully installed selenium-3.141.0 urllib3-1.24.1

C:\WINDOWS\system32>pip install BeautifulSoup
Collecting BeautifulSoup
  Downloading https://files.pythonhosted.org/packages/1f/ee/295988deca1a5a7accd783d8dfe14524867e71abb0b8c0eeceee49c759d/BeautifulSoup-4.7.1-py2.py3-none-any.whl (94kB)
  100% |#####| 102K 516K/s
Complete output from command python setup.py egg_info:
Traceback (most recent call last):
  File "<string>", line 3, in <module>
  File "C:\Users\ALISID-3\AppData\Local\Temp\pip-install-zb0ix7e1\BeautifulSoup\setup.py", line 22
    print "Unit tests have failed!"
    ^
SyntaxError: Missing parentheses in call to 'print'. Did you mean print("Unit tests have failed!")?

C:\WINDOWS\system32>pip install BeautifulSoup4
Collecting BeautifulSoup4
  Downloading https://files.pythonhosted.org/packages/1d/5d/2b0694a59df9ec32f8b4883f5d23b130bc2376021411fan70ee12351e/BeautifulSoup4-4.7.1-py3-none-any.whl (94kB)
  100% |#####| 102K 516K/s
Collecting soupsieve>=1.2 (from BeautifulSoup4)
  Downloading https://files.pythonhosted.org/packages/c9/f8/e54bd771ed4fab6b3fa1c178e1373c73d84f6f643299d4fd6da5a171c/soupsieve-1.9-py2.py3-none-any.whl
Installing collected packages: soupsieve, BeautifulSoup4
Successfully installed BeautifulSoup4-4.7.1 soupsieve-1.9

C:\WINDOWS\system32>
```

Figure 4.7: Installation of Python Tools

After installing python tools, keyword for defined categories have been queried for searching the specific blogs. In order to do so, a content search engine has been utilized. The search engine used was known as “Buzz Sumo”. It has many features and it provide the free trail for 7 days with limited features. Initially, an account has been created on content search engine and account has been used for fetching targeted data. While using account, there came a lot of limitations that it didn’t allow to save data if the account was not purchased from official website. It also doesn’t show additional pages to look. The only solution was to build a scrapper and fetch the data from it. Fig. 4.8 shows the keyword search and respective results obtained.

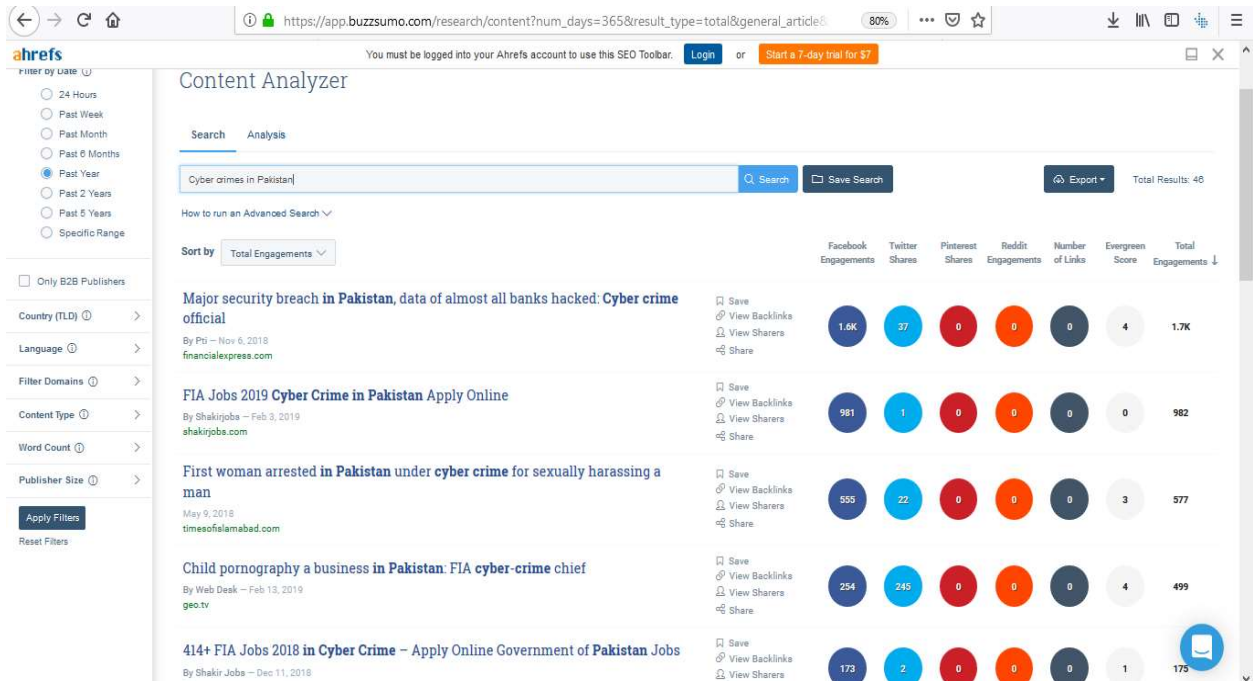


Figure 4.8: Keyword Search and Results on Buzz Sumo

For example, a keyword was searched named as “Cyber crimes in Pakistan” in search dialog box and after search, 20 entries were displayed in case if the account has been made otherwise only 5 entries were shown. This page includes the title of blog, author name, publishing date and stats. The purpose was to fetch this information. Scrapy tool was used to fetch the information; but somehow it was not possible. On further research, it was found that it is dynamic site and content is hidden in this case. So there arose a need to establish a new session that is automated and controlled by the automated software which will bypass the security policies of the Google. Thus, selenium was used for this purpose and chrome drivers were utilized to establish such kind of automated session. Fig. 4.9 shows the automated session that has been created by using selenium tools along with chrome web drivers.

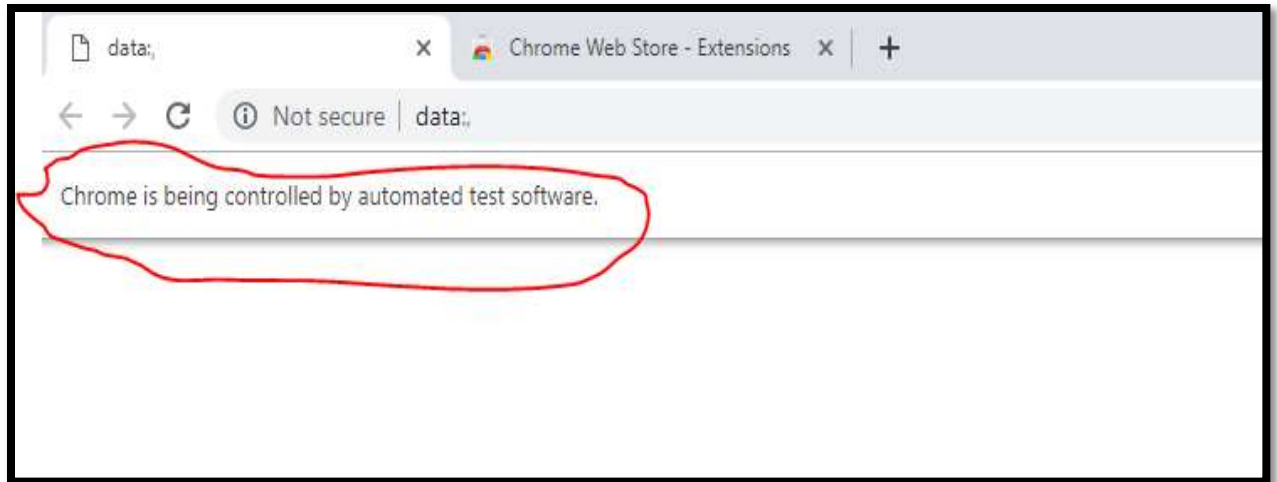


Figure 4.9: Controlled Session of Chrome

After session creation, BuzzSumo website was loaded for which it required few seconds as depicted in Fig. 4.10.

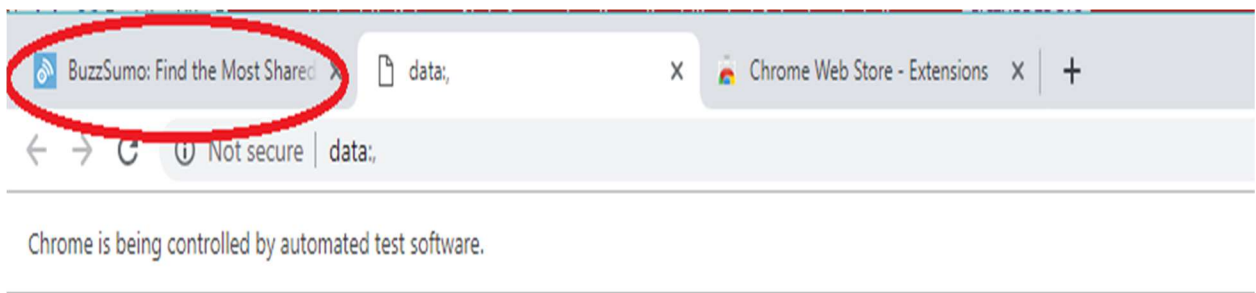


Figure 4.10: Loading of BuzzSumo Website

Next step is to perform HTML DOM analysis for each domain. HTML DOM analysis was done by clicking the Inspect element section of site and then searching for the required tags. The tags were present in divisions while some are referred as anchor tags. Anchor tags are mostly used for authors. After collecting a large number of domains, tags were required to be found out. Here, the focus was to find the tags of title of blog, author name, publishing date, URL link and stats. Buzzsumo presents stats for Facebook, Twitter, Reddit, Pinterest number of links evergreen scores and total scores. All the stat values were hidden thus indicating the dynamic nature of site as stats vary with passage of time, thus Scrapy tool is not recommended for such purpose. Therefore, selenium was the best choice to carry out the task. Social animal is another content analyzer that can be used in place of Buzzsumo. However, in this research, data is only collected from Buzzsumo.

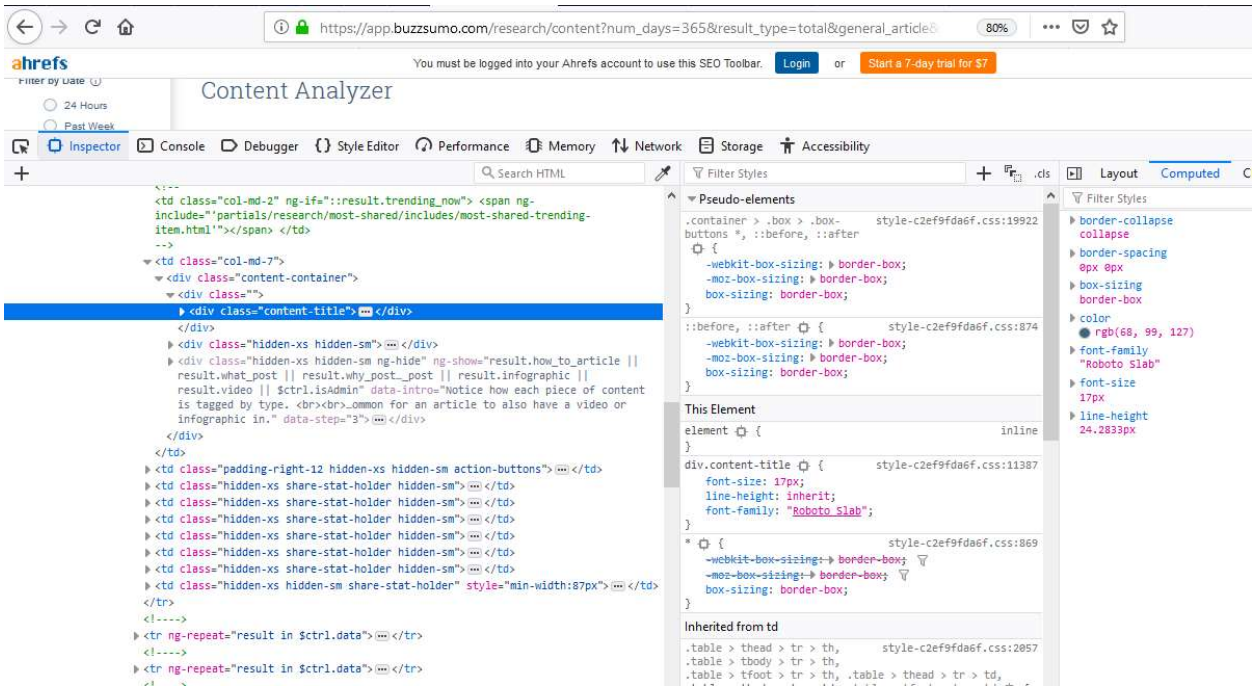


Figure 4.11: Html Dom Analysis for BuzzSumo

Next step inculcates fetching URL in chrome session and initializing parametric arrays. There are 11 parametric arrays which were filled on after the execution of code. A timer of 1 minute was set in code so that site can be fully loaded and account credential can be made so that instead of fetching the wrong information site must wait and fetch the required information. In the end, beautiful soup is used to extract the required information from the loaded site. Fig. 4.12 shows the csv file that has been created with following entries.

	A	B	C	D	E	F	G	H
1	Title	Author Name	Domain	Date	Facebook Engagements	Twitter Engagements	Pinterest Engagements	Reddit Engagements
2	With Over \$6 Million Missing	Biya Haq	mangobaaz.com	29-Oct-18	703	3	0	0
3	Major cyber attack hits private Bank in F	Kashif Rafiq	siasat.pk	29-Oct-18	469	14	0	0
4	Worst ever Cyber Attack on Pakistan - Ir	Editor	theoctobersky.com	17-Feb-19	439	0	0	0
5	BankIslami lost \$6m within 22 minutes i	Staff Report	pakistantoday.com.pk	22 Dec 18	155	124	n	n

Figure 4.12: Scrapped Data in Excel Format

4.3.3 Database Generation

After carrying out different keyword based searches, database was created. This database contains a lot of entries. Details of database will be discussed in next chapter. The header of the database contains following entries.

- i. Title
- ii. Author
- iii. Date
- iv. Blog Domain
- v. Blog - url
- vi. Facebook Engagements
- vii. Twitter Shares
- viii. Pinterest Engagements
- ix. Reddit Engagements
- x. Total
- xi. SEO score -Domain
- xii. Moz Rank
- xiii. Https test
- xiv. Server signature test
- xv. Global Rank
- xvi. Country Rank
- xvii. FINAL SCORES
- xviii. Authenticity

The header of database is shown in Fig. 4.13

Database comprising of 11 columns was generated and 20 entries were fetched at a time in the case of adding the account credential details on opening the new session of the chrome.

Algorithm 1: Code to Fetch 20 Entries

