Systematic Analysis and Visualization of Privacy Policies of Online Services

Dhotre, Prashant Shantaram

DOI (link to publication from Publisher):
10.5278/vbn.phd.tech.00025

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Dhotre, P. S. (2017). *Systematic Analysis and Visualization of Privacy Policies of Online Services*. Aalborg Universitetsforlag. Ph.d.-serien for Det Tekniske Fakultet for IT og Design, Aalborg Universitet https://doi.org/10.5278/vbn.phd.tech.00025

SYSTEMATIC ANALYSIS AND VISUALIZATION OF PRIVACY POLICIES OF ONLINE SERVICES

BY
PRASHANT S. DHOTRE

DISSERTATION SUBMITTED 2017

AALBORG UNIVERSITY
DENMARK

SYSTEMATIC ANALYSIS AND VISUALIZATION OF PRIVACY POLICIES OF ONLINE SERVICES

by

Prashant S. Dhotre

AALBORG UNIVERSITY
DENMARK

Dissertation submitted

.

Dissertation submitted:	September 2017
PhD supervisor:	Associate Prof. Henning Olesen, CMI Aalborg University
PhD committee:	Associate Professor Lene Tolstrup Sørensen (chairman) Aalborg University Professor Jari Porras Lappeenranta University of Technology Reader Dr. James M. Irvine University of Strathclyde
PhD Series:	Technical Faculty of IT and Design, Aalborg University
Department:	Department of Electronic Systems

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-077-7

Published by:
Aalborg University Press
Skjernvej 4A, 2nd floor
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2017

Dedicated to…
my mother (Shashikala) and my wife (Deepali)

CV

Mr. Prashant. S. Dhotre has obtained his B.E degree in Computer Science and Engineering from Shri Ramanand Teertha Marathwada University, Nanded, India in 2004 and M.E. degree in Information Technology from Savitribai Phule Pune University, Pune, India. He has more than 13 years of teaching and research experience.

He has authored 2 books on a subject "Theory of Computations" for UG students of Computer Engineering.

He loves to interact with the students through a series of guest lectures on "Design and Analysis of Algorithms", "Theory of Computation", and "Data Structures" in various Engineering Colleges across the Pune region.

As a part of his research, he has delivered several talks on "research, how to write a research proposal, how to write research papers, Use of reference manager and citation tools, etc." at Institute and at University level.

Currently, he is working as Assistant Professor in Department of Computer Engineering, STES's Sinhgad Institute of Technology and Science, Pune, India. He has guided more than 50 undergraduate students for projects. His research interests include Algorithms, Privacy, and Turing Machine. He has visited a few countries like Denmark, Sweden, China, and Germany.

ENGLISH SUMMARY

Due to the advancement in mobile and wireless communications in today's digital world, Internet services like social networks, search engines, etc., have brought many benefits to the users. However, most of the services also collect excessive information about the users and their day-to-day activities online. Using "Big Data" technologies, the user information is collected and analyzed by the service providers to improve their services, an approach giving rise to several privacy concerns.

For service providers, user information has become an important part of their business model and an economic asset. On the service provider side, the user information is collected, stored, processed, and analyzed to get additional value from it, often without the users' consent, which constitutes a major privacy risk. Once the information has been disclosed, the users have no control over it.

Although the business practices of the service providers are usually specified in the form of privacy policies (terms of use), these documents are time-consuming to read and complicated to understand, and users do not really know what happens to their data. Hence, increasing the privacy awareness is an important means to empower the users towards the service providers.

Presently, several privacy awareness tools, e.g. website rating tools (based on users' experience), and blocking tools (blocking hidden data trackers, advertisers, third parties, etc.) are available. However, there is still a clear need to increase the user's privacy awareness and assist them in understanding the content of privacy policies. This was confirmed by a comprehensive survey with Indian users, which was carried out during this project.

In the thesis, a new Privacy Policy Elucidator Tool (PPET) is proposed and implemented. It is capable of classifying, summarizing and visualizing the contents of privacy policies of service providers. Using a Naïve Bayes approach, the PPET tool classifies the contents of privacy policies into different sections, dealing with the collection, sharing, usage, protection, and management of user information. The tool extracts and summarizes the policy content, provides a graphic visualization of it, and thereby assists the users to learn and understand the practices of service providers.

For test and performance evaluation of the PPET, a number of training and testing records in the form of a matrix was used. The PPET achieved more than 95 % accuracy for classification of privacy policies into predefined sections. This accuracy is also well supported by the analysis of the user feedback on the PPET. According to the user feedback, the PPET served to motivate users to read the privacy policies and helped them in enhancing their privacy awareness.

Another important result of this work is the detection that the current unstructured privacy policies do not comply with the general privacy design guidelines and privacy regulations. Hence, this thesis also proposes a standardized uniform template for the privacy policies, which is aligned with the upcoming EU General Data Protection Regulation (GDPR), which will be enforced from 2018.

DANSK RESUME

Som følge af udviklingen inden for mobil og trådløs kommunikation i dagens digitale verden har internet-tjenester, såsom sociale netværk, søgemaskiner, etc., medført mange fordele for brugere. Men samtidig indsamler de fleste tjenester en mængde information om brugerne og deres daglige aktiviteter online. Ved hjælp af 'Big Data'-teknologier indsamler og analyserer tjenesteudbyderne brugernes information med henblik på at forbedre deres tjenester, en fremgangsmåde, der potentielt giver reel mulighed for krænkelse af privatlivet.

For tjenesteudbyderne er information om brugerne blevet en vigtig del af deres forretningsmodel og et økonomisk aktiv. Tjenesteleverandøren indsamler, opbevarer, behandler og analyserer brugeroplysningerne for at skabe yderligere værdi, ofte uden brugernes samtykke, og dette udgør en væsentlig risiko for privatlivets fred. Når oplysningerne først en gang er blevet delt, har brugerne ikke længere kontrol over dem.

Selvom tjenesteudbyderne normalt specificerer deres forretningspraksis i form af ”privacy policies” eller ”terms of use” dokumenter, er disse ofte tidskrævende at læse og svære at forstå, og brugerne ved i realiteten ikke, hvad der sker med deres data. Øget opmærksomhed om privatlivsbeskyttelse er derfor et vigtigt middel til at styrke brugernes position i forhold til tjenesteudbyderne.

I dag findes der adskillige værktøjer, der bidrager til øget opmærksomhed omkring beskyttelse af personlige oplysninger, f.eks. værktøjer til vurdering af websteder (baseret på brugernes oplevelse) og blokeringsværktøjer (som blokerer skjulte data-sporingssystemer, annoncører, tredjeparter osv.) Der er imidlertid stadig et klart behov for at øge brugernes bevidsthed omkring privatlivsbeskyttelse og hjælpe dem med at forstå indholdet af tjenesteudbydernes privacy policies. Dette bekræftes af en omfattende brugerundersøgelse blandt indiske brugere, som blev foretaget i løbet af dette projekt.

I afhandlingen præsenteres et nyt værktøj, Privacy Policy Elucidator Tool (PPET), som er udviklet og implementeret i løbet af projektet. Det er i stand til at klassificere, opsummere og visualisere indholdet af privatlivspolitikker fra tjenesteudbydere. Ved hjælp af en ”Naïve Bayes” tilgang kan PPET-værktøjet klassificere indholdet af privatlivspolitikker i forskellige sektioner såsom indsamling, deling, brug, beskyttelse og håndtering af brugerinformation. Værktøjet uddrager og sammenfatter indholdet af en privacy policy, foretager en grafisk visualisering af det, og hjælper derved brugerne med at forstå tjenesteudbydernes praksis.

For at teste og vurdere ydeevnen af PPET er der udvalgt en række trænings- og test records, organiseret i form af en matrix. Resultaterne viste en nøjagtighed på over

95% fra PPET, hvad angår klassificeringen af privatlivspolitikker i relation til foruddefinerede sektioner. Denne nøjagtighed understøttes ligeledes af analysen af brugernes feedback på brugen af PPET. Ifølge brugernes feedback har PPET bidraget til at motivere brugerne til at læse privatlivspolitikkerne og hjulpet med at forbedre deres bevidsthed omkring privatlivsbeskyttelse.

Et yderligere vigtigt resultat af dette forskningsprojekt er, at de nuværende ustrukturerede privatlivspolitikker kun i ringe udstrækning overholder de generelle designprincipper og love for beskyttelse af persondata. Afhandlingen præsenterer derfor et forslag til en standardiseret og ensartet skabelon for privatlivspolitikker, som er struktureret i lighed med den kommende EU Persondataforordning (GDPR), der træder i kraft i 2018.

ACKNOWLEDGEMENTS

I am extremely grateful to my supervisor, Associate Professor Dr. Henning Olesen for his valuable guidance, scholarly inputs, motivation, and constant encouragement for completion of my doctoral thesis.

Dr. Henning Olesen is a very passionate supervisor with a positive disposition. Despite his busy schedule, he was always available to elucidate my doubts. He has supported very well during my ups and downs in the journey of the Ph.D. program. I consider it as an amazing opportunity to do my Ph.D. under his valuable guidance and to learn from his research expertise. Thank you, Professor Henning, for all your unconditional help and support.

I thank Professor Dr. Knud Erik Skouby, Director, CMI, Aalborg University, for the academic facilities and support provided to carry out my research at CMI. I would like to thank Associate Professor Dr. Samant Khajuria, who has been very supportive and encouraging, at various steps of Ph.D. I express my gratitude to both.

I would like to thank Prof. M. N. Navale, Hon’ble President of STES, Dr. Mrs. S. M. Navale, Secretory, STES, Vice President (HR), Mr. Rohit M. Navale, Vice President (Admin), Mrs. Rachana M. Navale-Ashtekar, Vice President (Admin), and Dr. A. V. Deshpande sir, Dean, STES for giving me the opportunity to do a doctoral research in Alborg University, Denmark.

I am thankful to Prof. Ramjee Prasad Sir, Dr. S. S. Inamdar Sir for being a source of inspiration and initiating the GISFI Ph.D. programme for the teachers of STES. I would like to thank Dr. S. N. Mali sir, for motivation and support for completion of the doctoral program. I am thankful to Dr. Rajesh Prasad sir, for encouragement and support.

I honestly acknowledge Dr. P. N. Mahalle, for being always a motivator and my elder brother. I would like to thank Dr. Mrs. V. M. Rohokale for her kind support. I also thank my Heads of the department Prof. Mrs. Geeta Navale and Prof. Nihar Ranjan for supporting and understanding me. I am thankful to all faculty and staff of

Department of Computer Engineering, SITS, Narhe, Pune for their valuable support and help.

I am obliged to my friends, Mahendra Salunke, Geetanjali Shinde, Sushant Gote Prateek and Ambuj, Rahul Madke, Vivek Jog, Anurag for their encouragement and suggestions during the Ph.D. studies.

I thank my beloved parents- Dr Shantram Dhotre and Mrs. Shashikala Dhotre, my beloved in-laws- Mr. Arunrao Lokare and Mrs. Chandrakala Lokare for blessing and support. Last but not the least, I would like to thank my two kids *Chaitanya and Harssh* for cheerful smiles and my wife *Deepali* who is the best co-author of my life.