```
1. for i in range(20):
2.   for j in range(11):
3.     if j == 1:
4.       row.append(column[j][i])
5.     else:
6.       row.append(column[j][i].get_text())
7.   writer.writerow(row)
8.   row=[]
```

In case of not fetching data by using the account credentials, then the program have been reustomed for fetching out 5 enteries as displayed on the screen.

Algorithm 2: Code to Fetch 5 Entries

```
1. for i in range:(5)
2.   for j in range:(11)
3.     if j == 1:
4.       row.append(column[j][i])
5.     else:
6.       row.append(column[j][i].get_text())
7.   writer.writerow(row)
8.   row[]=
```

The name of author was not simply taken as name of author as the author tag has many whitespaces and it starts with the phrase "By".

Algorithm 3: Extracting Information from Tags

```
1. author_div = soup.find_all('div',attrs={'class':'author-name'})
2. for div in author_div:
```

3. `author_a.append(div.find('a'))`
4. `for author in author_a:`
5. `author_names.append(author.get_text().strip()[3:])`

4.4 Source Verification

This research presents source verification model having quintuple elements. These elements play a major role in source verification. This model is never presented before as far as this research is concerned. This model contains five elements that are listed below.

- i. Social Stats
 - a. Facebook Engagements
 - b. Twitter Shares
 - c. Pinterest Shares
 - d. Reddit Engagements
- ii. Domain Moz Rank
- iii. Domain SEO Score
- iv. Domain Rank
 - a. Global Rank
 - b. Country Rank
- v. Site Authentication
 - a. Https Test
 - b. Server Signature Test

These elements were acquired through programmed code and through Google SEO small tools. The details have been shown in the Fig 4.14.



Figure 4.14: Technique for Source Verification

SEO small tools contains many tools out of which we have only used two tools that are as follows:

- i. **Website SEO Score Checker:** Website SEO Score checker displays the SEO score of website as name depicts clearly. Google perceives security feature as of great importance. SEO score checker tests a large number of parameters (each parameter is kind of Big data itself such as Meta description, Keywords density test, Robots.txt test, Sitemap & broken links test etc.) & shows the detailed information of each parameter along with strengths & weaknesses regarding of targeted website.

- ii. **Alexa Rank Comparison:** Alexa is a global ranking system that utilizes web traffic data compiling a list of the most famous websites, the Alexa Rank. Lower Alexa Rank indicates the popular sites (for example, a site with the rank of 1 has the most visitors on the internet). The span is the number of clients that visit a particular website in a day. Site hit, as the name suggests, is the circumstances the clients view a specific page. Alexa ranking makes sure that if a particular client visits a similar page a few times around the same time, at that point, it will consider each one of those visits one.

The elements that have been taken by using these tools are as follows

- i. Website SEO Score Checker
 - a. Domain SEO Score

- b. Site Authentication
- ii. Alexa Rank Comparison
 - a. Domain Moz Rank
 - b. Domain Rank

Each element has further sub elements. Details of each element have been described below

4.4.1 Social Stats

Social stats have been extracted through the content analyzer. The coded program extracts the statistical information. The social stats that we have incorporated in our database for source verification are as follows

- a. Facebook Engagements
- b. Twitter Shares
- c. Pinterest Shares
- d. Reddit Engagements

These social stats vary for each blog. These stats shows the interest of people on that particular blog and how much popular among the folks. The stats show that blog's topic had been a hot topic.

4.4.2 Domain Moz Rank

MozRank is one of the most popular & dependable for purpose of measuring authority of a web page/site. Many webmasters and SEO experts are using MozRank, nowadays, as a point of reference for optimizing search engines. This tool is created by an organization named as Moz, a company that provides tools for search engine optimization. It is a measure of domain linking authority or popularity of a given domain, hence, reflecting the importance of a domain on the internet comparatively.

MozRank of web pages is based on the similar pages on the web that are linked to them as well as the MozRank of those pages with links. This would also mean that if the MozRank of the linking pages is high, there is a greater chance that the MozRank of the receiving page of those links will also be high. Fig. 4.15 shows the Moz Rank checker.

MOZRANK CHECKER



Figure 4.15: MOZ Rank Checker

4.4.3 Domain SEO Score

SEO stands for “Search Engine Optimization.” Search Engine Optimization (SEO) is the process of influencing the visibility of a website in a search engine’s “natural” or “organic” search results. Search engine optimization aims to maximize the number of visitors to a website thus, ensuring that the web site appears high on the list of search results from search engines. It is the measure how well the user-facing and technical aspects of the domain contributes to SEO. If the organic traffic towards a page is high, it implies that rank of that site will also be high & as a result SEO score will be high too. Fig. 4.16 shows the SEO checker.

WEBSITE SEO SCORE CHECKER

If you want to check the SEO score for your site or your competitor's. Enter the domain name or URL into the given field and click "Check SEO." You will get the results in no time!

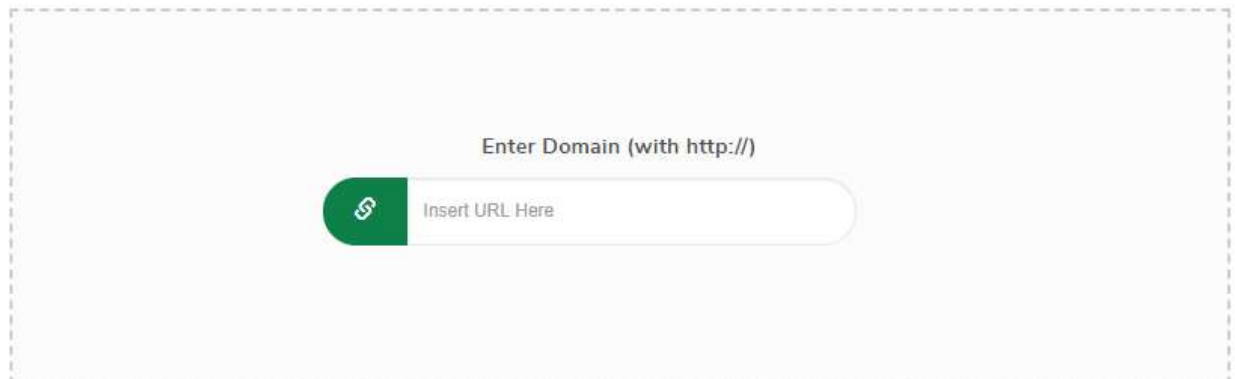


Figure 4.16: Website SEO Checker

4.4.4 Domain Rank

Domain Authority (DA) is a search engine ranking score developed by Moz focused on predicting how well a website will rank on search engine result pages (SERPs). A Domain Authority score lies within range of 01 to 100, with higher scores referring to high chances of ranking. Domain Authority is based on evaluation results of multiple factors, including linking root domains, number of total links, merging into a single DA score. This score can then be used while comparing websites or tracking the "ranking strength" of a website with passage of time. It shows that how well the website will be ranked on different search engines. It has following two elements

- i. **Global Rank:** It tells how well the website altogether rank on different search engines globally
- ii. **Country Rank:** It tells how well the website altogether rank on different search engines with respect to country

4.4.5 Site Authentication

It includes https test and server signature test

- i. **Https test:** It tells whether the communication of this domain with everyone else is encrypted i.e. secure or not
- ii. **Server signature test:** A server signature represents the public identity of a web server. It contains sensitive information that that can be used to exploit any known vulnerability. Thus, as per standard security practices, it should be turned off so to avoid disclosure of information such as software version running etc.

4.5 Conclusion

This chapter discussed the design & research methodology of proposed model. It includes few major steps; the first step being query based search model followed by process of information extraction. After extracting targeted information, database is generated with initialized parametric arrays. The information extracted from online resources was also authenticated by a novel procedure of source verification.

Results and Analysis of Proposed Model

5.1 Introduction

This chapter presents the implementation results of query based search model proposed in this research. Further, analysis is done on the results obtained for getting deep insight of evaluation of model.

5.2 Overview of Proposed Model

Database has been generated by fetching the data from the online sources. The code has been programmed using python language. Data has been fetched from the online blog/articles using html tags. After fetching the data, all the fetched data is saved in excel format. The obtained data is not completely in usable format thus it's made so by taking some necessary measurements. Further data is not only obtained by the websites only. In order to perform source verification which is very important element of our thesis, data is obtained by using SEO small tools build by Google.

A source verification model having quintuple elements was presented. These elements play a major role in process of source verification. This model is never presented before as far as our research is concerned. The analysis is done by using different formulas. This model contains five elements that are already listed. This chapter discusses each element in detail and how they have been modelled to get the desired results.

5.3 Defined Authentication levels

In the model of source verification, five authentication levels were defined as listed below (also shown in Fig. 5.1):

- i. High
- ii. Medium High
- iii. Medium
- iv. Medium Low
- v. Low



Figure 5.1: Levels of Source Verification

These levels have been defined according to the scale. Table 2 describes the criterion of the scale set for authenticity evaluation.

Table 2: Authenticity Criterion

<i>Authenticity Criterion</i>	
<i><20</i>	<i>LOW</i>
<i>>20<40</i>	<i>MEDIUM LOW</i>
<i>>40<60</i>	<i>MEDIUM</i>
<i>>60<80</i>	<i>MEDIUM HIGH</i>
<i>>=80</i>	<i>HIGH</i>

The criterion has been defined by fixing the parameters. The values that lie below 20 are regarded as "LOW". The values that lie between the range of 20 and 40 are regarded as "Medium Low". The values that lie between the range of 40 and 60 are regarded as "Medium High". The values that are above 80 are regarded as "High".

5.4 Feature Scaling

Feature scaling (also known as data normalization) is methodology used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms. Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important. Feature scaling is often used to

standardize the range of independent variables or certain features of data. In task of data processing, it is also termed as data normalization and is generally executed during the data pre-processing step.

Feature scaling technique of Normalization of data is described in formula as

$$X' = a + \frac{(X - X_{\min})(b - a)}{X_{\max} - X_{\min}}$$

Where

$X \in [X_{\min}, X_{\max}]$ is measurement to be scaled

X_{\min} = minimum of the range to be scaled

X_{\max} = maximum of the range to be scaled

a = minimum of the range to be scaled

b = maximum of the range to be scaled

Data processing is, generally, the collection and manipulation of items of data to produce meaningful information. Therefore, it is considered a subset of information processing, the change (processing) of information in any manner detectable by an observer. The term Data Processing (DP) also corresponds to a department within an organization responsible for the operation of data processing applications. Feature Scaling or Standardization is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm, hence increasing computational speed of system.

Scale each of quintuple elements ranging from 1 to 20

We have defined 5 marks for each social stat. Thus distribution of each social stat out of 20 is equivalent.

- i. **Sum of social stats = (Facebook Engagements + Twitter Shares + Pinterest Shares + Reddit Engagements)**
 - a. Facebook of Engagements – mark on the scale 1 to 5
 - b. Twitter Shares – mark on the scale 1 to 5
 - c. Pinterest Shares – mark on the scale 1 to 5
 - d. Reddit Engagements – mark on the scale 1 to 5

- ii. **SEO Score Domain – mark on the scale 1 to 20**
- iii. **Moz Rank – mark on the scale 1 to 20**
- iv. **Authenticated Site**
 - a. Https test - if https test is passed then mark it 10 score otherwise 0
 - b. Server signature - if server signature is off then mark it 10 score otherwise 0
- v. **Domain Rank**
 - a. Global Rank – mark on the scale 1 to 10
 - b. Country Rank - mark on the scale 1 to 10

There are some parameters that can't be scale by the feature scaling way because the data has a large set of variant values. These values had caused a large diversion when calculating through formulas and giving the range in less than 0. Even the maximum value did not get higher than 0 in some cases. Thus, a manual way is adopted to do so. Here, the values having large variations and having a huge difference in minimum and maximum values are given range in a manual way. Social stats and domain ranking are the features that have such variations and needed a way other than formulized feature scaling. Table 3 shows the scaling criterion for social stats and domain ranking.

Table 3: Social Stat and Domain Ranking

Social Stat Scaling		Global Rank Scaling		Country Rank Scaling	
1	<=50	10	<=100	10	<=50
2	<=500	9.5	<=500	9.5	<=250
3	<=1000	9	<=1000	9	<=500
4	<=1500	8.5	<=1500	8.5	<=1000
5	<=2500	8	<=3000	8	<=1500
6	<=3500	7.5	<=5000	7.5	<=2500
7	<=5500	7	<=8000	7	<=3500
8	<=7500	6.5	<=15000	6.5	<=4500
9	<=9500	6	<=25000	6	<=5500
10	<=10500	5.5	<=50000	5.5	<=6500
11	<=12500	5	<=80000	5	<=7500
12	<=13500	4.5	<=100000	4.5	<=8500

13	<=14500	4	<=200000	4	<=10000
14	<=15500	3.5	<=500000	3.5	<=25000
15	<=16500	3	<=800000	3	<=50000
16	<=17500	2.5	<=2000000	2.5	<=100000
17	<=18500	2	<=5000000	2	<=150000
18	<=19500	1.5	<=8000000	1.5	<=250000
19	<=20000	1	<=10000000	1	<=350000
20	>20000	0.5	>10000000	0.5	>350000

5.5 Data Analysis

We have analysed the data using different techniques. The first way that is adopted is to find out how many of our scrapped articles found to be verified and then verify the domains to which they are linked. We have found out that none of the blogs scrapped reached to the high level of source verification as per our defined criterion. This seems true as we haven't added the factor of verification from the legal site or governmental site. We haven't found out the legal proceeding or affiliation of the article with respect to daily broadcasting news of highly reputable news channel. This can be done as future project but as per our finding and laid down criterion, 230 blogs are ranked as "Medium". This is quite reasonable as our parameters lies on the fact of domain authentication or validity of domain. Relying totally on domain is not an ultimate solution; legal proceeding and their results must be considered to declare the validity of an article.