I would like to thank all those who have helped me directly or indirectly.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CCTV	**Closed-Circuit Television**
CLIP	Center On Law And Information Policy
DM	Data Mining
DPO	Data Protection Officers
EFF	Electronic Frontier Foundation
ENISA	European Network And Information Security Agency
EU	European Union
FIP	Fair Information Practices
FTC	Federal Trade Commission
GDPR	General Data Protection Regulation
IR	Information Retrieval
KDD	Knowledge Discovery In Databases
KNN	K Nearest Neighbors
LMP	Life Management Platform
LSA	Latent Semantic Analysis
ML	Machine Learning
NIST	National Institute Of Standards And Technology

NLP	Natural Language Processing
OECD	Organization Of Economic Cooperation And Development
P3P	Platform For Privacy Preferences
PbD	Privacy By Design
PET	Privacy Enhancing Technologies
PI	Personal Information
PII	Personally Identifiable Information
PIP	Personal Information Privacy
PPET	Privacy Policy Elucidator Tool
PPMLP	Privacy Policy Modeling Language Processor
SSL	Secure Socket Layer
SVM	Support Vector Machine
TC	Text Classification
TF-IDF	Term Frequency Inverse Document Frequency
TSV	Tab-Separated Values
VPN	Virtual Private Network
WOT	Web of Trust
XML	Extensible Markup Language

CHAPTER 1. INTRODUCTION

This chapter describes the discussion on the research context which will help to understand what is going and what is missing in the information flow between the users and the service providers. The challenges and the research questions are represented which is followed by the important research objectives. The chapter also discusses the research methodology. The chapter ends with the thesis organization.

1.1. INTERNET, USER, AND SERVICE PROVIDERS

Through mobile and wireless technologies, communicating, collaborating, and computing has resulted in a fast and easy access to a variety of services by the users. Significant benefits have been accrued, thereby transforming the lives of the users. More than 10 billion users have registered for more than 200 online social networking sites over the globe (Kumar, Jain, & Srivastava, 2016) that result in a large amount of communication about Internet users. The term "mass self-communication" by Manuel Castells denotes online sharing of a huge amount of text messages and posts by the Internet users (Castells, 2009). It is observed that 10 billion text messages and 1 billion posts are shared online every day.

The benefits of modern technologies are transforming and impacting our daily lives. A user starts his morning with e-news using *Inshorts*[1] App, birthday greetings are posted on *WhatsApp*, wedding/extraordinary events' photos are being shared on *Facebook*, *Google* Maps helps to find the destination, success stories are posted on blogs, information is stored on *Google* Drive, we make travel plans using *MakeMyTrip*[2], etc. (FTC, 2012). Hence, the services have made a huge transformation in human lives that has provided significant benefits to them.

On the one hand, these technologies are carrying us to the new era of a digital world where cars are driverless, patients are monitored remotely via wearable medical devices, smart refrigerators tell us list of items as per the expiry dates, etc. (Rule & Greenleaf, 2010). Users' are tracked 24x7 by other entities using modern technologies. This includes constant surveillance by security cameras or CCTV when we are at shopping malls or at holiday destinations. So, the fact is that the users are an integral part of the digital world. Hence, the wireless and mobile communication experience generate efficient and convenient services. Hence, the wireless and mobile communication experience yield into efficient and convenient services.

[1] *https://www.inshorts.com/*

[2] https://www.makemytrip.com/

On the other side, these technologies collect extensive information from users. Across the globe, the service providers are collecting users' information and extract additional value from it. As stated by the Meglena Kuneva (Kuneva, 2009) "Personal data is the new oil of the Internet and the new currency of the digital world", the users' information is vital and driving the global economy.

The information collected over social media is of 9 types like login details, messages, profiles, preferences, behavioral information, etc. (Beye et al., 2010). L. Salvatori et al. (Salvatori, 2015) claim that the information collected from users also includes movement of the cursor, websites visited, mouse clicks, friends' details, political orientation, search history. Further details include identifiable information, device details, payment details, user interaction, etc. (Zeadally & Winkler, 2016).

Therefore, the excessive information collection from the users gives rise to several privacy and security concerns. But, it is important to know what is privacy and security.

1.2. PRIVACY IS NOT ABOUT SECRECY

The terms privacy and security look similar. But, privacy should not be confused with security (Ortlieb, 2014). In naturalism, these two terms have different meanings based on the country laws, organizations, legal power, market sector, service providers.

In the digital world, we want to be protected from unauthorized access, the disclosure of secret information, etc. For the protection of users' information, the goals of security include- access control, authentication, integrity, etc. (Lou & Ren, 2009).

Security of information is important but it is comparatively well studied. However, privacy is a complex and multi-faceted concept which has become a fundamental issue is privacy (Lee & Janna, 2014).

Depending on the context, there are a variety of privacy concepts and understanding. Digging of personal or health information, or information is spied by third parties, misuse of personal information, etc. are few concerns about users' privacy. The concept of privacy, privacy principles, privacy policies, and laws are not the same across different countries (PrivacyPolicies.com, 2015). Hence, there is no specific standard for privacy.

The essential of privacy includes anonymity (hiding identity), limited access to the context, confidentiality, etc. As stated by Ann Cavoukian (Cavoukian, 2008), the privacy is about consent, choice, context, and respect. The personal information should be handled fairly for the said purposes with explicit user's consent.

In this thesis, the focus is given to privacy, privacy threats and issues, privacy enhancing technologies, and open problems.

1.3. MOTIVATION- THE “ACTUAL SCENE”

A typical interaction between the online user and the service provider is represented in fig 1-1.

Fig 1-1 consists of three things mainly a user, a service provider, and the guidelines or set of rules for the protection of privacy. In the beginning, the user requests a service from the service provider. Upon receiving the request, the service provider asks the user to register by sharing several personal attributes. The access to the services will be granted to the user once (s)he agrees to share those requested attributes. Each resource owner.

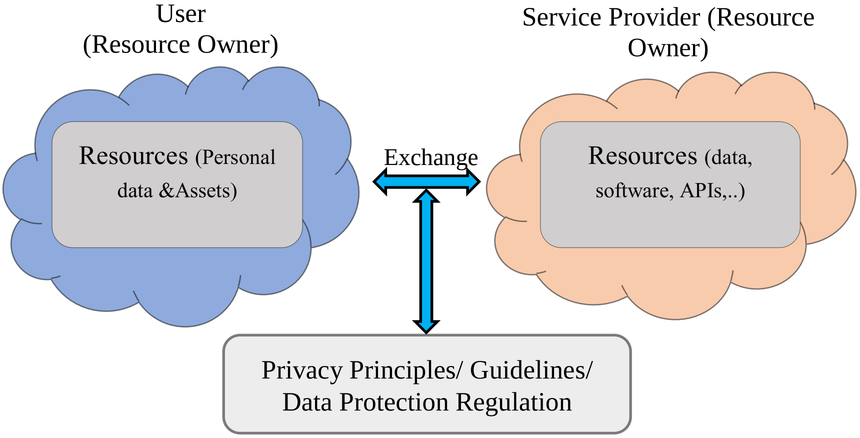

User
(Resource Owner)
Service Provider (Resource Owner)
Resources (Personal data &Assets)
Exchange
Resources (data, software, APIs,..)
Privacy Principles/ Guidelines/ Data Protection Regulation

Figure 1-1 Typical interaction with resource owners. This shows the flow of information as per the privacy guidelines

- Resource owner (user): Personal information of a user consists of attributes, choices, interests, context information, etc. This information is owned by the users and is valuable for service providers. To enable personalization, the user should provide certain information to service providers.

- Resource owner (service provider): A service provider offers the services to the users and gains an access to the users’ data. Before accessing the services, the service provider asks the users to accept the conditions. Service providers request users a few user attributes before allowing an access to the services. The set of attributes to be provided by the users are mentioned in their privacy policies.

- Privacy Principles/Guidelines/Data protection regulations: The interaction between user and service provider should be governed by the rules and regulations provided by the county law. Users' privacy should be maintained by following the guidelines that include user control and consent, minimum disclosure, transparency, purpose specification, data protection by design and by default, breach notifications etc.

  Privacy is governed by design principles and data protection laws. The principles of privacy given by Privacy by Design (PbD) (Cavoukian & Reed, 2013), the Organization for Economic Co-operation and Development (OECD) (OECD, 2013b) focus on the user control, notification, choice and consent, information use, minimum data disclosure, information quality and integrity, policy enforcement, etc. ("A structure for privacy - Teradata Magazine," 2007).

The service providers describe or characterize the users in terms of user's personal information, online behavior, assets, and interest, etc. The gathered information can be used in two ways.

1.3.1. POSITIVE WAY OF INFORMATION COLLECTION

The use of personal information is valuable for the Internet users. As per the users' interest, the service providers can provide personalized content and services. For example, if a person is checked into a hotel, the staff would give more attention about what he wants to have in a food and how should be making it available. The personal information will also be useful to provide the customizing user interface. Filtering of irrelevant information, identification and authentication are few examples where personal information is used for the purpose.

Using Big data technologies (Cavoukian & Jonas, 2012), the service providers analyze the collected information and accordingly the services are personalized as per the users' choice or interest. These technologies help the service providers to find which ads to show to users, who need what type of loan, which health tips to give, what is the next movie to watch, etc. There are lots of benefits of Big Data technologies that the users are eagerly enjoying by availing convenient and efficient services by service providers. This is a positive use of information collection.

1.3.2. NOT SO POSITIVE WAY OF INFORMATION COLLECTION

The escalation in the collection of users' information and its usage by the service providers have raised the privacy issues for the Internet users who own the data. personal information now has a social and an economic value. Organization for Economic Co-operation and Development (OECD) examined the prices of personal information like Social security number (8US$), driver's license number (US$ 3), date of birth (2 US$) [1]. The trade of user's data has given rise to a booming and highly lucrative market. The trade of personal information causes unease to the user as regards their privacy (OECD, 2013a).

Another trading example happened recently where a data breach incident was experienced by the Mail.ru, the most popular email service in Russia. Users' personal information has been accessed by compromising more than 270 million email accounts (David Lawler, 2016). The passwords were exchanged for performing criminal activities. Also, the stolen information was offered for almost $1.

Hence, Big data technologies and analytics sectors are increasing rapidly at the high rate and predicted to reach $16 billion by 2025[3]. It is well supported by the Tom Cochran, who said that "Personal Information is the currency of the 21st century" (Cochran, 2013).

1.3.3. WHAT HAPPENS TO USERS' INFORMATION

Based on the users' information and interaction with service providers, the groups of users' information-profiles- are created accurately. The profiles are helpful for service providers and other associated entities for targeted advertising. Further, such profiles are ready to sell to third parties without users' knowledge or explicit consent. These profiles are shared with other unknown entities without informing the users.

What most of us need to understand that the greatest threat to our privacy is arising from those companies/entities that we never heard of (Steve Kroft, 2014). The service providers share and sell users' information to the data brokers[4.5] The data brokers are the companies who gather users' useful information (users' education level, income, profession, habits, marital status, location, etc.) and based on collected information, they create the segments.

The segments are identifiable records like married customers, rich persons, educators, frequent buyer, travelers, etc. The data brokers keep such segments for sharing and trading it to other data brokers, government sectors, or advertisers, publishers, etc. So, anyone can buy the information like usernames, income details, habits, sexual orientation from such brokers and use it for any purposes.

[3] http://economictimes.indiatimes.com/tech/ites/big-data-analytics-to-reach-16-billion-industry-by-2025-nasscom/articleshow/52885509.cms

[4] http://www.cbsnews.com/news/data-brokers-selling-personal-information-60-minutes/

[5] http://www.economist.com/news/special-report/21615871-everything-people-do-online-avidly-followed-advertisers-and-third-party

Hence, the data brokers make money from sharing and selling the users' personal information to advertisers and publishers. Few examples of data brokers are Response solutions[6], Acxiom[7], Zifzi[8] who have lists of people having health issues, job seekers, credit card holders, etc.

Also, the users' information is accessed by the third parties. J. R. Mayer et al.(Mayer & Mitchell, 2012) demonstrate threats to users' privacy arises due to third parties. When a third party is embedded in the first party page, has access to users' information like users' personal, financial, medical, habitual, sexual information; resulting in information leakage and harm to others. Thus, information leakage lead to social or economic loss to the users.

1.3.4. ISSUES AND CHALLENGES

The interaction mentioned in fig 1-1 is monitored by third parties and indirectly by the data brokers. So, the "actual scene" is different. This "scene" is not known completely to all the users. Surprisingly, most of the users are not aware of the entities like data brokers and third parties, their practices, and associated risks to the users' privacy that arises from them. Hence, there is the need to spread awareness among Internet users about privacy.

Despite a fact that every individual has the right to privacy, a violation of users' privacy is observed in the developing countries like India. There are some instances where the nominal action has been taken over malpractices. The allegation on the *Unicommerce* (a firm owned by *Snapdeal* for inventory management) was illegally accessing confidential business information of the *PayTM* (Vishal Singh, 2016). The information was accessed from the registered *PayTM* sellers. The Delhi High Court ordered *Unicommerce* not to access and refrain from using the logo of advertisement.

The service providers should also inform the users about their strategies for personal information management. It must emphasize on the how effectively, they follow the privacy principles, guidelines, and laws. But a reality is quite different. For example, it is questionable to say that the services are secure (Nathaniel Mott, 2017). The WhatsApp users are more worried that their communication is not secure. Also, the sharing of WhatsApp users' data to Facebook (a parent company) has been leading to breach national laws of data protection in many countries like Germany (Natasha Lomas, 2016). Hence, the challenge is to enhance users' awareness about personal information management or practices stated in the form of privacy policies.

[6] http://www.cbsnews.com/news/data-brokers-selling-personal-information-60-minutes/

[7] https://www.acxiom.com/

[8] http://www.zifzi.com/index.php?route=product/category&path=20

If a user wants to avail a service offered by a service provider, the user has two choices. The first choice is that they must accept the privacy policy. The second choice to reject the policy and user cannot access the services. This strategy forces the users to accept all the terms and conditions even if the users may not agree with some of them. This situation for users is called as “Take it or Leave it” (Khajuria, Sørensen, & Skouby, 2017).

Also, during the installation of apps or accessing the services, the users are asked to share a bunch of specific users’ personal attributes to the service provider. The users are not informed clearly about the importance and purpose of disclosing it. Hence, it is important to focus on privacy policies.

The challenge includes how to limit information disclosure, it's processing, and misuse, and selling of personal information (Mowbray & Pearson, 2012). The lack of importance of knowledge about privacy and the inability to control the use of information results in increased privacy risk to users (Zengjie Cao, Yuanyuan Lin, & Cong Zhao, 2011).Hence, privacy awareness is important.

Therefore, most of the recently found issues are concerned to a user’s privacy and harvesting that lead to the changes in privacy rules and regulations (The European Parliment and The Council of the European Union, 2016). The new legislation is taking steps not only to secure user information but also protect its privacy.

Hence, the challenges on the users’ side are (Dhotre, Olesen, & Khajuria, 2017):

- How to deal with the privacy policies?
- How to enhance privacy awareness about privacy among users?
- How to empower users to control disclosure of personal information?

1.4. AIM, STATEMENT, AND SCOPE OF THIS RESEARCH

This thesis focuses on the systematic analysis of the privacy policies and its visualization. According to the state of art presented in the next chapter, there is a necessity to develop a tool that will help the Internet users to enhance their privacy awareness. The European General Data Protection Regulation (GDPR) is clearly envisioned to accomplish freedom, consent, justice, and security (The European Parliment and The Council of the EU - GDPR, 2016).

The principle of consent in the GDPR says that the agreement (written or oral, or electronic means) to the personal information processing relating to users should be explicit, informed and unambiguous. This could include ticking a box of the privacy policy when visiting a service provider. The privacy policy or similar statements should clearly mention information on user personal information handling activities agreed for the purposes. This principle has clearly revealed that there is a need to present a work for the analysis and visualization of the contents of privacy policies. Hence, considering this need from GDPR and literature review, it shows that the work discussed in this thesis is necessary.

1.4.1. RESEARCH QUESTION

The research question (RQ) of this research is:

> *"How to improve users' understanding of privacy policies while interacting with service providers in a variety of contexts?"*

Considering the fact that users do not read the privacy policies of service providers, it is important to understand the reasons of overlooking the privacy documents (McDonald & Cranor, 2008). Hence, the question is how to help the users to read the contents of privacy policies to enhance their privacy awareness.

1.4.2. RESEARCH OBJECTIVES

This research presents four research objectives (RO) to be achieved later in this thesis.

RO1: To understand Internet users' privacy awareness and current online practices.

In the first place, the first objective is to understand the most important privacy risks involved when the user communicates and access variety of online services. As this research revolves around privacy awareness, so it's vital to know the understanding of the users towards privacy. Considering privacy issues in India, it is essential to understand the user behavior in the different online scenarios. This will help to understand the requirements to design motivation and mitigation plan for personal information privacy awareness.

RO2: To interpret and analyze the privacy and data policy set by the service providers.

In order to enhance privacy awareness, it is important to know the role of the privacy policy of the service providers. The privacy policy involves the collection of users' information and its management. So, it is further important to know the user's personal attributes that are being shared with service providers. This will lead to another area where users personal attribute and its use mentioned in privacy policy need to be understood in detail.

The types of user attributes and its use are vital in understanding the goal of the policy of service providers. It is also important to know the purpose of information collection and the methods of the collection while interacting with the services offered by the service provider. How service providers ensure the user about information privacy when it comes to the agreement with the user on personal information collection, processing, and sharing, is the question to be answered here.

RO3: To improve users' understanding of privacy policy through visualization tool.

The interpretation and analysis of the privacy policies (RO2) lead to many findings. The next challenge is how to bring these findings to the notice of Internet user. Hence, the objective is to visualize the analyzed contents of privacy policies using visual aids. To what extent the contents of privacy policies are analyzed, visualized and summarized to enhance users' understanding of privacy policy is the core objective of this research.

Involving automated process for visualization will help users to improve the focus on privacy policy readability and understandability. The visualization tool will help users to gain deeper and clear knowledge of the privacy policy.

RO4: To develop a method to collect user review of services based on user experiences.

Users' knowledge will help to rate the services based on their experiences. So, the question is how to gather user experience based on the privacy policies. Hence, the next objective is to know the rating of the websites or services from the users who have experienced the visualization tool.

The experiences of users will help to know the rating to the visualization tool. This would help to justify the accuracy achieved by the tool. The feedback on the tool and its analysis help to define and describe a template for the privacy policies.

To establish a platform for analysis and visualization of the privacy policies, the above-mentioned objectives, introduces a plan to get a necessary action plan. RO1 helps to understand the privacy awareness and online practices of the users. This introduces the need for monitoring and mitigation factory for privacy protection and awareness. RO2 helps to focus on difficulties in reading the privacy policies set by the service providers. RO3 assist in defining a way/approach to simplify the contents of privacy policies. The final RO4 requires the resulting approach is accepted and acts as a base for enhancing privacy awareness.

1.4.3. SCOPE OF THE RESEARCH

Privacy policy analysis, visualization, and summarization by developing a privacy awareness tool are the key work presented in this thesis. The scope of this work is limited to address the challenges in reading privacy policies of the service providers.

Considering privacy issues in India (Dhotre & Olesen, 2015), the privacy policies were selected from various service providers that are frequently used in India.

The best way to demonstrate the work is to implement a tool that reached to the users in quickest and easiest way. This tool is implemented as a browser extension that can be easily installed manually. This tool will present the contents of privacy policy when the user visits the website and by clicking on the browser extension icon.

The combination of the manual and automatic process presented in this thesis is intended to enhance privacy awareness among users. The results produced from this awareness tool can spread the awareness of user information handling methods mentioned in the privacy policy of service providers. Also, contributing to the work carried out Web of Trust[9] is attained by integrating their work to the tool mentioned in this thesis later.

1.5. CONTRIBUTION OF THIS THESIS

This thesis describes a solution for the problem in understanding the privacy policy set by the service providers. The solution is represented in the easiest way so that a layman can understand it. However, this solution is stressing on privacy awareness and not privacy protection.

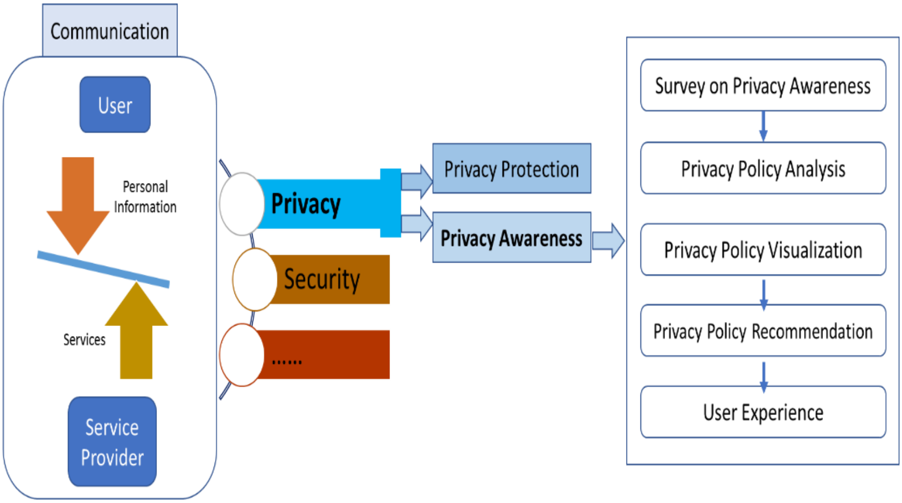


Figure 1-2 Problem evolution and objectives

The work represented in this thesis has five contributions (fig 1-2). The contributions start with the identification of privacy issues to the users whenever the user is involved

[9] Mywot.com

in online activities. This helps to understand the requirements for proposing a solution to some of the privacy issues. Fig. 1-2 illustrates the summary of problem evaluation to objectives.

- Why privacy awareness

  Privacy protection and privacy awareness are the two further areas to address privacy issues. It is important to empower the users by spreading the awareness about privacy. Privacy-aware users make the right decision while sharing the information on the web. Using some mechanism/tool, the privacy awareness can be disseminated among a set of Internet users. Through interactions, the users can empower other users. Therefore, the service providers either should stop pricing the personal information or the users should get the returns on the personal information sharing (Li, Li, Miklau, & Suciu, 2012).

Hence, privacy awareness will try to make the equal balance between users and service providers. Privacy awareness (a privacy innovation or privacy management) starts with a small group of users that helps to propagate at larger scale (Avgerou & Stamatiou, 2015)

The proposed work is more focused on privacy awareness than privacy protection. The privacy awareness can be realized by developing and implementing a tool that takes the requirements of privacy. The user survey helps to understand the requirements from the user for privacy awareness as well as protection. User consent, informed decisions are the key parameters to look further into solving problems of privacy awareness.

The privacy policy is the key element of communication between users and service providers. This is a text document that speaks on personal information harvesting. Visualization of privacy policy enables the user to understand privacy policy very well and can make the informed decisions.

On the other hand, present privacy policies are unstructured, lengthy and difficult to understand. Recommending standardized and common structure for privacy policy contents is the key work of this research. The recommendations are experienced by the users and it illustrates that the privacy policy visualization hopefully enhances privacy awareness.

1.5.1. CONTRIBUTION 1: SURVEY ON PRIVACY AWARENESS AND ONLINE PRACTICES

The initial part of this thesis is discussed as an attempt to identify real problems and understand the user's privacy knowledge, perception and their online practices using a survey. This survey can be used to perceive users' realization towards Personal

Information Privacy (PIP). The focus of the survey to find the factors that help the users to enhance privacy awareness. This study can assist to understand users' privacy concerns, and practices when they are online and interacting with services offered by service providers. Also, the outcome of this study is to know possible threats to user privacy.

This study acts as a base to understand the basic challenges for user empowerment like the visualization of information flow, user consent, controlled bargain, readability of privacy policy and its visualization, etc. The responses of the survey suggest the motivation and mitigation plan that is used to identify the requirements to design a solution that will help to enhance user privacy awareness. The survey gives more opportunities to enhance privacy awareness.