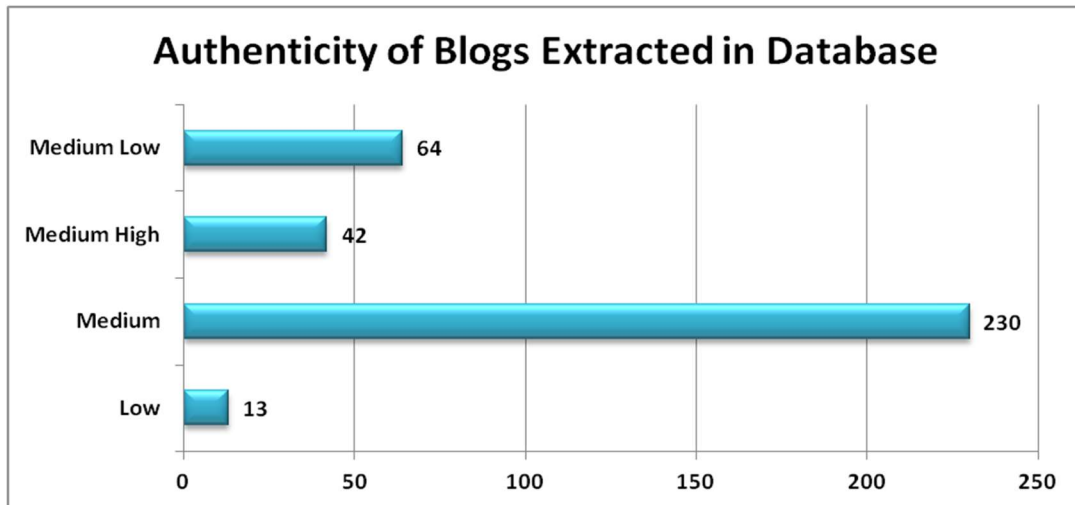


Figure 5.2: Authenticity of Blogs Extracted in Database

The graph displayed in Fig.5.2 further depicts that 64 articles are ranked as medium low and 42 articles have been ranked as medium high, however only 13 articles are those where source is found to be very weak thus ranked as low.

In the Fig.5.2, analysis have been done on the basis of individual blogs. The same work has been done by considering the domain, instead of articles. In order to get which domain has produced largest number of low verification level articles. This might help in future indication that these sites are not reliable for trusting the information. Fig. 5.3 displays the domain graphs as per the verification level.

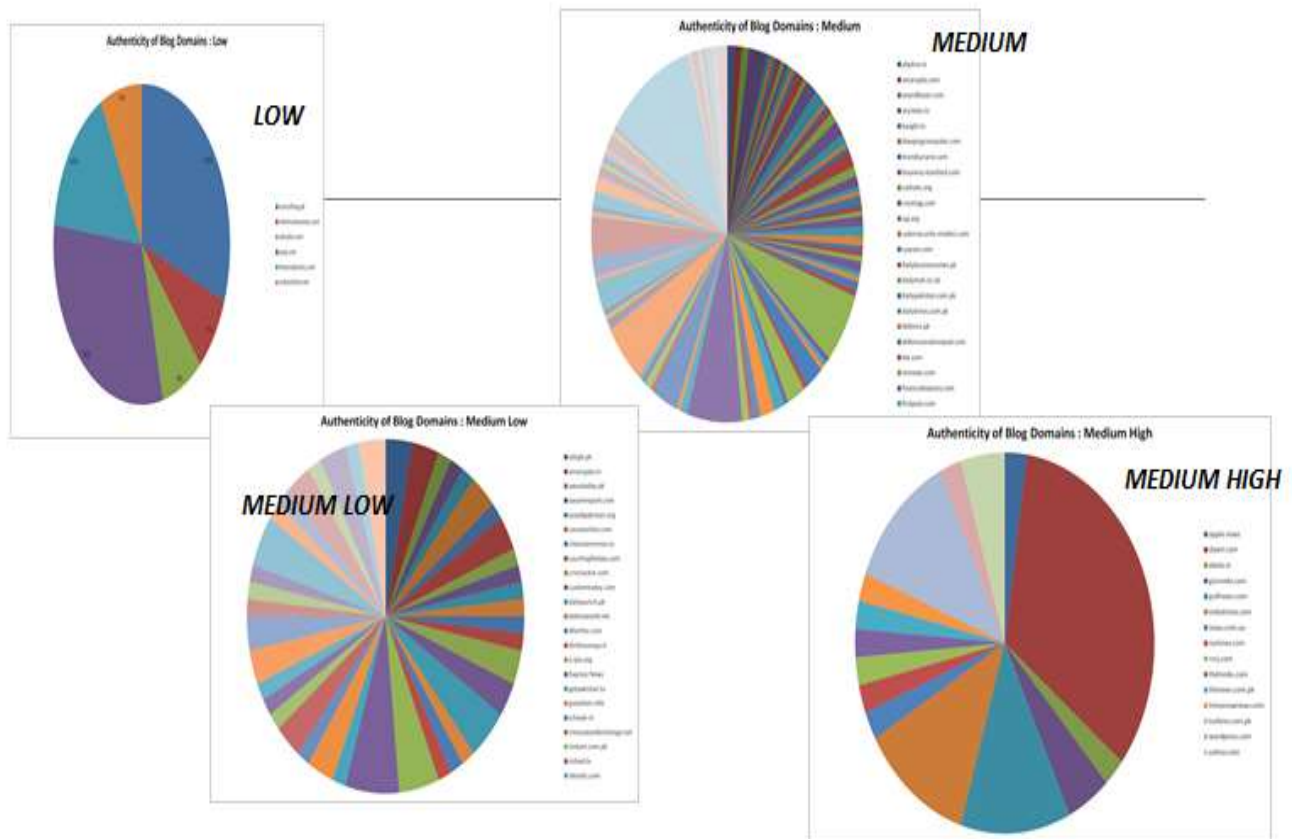


Figure 5.3: Domain Graphs as per the Verification Levels

In the graphs shown in Fig. 5.3, data is given as per the domain ranking, as the trend of articles shows that maximum articles lie down in the “medium high” category. Thus maximum domain lies in “medium high” category of source verification. We have obtained the graph of highest domains falling in “Medium High” category of source verification. Fig. 5.4 shows the highest domains falling in “Medium High” category of source verification.

The Fig. 5.4 shows that “dawn news” presents the highest number of cases having the “medium high” level. These stats are as per the data of cybercrimes in Pakistan. The data clearly depicts that international media is almost silent in reporting the Pakistani local cybercrimes. However, local news has actively participated in reporting the cybercrimes that took place in Pakistan.

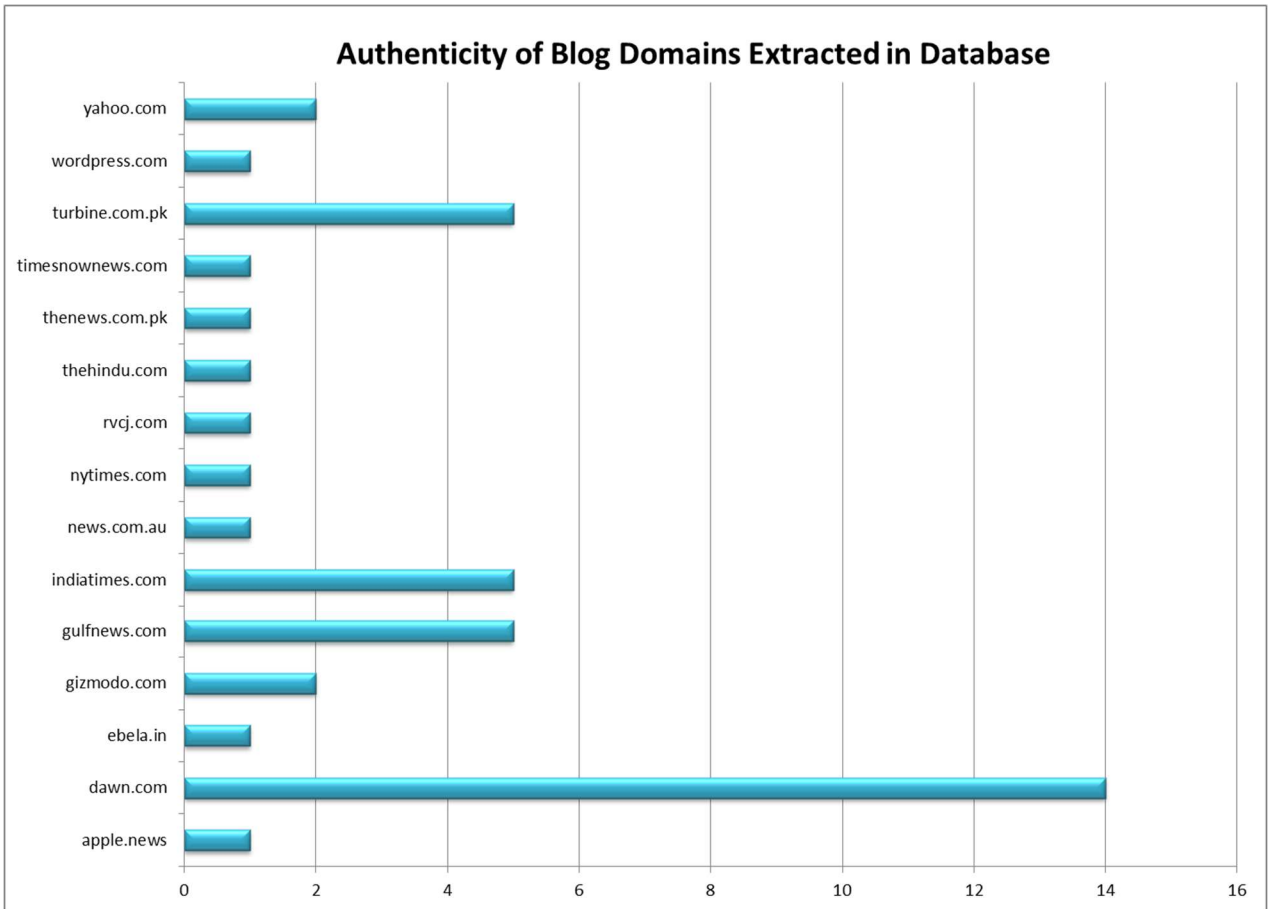


Figure 5.4: Blog Domains having Medium High Category

Further analysing the graphs showed in Fig. 5.3, where data is given as per the domain ranking, the trend of articles showed that huge number of articles lie down in the low category, thus, maximum domain lies in “low” category of source verification. We have obtained the graph of highest domains falling in “low” category of source verification. Fig. 5.5 shows the highest domains falling in “low” category of source verification.

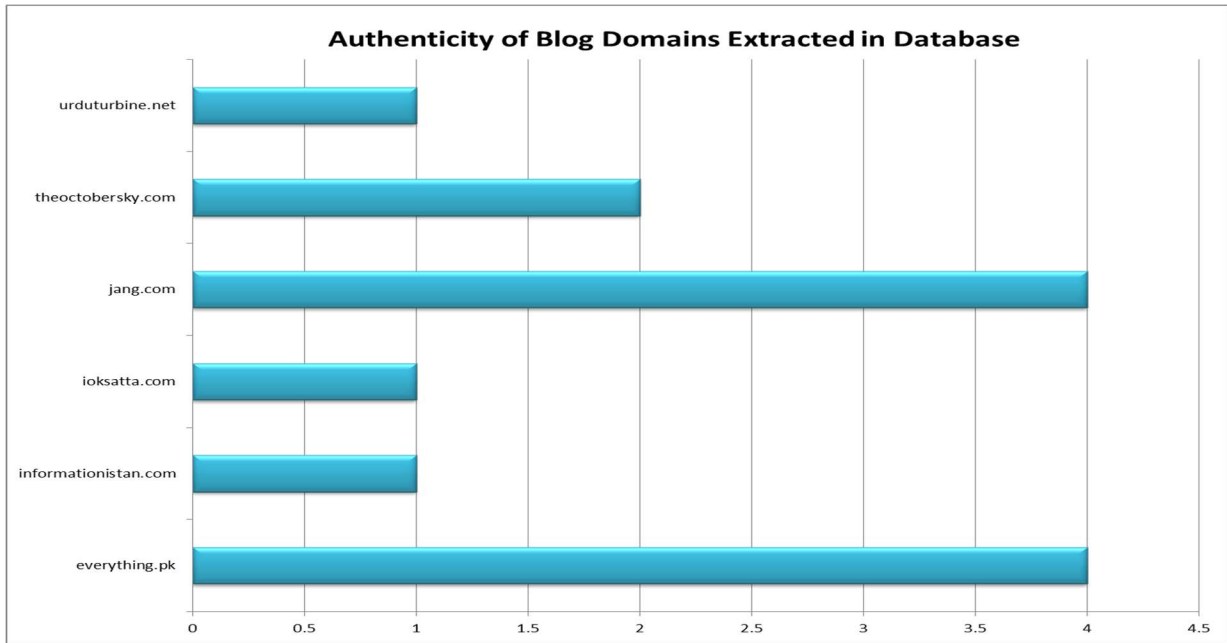


Figure 5.5: Blog Domains having Low Category

The Fig. 5.5 showed that “everything.pk” and “jang.com” presents the highest number of cases having the “low” level. These stats are as per the data of cybercrimes in Pakistan. The data clearly depicts that international media is almost silent in reporting the Pakistani local cybercrimes. However, local news has actively participated in reporting the cybercrimes that took place in Pakistan. The most important thing is here that data that has been obtained from non-reputable and ambiguous sites are mostly regarded as low level. It clearly indicates as the sites are not trust worthy and neither SSL (Secure Socket Layer) protected, their data can be altered at any time or there might be a chance that the data here entered is fake, just to attract the attention of viewers.

5.6 Limitations

Some factors that make it harder to scrape a site include

- i. Badly formatted HTML code with little or no structural information
- ii. Authentication systems that are supposed to prevent automatic access e.g. CAPTCHA codes and pay walls
- iii. Session-based systems that use browser cookies to keep track of what the user has been doing
- iv. A lack of complete item listings and possibilities for wildcard search

- v. Blocking of bulk access by the server administrators

5.7 Conclusion

Data analysis performed on extracted information from online blogs and articles revealed that they have different levels of authenticity and reliability as per our defined criterion. Most of the information lies in category of “Medium High”. It was due to the fact that these verification parameters didn’t belong to some governmental or legal proceeding of an incident. Further, the proposed model had some limitations. Some factors made scrapping process difficult such as badly formatted HTML code with little or no structural information. As every article has its own define HTML tags, therefore, scrapping each blog required detailed inspection of their HTML code discovering respective HTML tags. Further, bulk or multiple accesses were being blocked by server administrators considering user as a bot.

Conclusion and Future Considerations

6.1 Introduction

A brief overview of research work along with future directions is illuminated in this chapter.

6.2 Overview

With rapid technological advancements, cybercrimes are on a high rise especially in countries who are in list of developing countries such as Pakistan. This paper presents a query based search model for extraction of events related to cybercrimes in Pakistan. Information is extracted from crowdsourcing platforms such as blogs/articles and results are stored in form of well-organized database having fixed parametric arrays. We also presented a method of source verification of data acquired from online blogs, which, to best of our knowledge, has not been presented before.

Data is extracted from online blogs via two different approaches i.e. the first one being building customized scrappers for each blog domain & the other one was to extract data via content search engines. Source verification is performed by defining quintuple elements. The data required for process of source verification was obtained from SEO small tools (an open source tool built by Google) while some of the data (social stats) were acquired during information extraction process via content search engines. Five authenticity levels were designed including High, Medium High, Medium, Medium Low & Low. The values for quintuple elements were scaled down through feature scaling/normalization technique & cumulative results of these elements were represented in form of authenticity levels.

Results obtained have been analyzed with help of several graphs. The results displayed that maximum blogs reached to “Medium High” level of authenticity while none of the blogs scrapped reached to high level of source verification as per our defined criterion. It seems quite

true because we focused on data available from crowdsourced platforms (articles/blogs) & we haven't added the verification parameter from legal or governmental site such as NR3C which can be done as future project. The analysis results also indicated that international media is silent in reporting cybercrimes in Pakistan while these crimes are being reported by local news actively.

6.3 Future Considerations

The proposed model can be continued further for purpose of risk analysis. Risk analysis will help in finding out the most vulnerable cybercrimes in society. Moreover, legal proceeding and their results must be considered to declare the validity of an article. For instance, Federal investigation Agency (FIA) updates their official website regarding crime cases dealt by them in form of images published in newspapers. Extracting data from these images could be considered as a valid parameter for process of source verification of articles which could be done as a future work.

Source Code

Methodology 1:

#####Link Grabbing#####

```
import re
import linkGrabber
links_1 = linkGrabber.Links
('https://www.google.com/search?source=hp&ei=JajRXPqIEtCymwXsioTQBwandq=cyber
crime+in+pakistan+2019&btnK=Google+Search&doq=cybercrime+in+pakistan+2019a
ndgs_l=psy-ab.3..0i22i30.1302.6295..7128...0.0..0.365.7186.2-25j2.....0....1..gws-
wiz.....0..0i131j0i10j0i22i10i30.sd4A7o8DqwM')
gb_1= links_1.find (limit=1000, duplicates=False, pretty=True)
print(gb_1)
```

#####URL Cleaning#####

```
Query (keyword)
View (https://..)
{url, text_url, title_url} = Linkgrabber (site_url); //python coding
// url cleaning
If title_url  $\cong$  {set of keywords} || text_url  $\cong$  {set of keywords}
then
Save (url)
else
return
```

#####Spider for Articles#####

******Pakobserver.py******

```
# -*- coding: utf-8 -*-  
import scrapy  
class ThenationSpider(scrapy.Spider):  
    name = 'pakobserver'  
    allowed_domains = ['www.pakobserver.net']  
    start_urls = ['https://pakobserver.net/cybercrime-in-pakistan/']  
    def parse(self, response):  
        #Extracting the content using css selectors  
        titles = response.css('.entry-title::text').extract()  
        date = response.css('.entry-date::text').extract()  
        author = response.css("p").extract()  
        #Give the extracted content row wise  
        for item in zip(titles,date,author):  
            #create a dictionary to store the scraped info  
            scraped_info = {  
                'title' : item[0],  
                'date' : item[1],  
                'author' : item[2],  
            }  
            #yield or give the scraped info to scrapy  
            yield scraped_info
```

******thenation.py******

```
# -*- coding: utf-8 -*-  
import scrapy  
class ThenationSpider(scrapy.Spider):  
    name = 'thenation'  
    allowed_domains = ['www.nation.com.pk']  
    start_urls = ['https://nation.com.pk/27-Oct-2017/rise-of-cybercrime-in-pakistan']  
    def parse(self, response):  
        #Extracting the content using css selectors
```

```

titles = response.css('title::text').extract()
date = response.css('.meta-date::text').extract()
author = response.css('.authorname').extract()
#Give the extracted content row wise
for item in zip(titles,date,author):
    #create a dictionary to store the scraped info
    scraped_info = {
        'title' : item[0],
        'date' : item[1],
        'author' : item[2],
    }
    #yield or give the scraped info to scrapy
    yield scraped_info

```

Methodology 2:

#####Building a Scrapper using python language#####

```

import bs4
import selenium
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from bs4 import BeautifulSoup
import time
import csv

from selenium.webdriver.common.keys import Keys

url =
"https://app.buzzsumo.com/research/content?logged_in=1andnum_days=365andresult_type="

```

```
totalandgeneral_articleandinfographicandvideoandhow_to_articleandwhat_postandwhy_post  
andlistandtab=0andpage=1andq=cyber%20attacks%20in%20pakistan"
```

```
driver = webdriver.Chrome()
```

```
driver.implicitly_wait(30)
```

```
driver.get(url)
```

```
time.sleep(20)
```

```
soup = BeautifulSoup(driver.page_source,'lxml')
```

```
titles = soup.find_all('a',attrs={'class':'search-title'})
```

```
#Initialize stats arrays/lists
```

```
author_a = []
```

```
domain_a = []
```

```
date = []
```

```
facebook_stats = []
```

```
twitter_stats = []
```

```
pinterest_stats = []
```

```
reddit_stats = []
```

```
links_stats = []
```

```
evergreen_stats = []
```

```
total_stats = []
```

```
author_names = []
```

```
headers = ['Title','Author Name','Domain','Date','Facebook Engagements',  
           'Twitter Engagements','Pinterest Engagements','Reddit Engagements',  
           'Number of Links','Evergreen Score','Total Engagements']
```

```
#####Get social stats#####
```

```
#facebook
```

```
facebook_div = soup.find_all('div',attrs={'class':'stat facebook'})
```

```
for div in facebook_div:
```



```

facebook_stats.append(div.find('span'))