1.5.2. CONTRIBUTION 2: MANUAL ANALYSIS OF PRIVACY POLICIES

The privacy policy is an important document that allows companies/service providers to communicate with users on various aspects of personal data. This document describes harvesting personal information and its management. Preferably, the privacy policy should be readable and informative. So, this contribution is focused on knowing how descriptive, informative the privacy policy is? Interpretation and analysis of the privacy policy of online services are one of the major challenges.

The initial work of this contribution is to analyze user attributes collected by service providers. The contribution was developed with a classification of user attributes into the Personally Identifiable Information (PII) and other attributes. The graphical representation of user attributes will assist the users to understand its importance in the form of its ways of the collection, use, and sharing.

The manual analysis approach discussed in the thesis involves an interpretation and analysis of over 50 privacy policies of different service providers. This analysis provides the insights on how data are gathered, its purposes and data management. Also, the study is focused on knowing the readability of privacy policy, security measures and adherence of privacy policy contents to the privacy laws or regulations. The benefit of this study is to realize that there exists a gap between the service providers' assurance and the users' expectations. The study denotes that the privacy policy needs to be clear, shorten and should offer information using visual cues. The outcome of this analysis reveals the important and user concerned sections of the privacy policies.

1.5.3. CONTRIBUTION 3: DEVELOPMENT OF VISUALIZATION TOOL TO IMPROVE USERS UNDERSTANDING ON PRIVACY POLICY

This contribution is an extension to the previous contribution. An automated tool called PPET is developed to classify the contents of privacy policy among key sections. This tool uses a Naïve Bayes classification algorithm that helps the user to read classified contents of the privacy policy. To develop this tool, a corpus was necessary which helped in training the model. The large corpus of was developed that has more than 43544 records. Hence, this tool allows the user to quickly select a section of the privacy policy to read and understand.

The visualization and summarization of privacy policy contents can help the users to easily understand various approaches used by service providers on users' information management. This contribution is the answer to know how service providers handle users' personal information, a way of collection, security practices, and so on. The different panels provided in this tool will motivate to read the privacy policy. More than 600 privacy policies of frequently used websites in India are analyzed.

Also, the work presented here is a contribution to spreading the reputation of the website from WoT. The rating shown on this tool surely will give self-confidence to the user for trustworthy interaction with the service provider. So, hopefully, this unique tool enhances user awareness.

1.5.4. CONTRIBUTION 4: RECOMMENDATION- A TEMPLATE FOR PRIVACY POLICIES

The major challenge for service providers/companies is to quickly provide the right information and its purpose for the user. There should not be any loss to the service provider or the user while collecting, disclosing, and sharing user information. Hence, the service provider should follow the right steps to manage user information and must adhere the rules and regulations given by like country laws like Information Technology Act 20018 (IT Act) (The Gazette of India, 2009).

On the same line, the European GDPR acts an opportunity to build a faithful and reliable relationship between service providers and users. User's explicit consent is the major focus during user information handling. The GDPR accentuate on the principles of Privacy by Design (PbD) (Cavoukian & Jonas, 2012) where privacy right should be considered from service inception. As per the rules, there should be a clear record specifying user information processing, a list of beneficiaries, information retention time, etc.

So, this contribution can help the users to know the service providers' current practices and adherence to the rules and regulations. This research work presented in this thesis identifies that the service provider's strategy on user information management is not similar to the guidelines mentioned by regulatory authorities.

Hence, the result of this contribution is a recommendation of standard template for the privacy policies in India. Considering users' responses on a survey and analysis of privacy policies, the privacy policy should be structured, simple, short and should be easy to understand. The standard template encompasses important privacy aspects like user information collection, a way of collection, security measures, cookies, and other key sections.

1.5.5. CONTRIBUTION 5: DESIGNING INTERFACE FOR “RATING TO WEBSITE”

The work presented in this thesis is privacy awareness tool to address the readability issue of the privacy policy. The benefits of developed tool do not limit to the privacy policy, but also ratings of the website. This tool also acts as a platform to receive user feedback based on their experiences with the services provided by the websites. Users feedback and comments received using the interface of the tool. The acts a base to give a rating to the websites. The results of the proposed approach show that the PPET supports the users in understanding the privacy policy easily.

1.6. RESEARCH METHODOLOGY

This thesis has performed a quantitative survey (Kothari, 2004) by employing a questionnaire to understand Internet users’ privacy awareness and current online practices to explore the understanding of the users towards privacy, the user behavior in the different online scenarios, requirements to design motivation and mitigation plan. Randomly selected Indian users were invited to respond online quantitative survey. Responses were analyzed to understand the distribution of attributes of respondents as age, computer literacy, etc., to recognize motivation and mitigation factors to increase users’ privacy awareness.

To interpret and analyze the privacy and data policy set by the service providers, manual content analysis[10], (Krippendorff, 2004). of popular service providers in India, is performed. Privacy policies were downloaded and manually analyzed to understand various aspects including collected user-data, a method of data collection, comment on children policy, opted security measures, etc. The content analysis methodology helps to understand the meaning and relationship of privacy terms, sentences, etc.

To improve users’ understanding of privacy policy, a visualization tool-PPET- is implemented. Statements of the privacy policies of more than 600 service providers are classified into different sections. The classification is done by using the Naïve Bayes classifier. PPET represents privacy policies in easily understandable visual representation to enhance users’ privacy awareness. The feedback of 262 responses on PPET tool was analyzed and found that the tool is useful to users in reading and understanding the privacy policy. The PPET has also asked the users to give ratings of the services based on their experiences.

[10] http://www.worldcat.org/title/research-methodology-methods-techniques/oclc/395725716

1.7. PUBLICATIONS

A book chapter and the peer-reviewed publications were presented as a part of the publication. The details are as follows:

A. Book Chapter:

1. **Prashant S. Dhotre**, Anurag Bihani, Samant Khajuria, Henning Olesen**,** (2017), **"Take it or Leave it": Effective Visualization of Privacy Policies"**, Samant Khajuria, Lene Sørensen, Knud Erik Skouby, "***Cybersecurity and Privacy: Bridging the Gap"***, Denmark, River Publication, ISBN: 9788793519664

B. Conference Publications:

1. **Prashant S. Dhotre**, Henning Olesen**, "A Ph.D. Abstract presentation on Personal Information Privacy System based on Proactive Design"**, IEEE INDICON Yashada, Pune, India, December 11, 2014.

2. **Prashant S. Dhotre**, Henning Olesen, **"A Survey of Privacy Awareness and Current Online Practices of Indian Users: Motivating and Mitigating Factors for Improving Personal Information Privacy"**, Proceedings of WWRF Meeting 34, Santa Clara, CA, USA, Apr. 2015.

3. **Prashant S. Dhotre**, Henning Olesen, Samant Khajuria**, "Interpretation and Analysis of Privacy Policies of Websites in India"**, Proceedings of WWRF Meeting 36, Beijing, China, June 2016.

4. **Prashant S. Dhotre**, Henning Olesen, Samant Khajuria, **"User Privacy and Empowerment: Trends, Challenges, and Opportunities"**, *International Conference on Intelligent Computing and Communication ICICC - 2017.Springer Series on* MAEER's MIT College of Engineering, Pune, India. Conference Date: August 2017.

1.8. THESIS ORGANIZATION

Fig. 1-3 illustrates the outline of the thesis with a brief introduction to each chapter. Considering the evaluation of state-of-art presented in chapter 2, the main requirements to address the privacy awareness using a visualization tool. The development of the tool involves several challenges. The main challenge is to know the privacy knowledge and awareness among users.

The other challenge is to inspire and motivate the users to read the privacy policies as it is complicated and difficult to read. The challenges were addressed and discussed in several chapters.

Chapter 2: State of the art and research context

This chapter introduces the concept of privacy, privacy and user information, and privacy in the current situation. Further, this chapter discusses on privacy threats followed by privacy definitions.

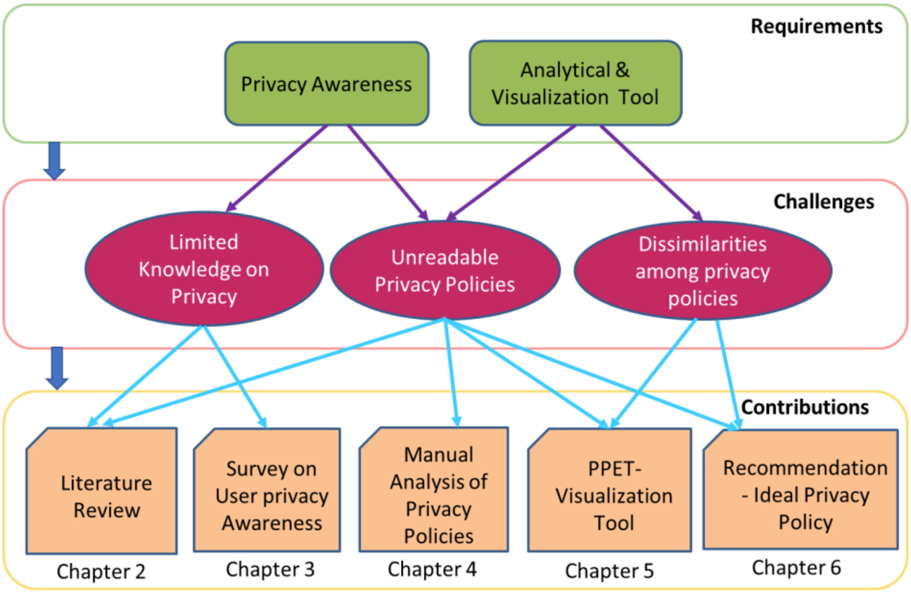

Requirements
Privacy Awareness
Analytical & Visualization Tool
Challenges
Limited Knowledge on Privacy
Unreadable Privacy Policies
Dissimilarities among privacy policies
Contributions
Literature Review
Chapter 2
Survey on User privacy Awareness
Chapter 3
Manual Analysis of Privacy Policies
Chapter 4
PPET-Visualization Tool
Chapter 5
Recommendation - Ideal Privacy Policy
Chapter 6

Figure 1-3 Thesis organization

Considering the definition of privacy, specific privacy awareness issues are presented in this chapter. Privacy protection principles, regulations, and guidelines are also discussed. This is continued with privacy-enhancing technologies and description of privacy awareness survey and visualization tools. This chapter summarizes the issues in the privacy policy of the service providers**.**

Chapter 3: Survey on privacy awareness and online practices

This chapter discusses the survey on user privacy carried in India. The principles of the survey are discussed, followed by the methods of a survey conducted. This is continued with the questionnaires and diversification of participants. Further, the results of the survey are elaborated in detail. The outcome of the survey is concluded at the end of this chapter.

Chapter 4: Extensive analysis of privacy policies

Manual analysis of privacy policies is described in this chapter. The chapter begins with the motivation for performing manual analysis of privacy policies. Further, it describes how the websites are selected. The interesting way of manual analysis method is discussed to gain important finding from privacy policies. Further, the extensive analysis, results, and discussions are presented. The chapter concludes with challenges in understanding privacy policies.

Chapter 5: Visualization of privacy policies using Privacy Policy Elucidator Tool (PPET)

This chapter begins with the description of the motivation for visualization of privacy policies to enhance privacy awareness. Further, introduction to visualization approaches is described. The need and the goal of the proposed visualization tool- the PPET- is presented. To continue, this chapter includes the big size corpus, architecture, mathematical modeling, machine learning based model. The client and server-side algorithms are also given in detail. The tool is applied to more than 600 privacy policies. The detailed discussion about the results obtained from the PPET tool is given in the last part of this chapter.