#twitter
twitter_div = soup.find_all('div',attrs={'class':'stat twitter'})
for div in twitter_div:
    twitter_stats.append(div.find('span'))

#pinterest
pinterest_div = soup.find_all('div',attrs={'class':'stat pinterest'})
for div in pinterest_div:
    pinterest_stats.append(div.find('span'))

#Reddit
reddit_div = soup.find_all('div',attrs={'class':'stat reddit'})
for div in reddit_div:
    reddit_stats.append(div.find('span'))

#Links
links_div = soup.find_all('div',attrs={'class':'stat links'})
for div in links_div:
    links_stats.append(div.find('span'))

#evergreen
evergreen_div = soup.find_all('div',attrs={'class':'stat evergreen'})
for div in evergreen_div:
    evergreen_stats.append(div.find('span'))

#total
total_div = soup.find_all('div',attrs={'class':'stat total'})
for div in total_div:

```

```

total_stats.append(div.find('span'))

#####Get date#####
date = soup.find_all('div',attrs={'class':'published-date'})

#####Get author names#####
author_div = soup.find_all('div',attrs={'class':'author-name'})
for div in author_div:
    author_a.append(div.find('a'))

for author in author_a:
    author_names.append(author.get_text().strip()[3:])

#####Get domains#####
domain_div = soup.find_all('div',attrs={'class':'content-url'})
for div in domain_div:
    domain_a.append(div.find('a'))

column = [titles,author_names,domain_a,date,facebook_stats,
twitter_stats,pinterest_stats,reddit_stats,links_stats,
evergreen_stats,total_stats]

#####Write to csv file#####
file = open('buzz.csv','w',newline="",encoding='utf-8')
writer = csv.writer(file)
writer.writerow(headers)
row=[]