Chapter 6: PPET impact and recommendations

This chapter begins with a feedback on the PPET. The feedback is necessary to know the several things about privacy policy from the users' point of view. The analysis of feedback given important things like ratings of the services, important sections of privacy policies, the usefulness of the PPET, etc.

This chapter also discusses the privacy principles and guidelines, privacy policies, and to what extent the privacy policies follow privacy principles and guidelines. In the end, this chapter recommends a uniform and standard template for the privacy policies.

Chapter 7: Conclusions and future work

The potential use of research work is summarized in this chapter. This chapter concludes the thesis. Here, concluding remarks on the whole contribution as well as results obtained using PPET tool is specified. Also, future work is discussed in this chapter that talks about analyzing semantic of privacy policies, risk assessment, bargaining mechanism, pay for privacy, etc.

1.9. LIMITATIONS OF THE RESEARCH

The work presented in this thesis is a development of privacy awareness tool (Marella et al., 2014) to address the problem of “how to understand the privacy policy in the easiest way”. The systematic analysis and visualization of privacy policy involve manual as well as the automated process. This semi-automated process is implemented to get important findings from privacy policy along with its readability. Involving the user in reading and understanding a section or the complete privacy policy is the motivation towards developing an automated tool described later in this thesis.

This research offers only an approach to analyze, visualize and summarize the contents of the privacy policy. The work proposed in this research is based on the contents of the privacy policies. This work doesn’t take any semantic value into the consideration as this is a part of future work.

This tool has implemented and tested over 600 privacy policy of the several websites.

Privacy and privacy policy are important aspects to building a strong relationship between the users and service providers. This work is not defining privacy in any context. This work is related to the privacy policy to enhance privacy awareness and not the privacy protection that will take research area towards user information security. Hence, this research doesn’t involve any security techniques or measures.

1.10. REFERENCES

A structure for privacy - Teradata Magazine. (2007). Retrieved June 12, 2015, from http://apps.teradata.com/tdmo/v07n02/Features/StructureForPrivacy.aspx

Avgerou, A. D., & Stamatiou, Y. C. (2015). Privacy awareness diffusion in social networks. *IEEE Security and Privacy*, *13*(6), 44–50. http://doi.org/10.1109/MSP.2015.136

Beye, M., Jeckmans, A., Erkin, Z., Erkin, Z., Hartel, P. H., Lagendijk, R., & Tang, Q. (2010). Literature Overview - Privacy in Online Social Networks. Centre for Telematics and Information Technology University of Twente. Retrieved from https://research.utwente.nl/en/publications/literature-overview-privacy-in-online-social-networks

Castells, M. (2009). Communication Power. *3 Great*. Retrieved from http://socium.ge/downloads/komunikaciisteoria/eng/comunication power castells.pdf

Cavoukian, A., & Reed, D. (2013). Big Privacy: Bridging Big Data and the Personal

Data Ecosystem Through Privacy by Design. Retrieved June 28, 2014, from https://www.ipc.on.ca/images/Resources/pbd-big_privacy.pdf

Cavoukian, A. (2008). Privacy in the clouds. *Identity in the Information Society*, *1*(1), 89–108. http://doi.org/10.1007/s12394-008-0005-z

Cavoukian, A., & Jonas, J. (2012). Privacy by Design in the Age of Big Data. Retrieved June 28, 2014, from https://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf

Cochran, T. (2013). Personal Information is the Currency of the 21st Century. Retrieved August 10, 2016, from http://allthingsd.com/20130507/personal-information-is-the-currency-of-the-21st-century/

David Lawler. (2016). Millions of email accounts compromised in massive data breach that includes Google and Yahoo. Retrieved March 1, 2017, from http://www.telegraph.co.uk/news/2016/05/04/millions-of-email-accounts-compromised--in-massive-data-breach-t/

Dhotre, P., & Olesen, H. (2015). A Survey of Privacy Awareness and Current Online Practices of Indian Users: Motivating and Mitigating Factors for Improving Personal Information Privacy. *WWRF*. Retrieved from http://vbn.aau.dk/files/218591293/A_survey_of_privary_awareness_and_current_practices_of_Indian_online_users.pdf

Dhotre, P., Olesen, H., & Khajuria, S. (2017). User Privacy and Empowerment : Trends , Challenges , and Opportunities. In *Springer*.

FTC. (2012). Sharing Information: A Day in Your Life | Consumer Information. Retrieved May 18, 2015, from http://www.consumer.ftc.gov/media/video-0022-sharing-information-day-your-life

Khajuria, S., Sørensen, L., & Skouby, K. E. (2017). *Cybersecurity and Privacy - Bridging the Gap*. River Publishers. Retrieved from http://www.riverpublishers.com/book_details.php?book_id=434

Kothari, C. . (2004). *Research Methodology: Methods and Techniques*. Delhi: New Age International Publishers, New Delhi.

Krippendorff, K. (2004). *Content analysis : an introduction to its methodology*. Retrieved from https://books.google.co.in/books/about/Content_Analysis.html?id=q657o3M3C8cC&redir_esc=y

Kumar, H., Jain, S., & Srivastava, R. (2016). Risk analysis of online social networks. In *2016 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 846–851). IEEE. http://doi.org/10.1109/CCAA.2016.7813833

Kuneva, M. (2009). Keynote Speech on Roundtable on Online Data Collection, Targeting and Profiling. *European Consumer Commissioner*, (March).

Lee, R., & Janna, A. (2014). The Future of Privacy | Pew Research Center. Retrieved February 27, 2017, from http://www.pewinternet.org/2014/12/18/future-of-privacy/

Li, C., Li, D. Y., Miklau, G., & Suciu, D. (2012). A Theory of Pricing Private Data. http://doi.org/10.1145/2448496.2448502

Lou, W. L. W., & Ren, K. R. K. (2009). Security, privacy, and accountability in wireless access networks. *IEEE Wireless Communications*, *16*(August), 80–87. http://doi.org/10.1109/MWC.2009.5281259

Marella, A., Pan, C., Hu, Z., Schaub, F., Ur, B., & Cranor, L. F. (2014). Assessing Privacy Awareness from Browser Plugins.

Mayer, J. R., & Mitchell, J. C. (2012). Third-Party Web Tracking: Policy and Technology. In *2012 IEEE Symposium on Security and Privacy* (pp. 413–427). IEEE. http://doi.org/10.1109/SP.2012.47

McDonald, A. M., & Cranor, L. F. (2008). The Cost of Reading Privacy Policies. *I/S - A Journal of Law and Policy for the Information Society*, *4*(3), 1–22.

Mowbray, M., & Pearson, S. (2012). Protecting Personal Information in Cloud Computing (pp. 475–491). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-33615-7_3

Natasha Lomas. (2016). WhatsApp's privacy U-turn on sharing data with Facebook draws more heat in Europe | TechCrunch. Retrieved February 28, 2017, from https://techcrunch.com/2016/09/30/whatsapps-privacy-u-turn-on-sharing-data-with-facebook-draws-more-heat-in-europe/

Nathaniel Mott. (2017). As WhatsApp becomes latest victim, are any messaging apps truly secure? | Technology | The Guardian. Retrieved March 1, 2017, from https://www.theguardian.com/technology/2017/jan/14/whatsapp-vulnerability-secure-messaging-apps

OECD. (2013a). Exploring the Economics of Personal Data. *OECD Digital Economy*

Papers, (220), 40. http://doi.org/10.1787/5k486qtxldmq-en

OECD. (2013b). OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data - OECD. Retrieved September 23, 2017, from http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

Ortlieb, M. (2014). The Anthropologist ' s View on Privacy. *IEEE Security & Privacy*, (June), 85–87.

PrivacyPolicies.com. (2015). Privacy Law &amp; Regulations by Country. Retrieved February 27, 2017, from http://privacypolicies.com/blog/privacy-law-by-country/

Rule, J. B., & Greenleaf, G. W. (Graham W. (2010). *Global privacy protection : the first generation*. Edward Elgar.

Salvatori, L. (2015). Social Commerce: A Literature Review. In *Science and Information Conference 2015 July 28-30, 2015 | London, UK.*

Steve Kroft. (2014). The Data Brokers: Selling your personal information - CBS News. Retrieved March 1, 2017, from http://www.cbsnews.com/news/data-brokers-selling-personal-information-60-minutes/

The European Parliment and The Council of the EU - GDPR. (2016). GDPR. *Official Journal of the European Union*. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

The Gazette of India. (2009). *Information Technology Act 2008,*. Retrieved from http://meity.gov.in/sites/upload_files/dit/files/downloads/itact2000/it_amendment_act2008.pdf

Vishal Singh. (2016). Delhi High Court Restrains Snapdeal Owned Firm Unicommerce From Using Paytm's Biz Data - Inc42 Media. Retrieved March 1, 2017, from https://inc42.com/flash-feed/delhi-high-court-restrains-unicommerce/?utm_source=inshorts&utm_medium=inshorts_full_article&utm_campaign=inshorts_full_article

Zeadally, S., & Winkler, S. (2016). Privacy Policy Analysis of Popular Web Platforms. *IEEE TECHNOLOGY AND SOCIETY MAGAZINE*, (june), 75–85.

Zengjie Cao, Yuanyuan Lin, & Cong Zhao. (2011). An empirical study of user's attitudes and behavior of privacy concerns. In *2011 International Conference on Computer Science and Service System (CSSS)* (pp. 2029–2034). IEEE.

http://doi.org/10.1109/CSSS.2011.5975032

CHAPTER 2. STATE OF THE ART AND RESEARCH CONTEXT

“Privacy, like an elephant, is... more readily recognized than described” (Young, 1978). Considering the varied contexts that the term privacy is used (context, social, legal, technological, etc.), it is tough to define privacy. Therefore, even more, it becomes necessary to examine the concept of privacy, its meaning, and different issues. In this chapter, the actual issues and threats are discussed which act as a base for the problem formation with this thesis as an answer. The chapter begins with privacy from different angles like user understanding, present digital era, etc. Personal data privacy protection and awareness are discussed by looking at the knowledge extracted from existing research, solutions. Also, the privacy protection and awareness tools are discussed along with their extensive comparative analysis. The chapter concludes with a discussion on the need for effective privacy awareness tools.

The few contents of this chapter have been already published in a Book Chapter (Paper no: Section 1.7, A-1). The book chapter gives a quick review of a few privacy awareness tool. However, this chapter provides the detailed description of most of the privacy awareness tools and tracking/blocking tools. The survey included in this chapter is extensive as compared to the publications presented in papers (Section 1.7, B-2, B-3 and B-4).

2.1. PRIVACY, CULTURE, AND UNDERSTANDING

Historically speaking, it is important to know and understand what is privacy, information privacy, and its meaning.The concept of privacy is a social concept (Lou & Ren, 2009),(Neill, 2001). Probably, privacy was first understood and published as, “the right to be let alone“ (Warren, S. and Brandeis, 2008). Privacy is a human right that acts as a prerequisite for protecting human beings from others and other entities with self-esteem and self-respect (Bruce, 2006). It is to be noted that main thought behind privacy is to offer liberty to the individual so that they can determine what they want, when and how. All stakeholders in society may not be in universal agreement to this idea of liberty.

2.1.1. PRIVACY IS NOT NATURAL

Humans invented devices to overcome the limitations imposed by their senses to know, understand and control the world. Human senses are limited to understand other humans and their abilities like feeling towards others. Considering our feelings about other devices, humans, or the environment, we create our own theory of understanding. Directly or indirectly, this theory is devised by human or group of humans. This theory involves a perception towards protection of ourselves from others (Warren, S. and Brandeis, 2008). Human beings have an innate need to understand and interact with the environment which in today's world includes not only other human beings but also devices which increase the scope of this interaction. Simultaneously, this need also brings along with it another risk - the risk to privacy. Over the years we have regularly devised theories for privacy protection to form a safety net around which this communication takes place.