for i in range(5):
    for j in range(11):
        if j == 1:

```

```
        row.append(column[j][i])
    else:
        row.append(column[j][i].get_text())
    writer.writerow(row)
    row=[]
file.close()
```

References

- [1] John Ombelets, "Crowdsourcing Cybersecurity"; [Online]. Available: <https://medium.com/cxo-magazine/crowdsourcing-cybersecurity-7fa834beafe6>. Accessed: April 11, 2019
- [2] Spencer Ackermen, "Data for the Boston Marathon Investigation will be Crowdsourced"; [Online]. Available: <https://www.wired.com/2013/04/boston-crowdsourced/>. Accessed: April 11, 2019
- [3] Daniel Dimov, "Crowdsourcing Cybersecurity: How to Raise Security Awareness Through Crowdsourcing"; [Online]. Available: <https://resources.infosecinstitute.com/crowdsourcing-cybersecurity-how-to-raise-security-awareness-through-crowdsourcing/#gref>. Accessed: April 12, 2019
- [4] Sultan Ullah et al., "Pakistan and cyber crimes: Problems and preventions", 2015 First International Conference on Anti-Cybercrime (ICACC), November, 2015.
- [5] Kundi et al., "Digital Revolution, Cyber-Crimes And Cyber Legislation: A Challenge To Governments In Developing Countries", Journal of Information Engineering and Applications, February 2014.
- [6] Thomas K.. "Crowdsourcing from its beginnings to the present"; [Online]. Available: <https://www.clickworker.com/2018/04/04/evolution-of-crowdsourcing/>. Accessed: April 21, 2019.
- [7] Alec Lynch, "Crowdsourcing is Not New: The History of Crowsourcing(1714 to 2010)"; [Online]. Available: <https://blog.designcrowd.com/article/202/crowdsourcing-is-not-new--the-history-of-crowdsourcing-1714-to-2010>. Accessed: April 21, 2019.
- [8] Daren C. Brabham, *Crowdsourcing* (2013), p. xix
- [9] Lalit Mohan S et al., "Crowdsourcing Security: Opportunities and Challenges", 2018 ACM/IEEE 11th International Workshop on Cooperative and Human Aspects of Software Engineering.
- [10] Anas Baig, "Crowdsourcing and Cybersecurity- A Supercell Against Cyber Threats"; July 2017 [Online]. Available: <https://crowdsourcingweek.com/blog/crowdsourcing-cybersecurity/>. Accessed: April 22, 2019.

- [11] Nicky Antonius and L. Rich, "Discovering collection and analysis techniques for social media to improve public safety", *The International Technology Management Review*, Vol. 3 (2013) No. 1, p. 43.
- [12] Ryan Goodrich, "What is Crowdsourcing?" [Online]. Available: <https://www.businessnewsdaily.com/4025-what-is-crowdsourcing.html>. Accessed: April 23, 2019.
- [13] Rumsfeld, John S., et al. "Use of mobile devices, social media, and crowdsourcing as digital strategies to improve emergency cardiovascular care: a scientific statement from the American Heart Association." *Circulation* 134.8 (2016): e87-e108.
- [14] Kate Starbird, "Digital Volunteerism During Disaster: Crowdsourcing Information Processing ", *The ACM Conference on Human Factors in Computing Systems (CHI)*, 2011.
- [15] Chul Hyun Park and Eric W. Johnston, "A framework for analyzing digital volunteer contributions in emergent crisis response efforts ", *Article in New Media and Society* · August 2017.
- [16] Barbora Haltofová, "Implementation of Geo-Crowdsourcing Mobile Applications in e-Government of V4 Countries: A State-of-the-Art Survey", *World Academy of Science, Engineering and Technology International Journal of Information and Communication Engineering* Vol:11, No:5, 2017.
- [17] Maisonneuve, Nicolas and Stevens, Matthias and Niessen, Maria and Steels, Luc. (2009), "NoiseTube: Measuring and mapping noise pollution with mobile phones", *Information Technologies in Environmental Engineering*. 215-228. 10.1007/978-3-540-88351-7_16.
- [18] P. Tran-Gia, T. Hossfeld, M. Hartmann, and M. Hirth, "Crowdsourcing and its Impact on Future Internet Usage", *it - Information Technology Methoden und innovative Anwendungen der Informatik und Informationstechnik*, Vol. 55, No. 4, 2013, pp. 139-145.
- [19] Hogenboom, FP and Frasinca, Flavius and Kaymak, Uzay and de FMG, Jong. (2011), "An overview of event extraction from text", *CEUR Workshop Proceedings*. 779.
- [20] Sunita Sarawagi (2008), "Information Extraction", *Foundations and Trends® in Databases*: Vol. 1: No. 3, pp 261-377.

- [21] Doddington, George and Mitchell, Alexis and Przybocki, Mark and Ramshaw, Lance and Strassel, Stephanie and Weischedel, Ralph. (2004), "The Automatic Content Extraction (ACE) program-tasks, data, and evaluation", Proceedings of LREC. 2.
- [22] Appelt, Douglas E. "Introduction to information extraction." *Ai Communications* 12, no. 3 (1999): 161-172.
- [23] Kuila, Alapan, and Sudeshna Sarkar, "An Event Extraction System via Neural Networks", *FIRE (Working Notes)*, 2017.
- [24] Changsheng Wan, Juan Zhang, and Daoli Huang, "SCPR: Secure Crowdsourcing-Based Parking Reservation System," *Security and Communication Networks*, vol. 2017, Article ID 1076419, 9 pages, 2017.
- [25] Tentilucci, Matthew et al. "Crowdsourcing Computer Security Attack Trees." (2015).
- [26] Klettner, Silvia and Huang, Haosheng and Schmidt, Manuela and Gartner, Georg. (2013), "Crowdsourcing affective responses to space", *Kartographische Nachrichten, Journal of Cartography and Geographic Information*. 63. 66-73.
- [27] M. Strohmeier, M. Smith, M. Schäfer, V. Lenders and I. Martinovic, "Crowdsourcing security for wireless air traffic communications," *2017 9th International Conference on Cyber Conflict (CyCon)*, Tallinn, 2017, pp. 1-18.
- [28] Michael White, Tanya Korelsky, Claire Cardie, Vincent Ng, David Pierce, and Kiri Wagstaff. 2001, "Multidocument summarization via information extraction", In *Proceedings of the first international conference on Human language technology research (HLT '01)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1-7.
- [29] Ji, Heng and Grishman, Ralph and Chen, Zheng and Gupta, Prashant. (2019), "Cross-document Event Extraction and Tracking: Task, Evaluation, Techniques and Challenges".
- [30] M. Granitzer *et al.*, "Analysis of machine learning techniques for context extraction," *2008 Third International Conference on Digital Information Management*, London, 2008, pp. 233-240.
- [31] Maisonneuve, Nicolas and Stevens, Matthias and Niessen, Maria and Steels, Luc. (2009), "NoiseTube: Measuring and mapping noise pollution with mobile phones", *Information Technologies in Environmental Engineering*. 215-228. 10.1007/978-3-540-88351-7_16.

- [32] Ingram, David G., Camilla K. Matthews, and David T. Plante. "Seasonal trends in sleep-disordered breathing: evidence from Internet search engine query data." *Sleep and Breathing* 19, no. 1 (2015): 79-84.
- [33] Mavandadi et al., "Distributed medical image analysis and diagnosis through crowd-sourced games: a malaria case study", *PLoS One*. 2012; 7(5):e37245
- [34] Oczan A, "Educational games for malaria diagnosis", *Sci Transl Med*. 2014 Apr 23; 6(233):233ed9.
- [35] Luengo-Oroz MA, Arranz A, Frean J, "Crowdsourcing malaria parasite quantification: an online game for analyzing images of infected thick blood smears", *J Med Internet Res*. 2012 Nov 29; 14(6):e167.
- [36] Freifeld CC, Chunara R, Mekaru SR, Chan EH, Kass-Hout T, Iacucci AA, et al, "Participatory epidemiology: use of mobile phones for community-based health reporting", *PLoS Med*. 2010;7:e1000376. doi: 10.1371/journal.pmed.1000376.
- [37] Mladenow, Andreas and Bauer, Christine and Strauss, Christine, "Crowdsourcing in Logistics: Concepts and Applications Using the Social Crowd", *International Conference on Information Integration and Web-based Applications and Services*, December 2015. 10.1145/2837185.2837242.
- [38] Kundi, GM. (2010), "E-Business in Pakistan: Opportunities and Threats", Lap-Lambert Academic Publishing, Germany.
- [39] Barr, R. and Pease, K. (1990), "Crime placement, displacement, and deflection", in: M. Tonry and N. Morris (eds), *Crime and Justice: A Review of Research*, 12(3): 12-23, University of Chicago Press, Chicago.
- [40] Levi, M. (1998), "Organized plastic fraud: Enterprise criminals and the side-stepping of fraud prevention", *The Howard Journal*, 37(4): 423-38.
- [41] Madhava S.S.P., and Umarhathab, S. (Eds.), (2011), "Information Technology Act and cyber terrorism: A critical review", *Cyber Crime and Digital Disorder*, Tirunelveli, India: Publications Division, Manonmaniam Sundaranar University.
- [42] Magalla, A. (2013), "Security, prevention and detection of cyber-crimes in Tanzania", *Doctoral Thesis*, Tumaini University Iringa University College, [Online]. Available at: http://www.academia.edu/3471542/the_introduction_to_cybercrime_security_prevention_and_detection_of_cybercrime_in_tanzania, (March 26, 2014).

- [43] Gupta, Ravi and Hugh Brooks, "Using social media for global security", John Wiley and Sons, 2013.
- [44] Tewksbury, Doug. (2012), "Crowdsourcing Homeland Security: The Texas Virtual BorderWatch and Participatory Citizenship. *Surveillance and Society*", 10. 249-262. 10.24908/ss.v10i3/4.3464.
- [45] Hui, Jennifer Yang. "Crowdsourcing for national security." *Nanyang Technological University* (2015): 1-15.
- [46] Shiffman, Gary M., and Ravi Gupta. "Crowdsourcing cyber security: a property rights view of exclusion and theft on the information commons." *International Journal of the Commons* 7.1 (2013): 92-112.
- [47] Imran Mukhtar, "Rising Cyber crimes become challenge for FIA"; [Online], Available: <https://nation.com.pk/14-Apr-2019/rising-cyber-crimes-become-challenge-for-fia>. Accessed: May 11, 2019.
- [48] Das, S. and nayak, T. (2013). *Impact of Cyber Crime: Issues and Challenges*. *International Journal Engineering Science and Engineering Technologies*, 6(2), 142-153.
- [49] Aima Nisar, "Facing online harassment? Pakistan law is on your side"; [Online], Available: <https://www.samaa.tv/opinion/2019/03/facing-online-harassment-pakistan-law-is-on-your-side/>. Accessed: May 11, 2019.
- [50] Awan, Jawad and Memon, Shahzad. (2016), "Threats of Cyber Security and Challenges for Pakistan", 11th International Conference on Cyber Warfare and Security, At USA, March 2016.
- [51] Afeera Firdous, "Cyber Security issues in Pakistan"; [Online]. Available: <https://ciss.org.pk/cyber-security-issues-in-pakistan/>. Accessed: May 11, 2019.
- [52] Anas Baig, "How Crowdsourcing Can Help Fight Against Cyber Threats"; [Online]. Available: <https://tech.co/news/cybersecurity-crowdsourcing-online-threats-2017-09>. Accessed: May 11, 2019.
- [53] Yannis Charalabidis, Euripidis N. Loukis, Aggeliki Androutopoulou, Vangelis Karkaletsis, Anna Triantafillou, (2014) "Passive crowdsourcing in government using social media", *Transforming Government: People, Process and Policy*, Vol. 8 Issue: 2, pp.283-308.

- [54] D. Galinec, D. Možnik and B. Guberina. "Cybersecurity and cyber defence: national level strategic approach", *Automatika*, 58:3, 273-286, DOI: 10.1080/00051144.2017.1407022, 2017.
- [55] E. Agichtein and L. Gravano, "Querying text databases for efficient information extraction," *Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405)*, Bangalore, India, 2003, pp. 113-124.
- [56] Steiner, Thomas, Ruben Verborgh, Rik Van de Walle, et al. "Crowdsourcing Event Detection in YouTube Video." *Proceedings of the 1st Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*. Ed. Marieke Van Erp et al. 2011. 58–67.
- [57] Steiner, Thomas and Verborgh, Ruben and Gabarró, J and Hausenblas, Michael and Troncy, Raphaël and Van de Walle, Rik. (2012). Enabling on-the-fly Video Shot Detection on YouTube.
- [58] NB Sristy, NS Krishna, DVLN Somayajulu, "Event Extraction from Social Media text using Conditional Random Fields", *FIRE (Working notes)*, 2017.
- [59] M. Imran, C. Castillo, F. Diaz, and S. Vieweg. "Processing social media messages in mass emergency: A survey". arXiv preprint arXiv:1407.7071, 2014.
- [60] H. Kwak, C. Lee, H. Park, and S. Moon. "What is twitter, a social network or a news media? ", In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
- [61] S. Lewandowsky, U. Ecker, C. Seifert, N. Schwarz, and J. Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106-131, 2012.
- [62] Zubiaga, Arkaitz, Maria Liakata, Rob Procter, Kalina Bontcheva and Peter Tolmie. "Crowdsourcing the Annotation of Rumourous Conversations in Social Media." *WWW*(2015).
- [63] Flora S. Tsai and Kap Luk Chan. 2007. Detecting Cyber Security Threats in Weblogs Using Probabilistic Models. In *Proc. PAISI'07*.
- [64] Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. 2016. Acing the IOC Game: Toward Automatic Discovery and Analysis of Open-Source Cyber Threat Intelligence. In *Proc. CCS'16*.
- [65] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams.

In ICWSM, 2009.

[66] Carl Sabottke, Octavian Suciu, and Tudor Dumitras. 2015. Vulnerability Disclosure in the Age of Social Media: Exploiting Twitter for Predicting Real-World Exploits. In Proc. USENIX Sec'15.

[67] Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. 2015. Weakly Supervised Extraction of Computer Security Events from Twitter. In Proc. WWW'15.

[68] Khandpur, R.P., Ji, T., Jan, S., Wang, G., Lu, C.T., Ramakrishnan, N, "Crowdsourcing cybersecurity: Cyber attack detection using social media", arXiv preprint arXiv:1702.07745(2017)

[69] Michael OvelgÅlonne, Tudor Dumitras, B. Aditya Prakash, V. S. Subrahmanian, and Benjamin Wang. 2016. Understanding the Relationship between Human Behavior and Susceptibility to Cyber-Attacks: A Data-Driven Approach. In Proc.TIST'16.