The meaning and definition of privacy differ. Since 1890 and especially in today's hyperconnected world, privacy and privacy issues have been in the spotlight by researchers, governments, and companies.

Even though the privacy is a 'fundamental right' ("EU Directive-GDPR," 2016), the differing interpretations on privacy among nations lead to the formation of different laws and regulations around the world. In many countries, privacy regulations specify the rights to personal data protection and control. Charter of Fundamental Rights of the European Union prominently states that protection of user privacy is a fundamental right. Such personal data should be treated ethically as per the specified purpose ("EU Directive-GDPR," 2016). Likewise, in the US, privacy is defined as basic rights of people (United Nations, n.d.).

Despite regulations in the US and Europe, there are major differences in the user understanding of privacy and the regulations. However, in some countries, privacy is either taken for granted (Rule & Greenleaf, 2010) and in others, it is inadequately implemented(Ryan, Merchant, & Falvey, 2011a). In many cultures, people prefer to seclude themselves from strangers and selectively share information. Mutually respecting each other's privacy is expected.

The notion of privacy is constantly evolving due to technological advances. Hence, there is no singular, universally accepted concept of privacy.

2.1.2. PRIVACY IN DIGITAL WORLD

Before the digitization of the society, the human's private premises or home was the world for him/her. Now, the digitization has brought about a rapid change in all situations (Lee & Janna, 2014).

Since people engage in social, cultural, commercial and financial life through the digital sphere, it is very difficult to define the boundaries of an individual's privacy. Today in the world of Big Data and Internet of Things (IoT), individuals are creating their own preferred space (online shopping, political preferences, Internet browsing, listening to own playlist, etc.) using the modern and digital equipment. Moreover, there is online monitoring and tracking of online activities, irrespective of a user’s geographical location.

In physical world privacy and the data is under the absolute control of the individual. Whereas in the digital world, the external interface allows data transfer between and among domains So, it’s challenging to protect user’s information which is easily accessible without any means of obstacles.

For example, the users’ data, who agree to the terms of service and privacy policy, are prone to misuse, abuse, and manipulation if they disclose more information than what is necessary. The ignorance about data management practices adopted by service providers once the terms of service and privacy policy are agreed upon could lead to potential exploitation of private data once shared the information in the digital domain is beyond the control of the users, and can be manipulated, displayed and mishandled with or without the owner’s consent.

The important thing is that the user information is in digital format and stored in the memory. Using Big Data technology (Cavoukian & Jonas, 2012), personal information is being processed, hidden patterns identified to extract a value from user data. This will lead to the issues of privacy and security.

The information stored by service providers is accessible to third parties and may be sold to data brokers. The data brokers make money by selling users' information to advertisers and publishers. Unknown to users, the biggest risk to the users’ personal information is from data brokers (Economist, n.d.).

Once the data is gathered at the service provider’s end, the information is categorized according to various segments (marital status, unemployed, political preferences, medical information, etc.). These segments are ready for sale to advertisers, government sectors or needy organizations.

2.1.3. PRIVACY PERCEPTION – IRREGULARITY

The understanding of privacy needs to be revised considering the amount of information collected and processed. This will enforce change in the existing laws to address contemporary needs.

The information collected by the surrounding devices or service providers is stored outside the boundaries of users’ control. Hence, there is an issue of ownership. The user should have more power over the raw data as well as processed data. Also, the information generated from the user action should belong to the users.

While defining personal information, it is stated that there is always a closed relationship between the event and the user (EU Agency for Fundamental Rights, 2014). Hence, the law is indecisive about the identity of individuals based on the information retrieved after data mining. It is necessary to consider various views on the privacy definition.

2.2. PERSONAL INFORMATION PRIVACY- DEFINITIONS

Privacy is considered in many ways. Interestingly and surprisingly, there are many views and definitions of privacy defined over the years. there are popular definitions by Westin in 1967 (Alan Westin, Osborne M., & Jr., 1967). One definition by Mike Bergmann (Bergmann, 2009) is:

> *"Privacy… means the right of self-determination regarding data disclosure, i.e., each user should be able to control how much personal information he is willing to give to whom and for what purpose."*

In 1975, a hypothesis from Atman about privacy states that it is a spectrum of two context based things i.e. "Openness "and "Closeness". The people will decide their spectrum in the process of optimizing their accessibility (Altman, 1975). According to Dourish and Anderson, the social boundaries are defined by the information flow (Dourish & Anderson, 2006). Information flow acts as an indicator to negotiate and strengthen identity in social groups. Achieving the desired state is the main goal of privacy along with Openness and Closeness. Hence, privacy is not only avoiding disclosure of information but also a context-based selective disclosure of information.

Another view mentioned is that privacy is a collection of four important domains: Security, Anonymity, Confidentiality, Safety (Abdullah, Conti, & Beyah, 2008). Security acts as a blocker to unauthorized parties from accessing the information. Anonymity helps to isolate the identity from information. Limiting access or disclosure of certain information is achieved using rules called confidentiality. Safety helps to protect user information from unwanted consequences.

An important aspect of personal information privacy is the term personal data or PII. The new Directive General Data Protection Regulation (GDPR) is quite complete and will have a major influence on services and business processes ("EU Directive-GDPR," 2016).The idea is to preserve personal information protected and lay emphasis on explicit user consent for using personal information. As per GDPR (article 4 (1)), personal data is defined as:

> *Any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more*

> *factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*

Here, the focus is given more on data identifier, not on data as such. The examples of the identifier can be name, identification, and location of data, address/email address, etc. There is another important definition of PII given by NIST (National Institute of Standards and Technology) (NIST, 2010). According to NIST, PII is defined as

> *"any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information"*

PII is used in US law, but it is used in EU regulation and defined as personal data (EU Agency for Fundamental Rights, 2014) and is given as:

> *"Data are personal data if they relate to an individual or at least identifiable person, the data subject"*

Even though the definitions are defined in different terms, however, the objective is the same. In the study by Daniel Solove (Solove, n.d.) the privacy is defined as:

> *"The term "privacy" is best used as a shorthand umbrella term for a related web of things. Beyond this kind of a use, the term "privacy" has little purpose. In fact, it can obfuscate more than clarity"*

Along with the above definition, the study showed that people in the USA understand privacy through security. However, the other angle to see privacy is provided by the privacy design goals and can be expressed as follows (Cavoukian, 2008):

> *"The goal of flexible, user-centric identity management infrastructure must be to allow the user to quickly determine what information will be revealed to, which parties and for what purposes, how trustworthy those parties are and how they will handle the information and what the consequences of sharing their information will be"*

In information privacy, identifiability is one of the key issues. In short, while designing the tools, emphasis should be on minimum disclosure, for the intended purpose. Also, users should be able to give informed consent.

With reference to the definition of privacy by (Bergmann, 2009) and in the context of privacy policy, the privacy awareness is defined as:

> *"Privacy Awareness… is seen as the user's ability to reflect the communication partner's privacy policy statements regarding purpose binding, transfer assertion and retention period applied for a certain data disclosure."*

This definition makes an obligation to the service provider to make a meaningful and complete privacy policy so that users can be aware of "what happens to their personal information".

According to the concept of "Protection motivation theory" (Rogers, 1975) (Conner & Norman, 2005), an Internet user is more motivated to follow protective measures if the user is properly educated about various parameters related to privacy issues and threats. This awareness includes possible threats, the impact of possible threats, the probability of occurrence of threat, protective approaches, and the efficiency of protective approaches. Therefore, there is a need to propose an innovative approach or tool to avoid potential loss due to limited awareness. This could be achieved by identifying the user attributes shared online, educating and notifying privacy threats and preventive measures to users, measuring user privacy risk level and improve users' perception of privacy risk (quantifying probability or vulnerabilities of occurrence of threats), etc.

Nuno Fortes et al. (Fortes, 2016) evaluated a study for internet users whose average age was 24 years and concluded that most of the users are not aware of privacy and are lazy to adopt approaches for privacy preservation. Educating users about privacy risk has resulted in more willingness of users to use privacy preservation approaches and a decrease in willingness to share some personal information.

2.3. PRIVACY AND ACTUAL THREATS

It has been found that service providers, especially social networking companies are giving adequate privacy control to their users. The users face privacy risks because the default privacy setting is not only not providing security, but also creating confusion. (Bonneau & Preibusch, 2010)(Alsagri & Alaboodi, 2015) (Karahasanovic et al., 2009).

Mobile and wireless communication have suffused our lives with ease. Through innovative technologies, it is possible to closely and directly monitor an individual's life for improving the quality of life, especially in the domain of healthcare. New advancements in technology can monitor heart rates, the sugar level in blood, and blood pressure[11].

Such technologies and instruments help to reduce costs, save time and enhance the precision levels of diagnosing which direct human intervention cannot. Each data set is stored on the service providers side and analyzed for providing services to the patients. Significantly, the colossal amount of data collected have become an asset for not only service providers but also other entities of the ecosystem which can be further used to mitigate problems and enhance human lives.

Information collection, monitoring, processing, and sharing are not limited to specific functions but have spanned new avenues which the user is unaware of. User profiling is another issue where capturing users' activities is required. Identification of a user can be served by examining information from several data sources like social media (Levchenko et al., n.d.). Users are attracted to publish information about themselves publicly which can be used for understanding users' expectations, likes/dislikes. This information later becomes user profiles that can potentially be used to cause harm to the users. Also, suitable tools are used to monitor data the game generates to know psychological traits of a player.

Users' detailed information is captured (browser information, IP address, cookies, browsing behavior) and such information is combined to define an information package of Internet users. Such information becomes a real problem for users if the information is shared with third parties without users' permission. This is a big problem.

Another prominent issue is the loss of user control. Once the data is shared, the user cannot exercise power over his personal information. Hence monitoring of users' activity with limited user control leads to escalating anxiety for the user.

2.3.1. CASE STUDY: AADHAR CARD IN INDIA

In India, it has now become mandatory to provide photographs and physiological information (especially fingerprints and iris) for beneficiaries of government schemes, compensation, benefits, student grants, etc. Recently, the users' information will be used in the development of policing (Jyoti Panday, 2017). Crime and Criminal Tracking Network & Systems (CCTNS)[12] is a plan under digital India that uses users' information for locating criminals and track them/missing people.

[11] https://www.theguardian.com/society/2015/may/19/digital-fitness-technology-data-heath-medicine

[12] http://ncrb.gov.in/BureauDivisions/cctnsnew/index.html

Indian citizen's biometric information is collected by Unique Identification Authority of India (UIDAI) and an AADHAR number, which is a 12-digit unique identity number, is issued. Until June 2017, 1.154 billion members have enrolled which forms the largest ID system in the world[13]. So, the huge amount of detailed user information is very vital for every enrolled member. Government officials and agencies have the access to AADHAR details whenever required. However, it is found that this information is not being handled very well considering privacy and security issues. Due to inefficient security measures, criminals and hackers take the opportunity of data leaks. Over 130 million AADHAR and bank details have leaked from websites of government departments[14].

So, in the context of privacy; data leak and misuse are a source of threat and worry for every individual. Even though the right to privacy has been denoted as a fundamental right, in reality, the sheer number of AADHAR cards as well as lack of strict digital security measures make the security of information a gigantic task. Similarly, the concept of the right to policy is considered as implicit rather than explicit. Considering these issues, it is debatable to make it compulsory for every individual to link AADHAR details to every scheme/bank transaction.

Concerns have been raised by opposing voices in the public domain on the government departments' insistence to link AADHAR with PAN. The critics argue that it could lead to a constant surveillance of the citizens by the state making them feel like "slaves"[15]. Not only the government but also, private companies can access the details of the individual using their AADHAR number. Some of the concerns identified that affect the privacy of users include user identity without consent, tracking and profiling through illegal access, etc.[16]

So, in short, it's a huge risk for individuals' privacy as the data leaks not only include personal or financial information, but also biometric information which is difficult to recover. Considering the guidelines and proposed act (Govt of India, n.d.), it is unclear about the notification and action against fraudulent entities or cases in terms of financial frauds or individuals whose data has been compromised for personal benefits. Moreover, the guidelines fail to cover the steps against illegal use of AADHAR for authentication, profiling or to be used as an asset for doing business for private companies. Hence, along with standard regulation formulation, a non-profit

[13] https://portal.uidai.gov.in/uidwebportal/dashboard.do

[14] http://indiatoday.intoday.in/technology/story/aadhaar-data-of-130-millions-bank-account-details-leaked-from-govt-websites-report/1/943632.htm

[15] http://timesofindia.indiatimes.com/india/mandatory-aadhaar-will-make-us-slaves-pil-in-supreme-court/articleshow/58407741.cms

[16] http://www.cse.iitd.ernet.in/~suban/reports/aadhaar.pdf

foundation or effective privacy protection/awareness mechanisms is missing in the countries like India, that not only protect user information privacy but also enhance user awareness about privacy.

2.3.2. CASE STUDY: NETFLIX

This online movie streaming service allows subscribed users to watch movies, TV shows, etc. The important thing about the Netflix is that the recommendations for movies or other watchable contents are based on users' ratings and watching history. Based on a significant and improvised algorithm (an outcome of a contest by Netflix-) that has improved recommendation which has turned Netflix to increase its subscriptions manifold[17],[18].

The consequences of the recommendation systems of Netflix have raised users' privacy issues. The ratings and users' information was easily accessed using the de-anonymization technique (Narayanan & Shmatikov, n.d.). According to this algorithm, the information is cross-related to the published non-anonymous ratings on the websites like the Internet Movie Database (IMD).

Overall, the mathematical model of de-anonymization requires little background information of users or secondary knowledge about customers to identify users and offer them required recommended movies or TV shows.

A lawsuit has been filed against Netflix for their failure in the protection of user privacy[19].

2.3.3. CASE STUDY: UNICOMMERCE VS PAYTM

Considering the recent advancements in online transactions and e-shopping in India, there have been cases of malpractices observed from famous online retailers which have an enormous number of customers[20].

Paytm (Payment through mobile) is an Indian e-commerce and electronic payment company founded in 2010 run by One97 Communication[21] and recently Paytm is

[17] http://www.nytimes.com/2009/09/22/technology/internet/22netflix.html?pagewanted=all&mcubz=3

[18] http://gigaom.com/video/google-schmidt-tv

[19] https://www.wired.com/2009/12/netflix-privacy-lawsuit/

[20] http://economictimes.indiatimes.com/industry/services/retail/snapdeal-owned-unicommerce-stole-business-data-paytm/articleshow/51921248.cms

[21] https://paytm.com/about-us

supported by Alibaba (a Chinese company). This company has 220 million registered users in the short span since its inception. Paytm offers Paytm Wallet used for many purposes, including online shopping, paying bills, etc.

Snapdeal is an online shopping site started in 2010 having a variety of the products for users, especially in India[22] [42]. It offers users some daily deals on the sets of products. Considering the number of daily transit users, the company aims to achieve to handle 20 million users in a day which is more than Flipkart and Amazon in India 2020[23].

As we see, both companies are dealing with a high number of users' transactions and in turn, have generated massive amounts of information out of those transactions. This could be one of the reasons for stealing business data as revealed recently.

A complaint was filed by Paytm against Unicommerce (Owned by Snapdeal; provides solutions to e-commerce vendors for inventory management) accusing Unicommerce for retrieving and storing confidential business data (including users' private information). The allegation also includes the use of Paytm name and logo without consent. The judgment of the Court in this matter it compulsory for Unicommerce to refrain from accessing any data that has been gathered by Paytm registered sellers. The verdict also asked the company to prevent the use of the name and logo for the advertisement.

2.3.4. LEARNINGS FROM CASE STUDIES

Looking at the three cases mentioned above, the main question arises is: How will modern technologies consider the privacy by design that helps to protect users' privacy?

In an ideal scenario, the overall ecosystem should be responsible for protecting users' information. But, in all cases, users' information is being shared, leaked, and re-used. Therefore, we can conclude that the data owners (users) will face the consequences of their threat to privacy and security.

So, this leads to the next point of discussion on the roles and responsibilities of all the elements in the ecosystem. The questions raised are: Is modern technology faulty? Have the business models overlooked users' data? or is awareness the answer to the

[22] https://www.snapdeal.com/

[23] http://economictimes.indiatimes.com/small-biz/startups/our-daily-transacting-users-exceed-flipkart-amazon-put-together-snapdeal/articleshow/53392532.cms

issues of privacy loss? Hence, there is a need to consider privacy as design or default and develop a new way to build a protective mechanism.

To start with, users' awareness is vital to understand the new requirements for designing modern technology/tools for privacy protection or spreading privacy awareness especially in countries like India where loss of privacy has become the main concern for users.

Another thing is to pass new privacy law and enforce compliance by the service providers or companies so that companies can provide control or dashboard to visualize information management.

2.4. PRIVACY PROTECTION PRINCIPLES AND GUIDELINES

Now we are living in a world where everything is connected and digitized. Every office, college, the organization is migrating from old system to a digital system. Users are now attracted towards cashless and digital transactions. However, our digital actions are monitored, tracked and analyzed for many undefined and undisclosed purposes. Personal information has become lucrative in the social and economic domain. Rendering by Tom Cochran "*Personal Information is the currency of the 21st century*" (Cochran, 2013). The noble purpose of users' information management is to make and produce better services for users. But, unscrupulous agencies could monitor users' activities and information using data mining or similar techniques to produce valuable information resulting in threats to security and privacy.

The purpose of mining users' information by service providers or external entities can be like spamming (bulk messaging), fishing, target advertising and to increase turnover in business, and understanding the users' behavior and interest (Huber, Mulazzani, Weippl, Kitzler, & Goluch, n.d.). In the health domain, there are set of applications where data mining is used for effective analysis of users' health information (Herland, Khoshgoftaar, & Wald, 2014). However, extensive analysis of users' information is a source of concern and has made users a worried lot (Ortlieb, 2014).

Latest applications and techniques must protect user information and should adhere to the legislation and privacy protection guidelines. However there is ambiguity regarding the identity of the real owner of users' information, how is the information processed and the purpose for which it is used. The following section describes such issues in detail.

2.4.1. LAWS OF IDENTITY

In 1993, Peter Steiner had made a short and interesting statement on Internet anonymity as a cartoon with the caption: "On the Internet, nobody knows you're a dog"[24]. The cartoon symbolizes that Internet communication between two anonymous entities is without understanding the concerns of, trust, safety, etc. In the framework of providing privacy and trust to Internet users, few critical issues with the Internet were identified by Kim Cameron who quotes (Kim Cameron, 2005):

> *"The Internet was built without a way to know who and what you are connecting to"*

The Internet is utilized on a large scale causing more exposure of information to the outside world. The users share their personal information on different websites as required to avail services. Users do not have an option but to enter personal information and accept whatever is given by service providers. The system should be designed to put the user in control. The deficiency of a framework to control confidential information is of primary concern. In today's universe of the Internet, it is necessary to empower the users' knowledge of what is privacy, security and who they are communicating with (Ann Cavoukian, 2012). The identity ecosystem should be built to provide the identity of a user considering the following elements of laws.

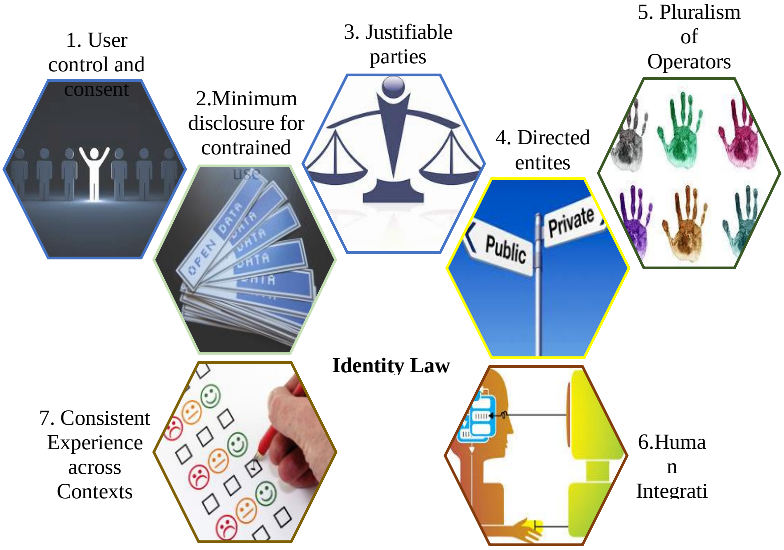

1. User control and consent
2.Minimum disclosure for contrained use
3. Justifiable parties
4. Directed entites
5. Pluralism of Operators
Private
Public
OPEN DATA
DATA
Identity Law
7. Consistent Experience across Contexts
6.Huma n Integrati

Figure 2-1 Identity law by Kim Cameron [49]

1. User Control and Consent: The simple and convenient system should put the user in control specifying which digital identities are defined, used and what information should be revealed. The user should decide what information should be exchanged in a variety of contexts. The system should reinforce the sense that the user is in control regardless of the context. Also, the user

[24] https://en.wikipedia.org/wiki/On_the_Internet,_nobody_knows_you%27re_a_dog

should have a control to know what information is being used for what kind of purposes.

2. Minimal Disclosure for Constrained Use: A system is said to be stable in the long term if it provides a unique solution that reveals the required minimum information, and defines its limited uses. This will help the systems to have the least damage. Important here to note is that the concept of "least identifying information" should be taken as meaning not only the fewest number of claims but the information least likely to identify a given individual across multiple contexts.

3. Justifiable Parties: A policy declaration on the use of information between two communicating parties should govern what exactly happens to the revealed information. Also, the identity of sharing parties should be explicitly mentioned in the policy document along with the conditions of data sharing

4. Directed Identity: Seeing entity types, the identity system should support universal and private entities. The public entities should be used in "omnidirectional" identifiers and private entities should be considered in "unidirectional". The identity system should facilitate the key method to prevent unnecessary disclosure of personal information and interrelated communications.

5. Pluralism of Operators and Technologies: Taking distinctive characteristics of contexts in the mind, there won't be a single, monolithic, and centralized identity system. So, the variety of channels should be provided with a system that enables the inter-networking of various identity systems. Hence, the identity system must support distinctive features.

6. Human Integration: Presently the communication between web servers and browsers are secured nicely. However, there should be another secure communication needed between the browser and the human mind. Hence, there is a need to extend identity system by integrating human-user.

7. Consistent Experience Across Contexts: The identity system must ensure users about the consistent experience in dealing with a variety of contexts. Based on the type of contextual identity, the system must enable digital identity. For example, in the case of the browsing context, users' real data should not be revealed. In the same way, the universal identity system must support various contexts like personal, professional, financial transaction, citizenship, etc.

The laws of identity should be obeyed by the identity systems. Hence, user consent and control are important principles that the service providers should follow while revealing the information.

2.4.2. PRIVACY BY DESIGN

To handle the issues of privacy and personal information, one of the ways to design solutions that consider privacy risks and prevent them is in the design phase. This will help the systems or solutions to avoid harm to user information and privacy breaches. Privacy by Design (Ann Cavoukian, 2012) is a framework to inculcate the principle of privacy by default in the design phase of the new product/system. To integrate privacy into the process, protocol and standards that preserve our control and freedom, 7 foundation principles are defined.

1. Proactive not Reactive: The proactive approach focuses on anticipation and prevention of privacy threats before they can happen. It is important to define and implement effective privacy protection strategies in the initial stage of system design and to continue them throughout its development.

2. Privacy as the Default Setting: There should be no requirement of user action to protect user privacy. There should be default rules and must deliver the maximum degree of privacy to users. IT systems or business practices should adhere to the information practices like collection limitation, purpose specification, minimum data disclosure and use.

3. Privacy Embedded into Design: The design and architecture of product should consider privacy as its center. Privacy should be embedded in the design phase in an all-inclusive way.

4. Full Functionality: The system’s functionality should not get damaged while integrating and employing Privacy by Design.

5. End-to-End Security: Privacy must be protected since the initial phase of data lifecycle to its last phase. At every part of the system, data and its privacy should be protected. Along with its privacy, the data security should be achieved using strong security measures and must be retained throughout the lifecycle of data.

6. Visibility and Transparency: All the operations, practices are following the promises and objectives according to as stated. All the components must be ensured, it is visible and transparent to the stakeholders. However, the visibility and transparency are achieved if information practices are followed (importantly like openness, compliance, and accountability). The privacy policies should be well written and communicated to the involved users and

the protection must be kept at a prominent level when it comes to transferring to external entities. The practices of personal information should be open to all users and should ensure compliance in Redressal mechanisms.

7. Respect for User Privacy: The user-centric systems should keep the user interests by means of appropriate measures like user control, notifications, etc. The user should be empowered to manage their own data.

The privacy guidelines emphasis on the transparency that each component of the system should be transparent. Similarly, the user should be aware of practices on personal information. This will help to enhance privacy awareness.

2.4.3. LIFE MANAGEMENT PLATFORM (LMP):

Most of our information in different domains like health, finance, insurance is vital and managing it with full security and privacy protection is the challenge. Also, losing the control over sensitive information is another issue to be addressed. LMP[25] will change the way individuals deal with personal information while interacting with different service providers. LMP provides a new way of not only storage of personal information but also, it's sharing in a secure manner.

“Informed Pull” and “Controlled Push” are the key features of LMP. Limited and required information are gathered in informed pull and information is securely shared with other interested entities in “Controlled Push”. In individual’s life, a lot of information needs to be managed in a more protected way. It's always a scaling for security and privacy against the type of information being shared and is as shown in Fig. 2-2.

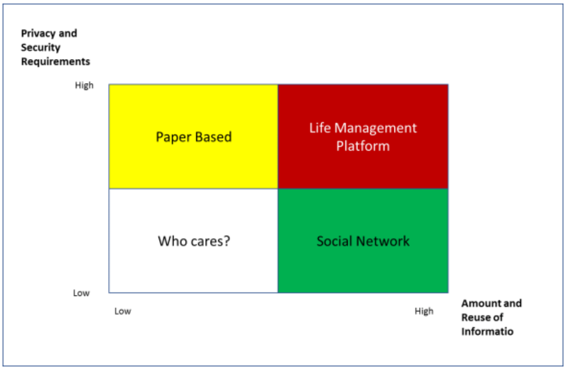


Figure 2-2 Life Management Platform[26]

[25] https://www.kuppingercole.com/report/an70608

A typical category of personal information is that which is very sensitive and remains to be in the form of paper. However, on social platforms, more and more public as well confidential data is routed. The LMP provides a set of tools to make information available using privacy-enhanced applications.

2.5. NEW LEGISLATIONS ON PERSONAL INFORMATION PRIVACY

A new Directive the EU is covering all the issues on personal information protection raised due to innovative and recent technological development. The old directive put forward in 1995 provides extensive choices of rights on personal information to the individual. According to the data protection law handbook (EU Agency for Fundamental Rights, 2014), an individual has the privileges to:

1. Access personal information of any controller
2. Correct personal information if found to be incorrect
3. Delete personal information
4. Protest unlawful use of personal information and its processing

Despite the directive, the new challenges require more clear, restrictive, definiteness, and an up-to-date law to protect the citizens of EU. Also, the challenge is to maintain and respect the right to privacy.

The new Directive GDPR is quite complete and will have a major influence on services and business processes. The idea is to preserve personal information protected and lay emphasis on explicit user consent for using personal information. The GDPR has defined the following terms which are important (The European Parliment and The Council of the EU - GDPR, 2016):

- Data subject: “An identified natural person or a natural person who can be identified, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person, in particular by reference to an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person.”

- Data controller: “the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes, conditions and means of the processing of personal data; where the purposes, conditions, and means of processing are determined by Union law or Member State law, the controller or the specific criteria for his nomination may be designated by Union law or by Member State law”

- Data subject's consent: "any freely given specific, informed and explicit indication of his or her wishes by which the data subject, either by a statement or by a clear affirmative action, signifies agreement to personal data relating to them being processed"

Talking about new regulation GDPR which was introduced in 2016 and expected to be in power from 2018, the "personal data breach" is an incident that leads to loss of users' privacy. This includes destruction, unauthorized disclosure or illegal access, storage, processing of personal information without user consent.

GDPR aims to present a new set of strict rules for handling personal information in recent and advanced information systems. The new rules are taking personal information breaches seriously and huge penalty will be marked if any organization or company uses the poor design or are negligent in their system. Based on the severity of personal information breaches, the amount of fine is calculated.

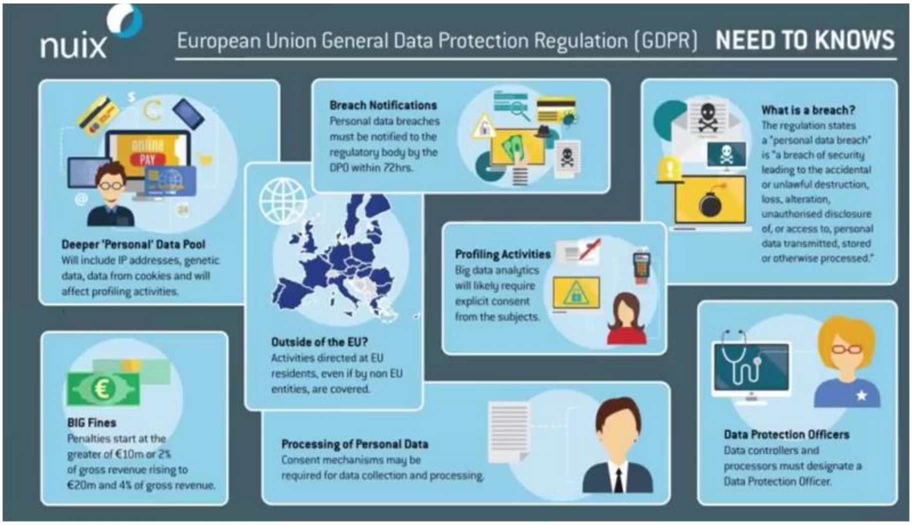

nuix
European Union General Data Protection Regulation (GDPR) NEED TO KNOWS
online PAY
Deeper 'Personal' Data Pool
Will include IP addresses, genetic data, data from cookies and will affect profiling activities.
Breach Notifications
Personal data breaches must be notified to the regulatory body by the DPO within 72hrs.
What is a breach?
The regulation states a "personal data breach" is "a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed."
Outside of the EU?
Activities directed at EU residents, even if by non EU entities, are covered.
Profiling Activities
Big data analytics will likely require explicit consent from the subjects.
BIG Fines
Penalties start at the greater of €10m or 2% of gross revenue rising to €20m and 4% of gross revenue.
Processing of Personal Data
Consent mechanisms may be required for data collection and processing.
Data Protection Officers
Data controllers and processors must designate a Data Protection Officer.

Figure 2-3 GDPR- Need to knows[26]

A quick insight[27] on GDPR is represented in fig. 2-3. A strict deadline is given to Data Protection Officers (DPO) to notify the regulatory body within 72 hrs. if there is data breach [52]. The service provider or Big data analytics must take explicit consent from data owners (users) before processing and perform profiling activities. This GDPR

[26] https://www.youtube.com/watch?v=67i_Uw8UeUE

imposes huge fines (punishment of minimum 10m Euro or 2% of gross revenue) to those companies who do not adhere to GDPR regulation.

Coming back to the user's explicit consent, the challenge is to make it simple and informative (Karahasanovic et al., 2009). To achieve this, the user must be informed about the collection, storing, processing, and sharing of personal information. Another challenge is to identify the problematic grey areas where the personal information measures can be improved and employed. GDPR also states that privacy should be considered as default in the design phase of product/service. Hence, it required following the PbD principles mentioned earlier in this chapter.

It is also equally important to follow the Organization of Economic Cooperation and Development (OECD) principles that are briefly defined here (OECD, 2013):

1. The information collection limit
2. The consent less use of personal information
3. Information should match with its use
4. Purpose of personal information collection should be stated clearly
5. There should be enough security measures to protect user information
6. In case of information privacy breaches, the controller is fully responsible
7. An individual should have complete access to his/her information
8. The change in personal information policy should be notified the individual

Considering a multi-faceted, complex and evolving term of privacy, the rules and regulations are always going to change in the context of the fast-growing digital world. The laws need to be refined at regular intervals to adopt the progress of technologies and services offered. However, it is challenging to improve the laws and make it available on time as it is a time-consuming process. So, there is need to do a research on enhancing privacy awareness and to define innovative ways for protection of personal information.

Seeing the limitations and essential duties of the data controller mentioned in GDPR (Article 23 and 30), the data controller or processor is obliged to specify the drive of the personal information processing, the type of personal information, retention time, and security measures to avoid unauthorized access or transfer. Each data controller should mention the activities on personal information and it should be recorded in writing and in electronic form. The content of the record should include details of the data controller, its usage, benefiting parties, release or deletion time, etc.

Similarly, the Information Technology Act 2008 in India, to which this researcher belongs, speaks on the protection of personal information. Considering privacy, two sections (43A and 72A) describe punishment (3 years custody or Rs. 5,00,000 fine) to the person who originates wrongful gain or the cause loss to the owner of the personal information. However, there is huge scope to improve and enforce in reality (Ryan, Merchant, & Falvey, 2011b).

In the present state, the controller or service provider should, but are not describing personal information practices clearly in the policy documents to enhance the privacy awareness and gain trust. This is forcing the user to be clueless about the content of policy documents. Hence, there is need to do extensive analysis of privacy policies and visualize them to enhance users' awareness.

2.6. SURVEY, PRIVACY ENHANCING TECHNOLOGIES AND RELATED WORK

The state of the art technologies and current work is classified into the following three sections as 1. survey on privacy awareness, 2. Use of PET, and 3. Developing tools.

2.6.1. APPROACH BASED ON SURVEY AND PRIVACY ENHANCING TECHNOLOGIES

A survey carried out by TRUSTArc[28] show that 96% of the companies believe that there is an increase in importance to managing privacy. However, the analysis also revealed that managing privacy is not so easy. 98% of the respondents said that it is becoming increasingly difficult to manage privacy.

A comprehensive nationwide study on the privacy and security habits of the Indian internet users was carried out (Kumaraguru, 2012). 10,427 responses were collected from various cities in India. The Study showed that the participants were more concerned about their privacy. An important finding was that the password was considered as the most personal information as compared to religion, mobile phone number, and health-related information which ranked lower. Another finding was that about 40% of the participants would never save/share personal information in/ through emails. Privacy seems to be the primary reason for this behavior.

College student group -mockup method was used for the device visualization technique (Abdullah et al., 2008). These visualization techniques have helped them to design and develop a Firefox extension whose basic idea is to help the end users to see their web search activities and enhance their awareness about online actions. This tool shows visited sites with their category, types of revealed information along with a timestamp (date and time). However, such work needs to be enhanced to enable self-monitoring.

[28] https://www.trustarc.com/products/iapp-gdpr-readiness-assessment/

Another study was carried out to understand users' knowledge of privacy policy. For this study, social websites were selected (Facebook, Twitter, LinkedIn, etc.). The important finding in this study was the conclusion that Internet users do not completely understand what they have agreed to while registration (Zeadally & Winkler, 2016).

In the US, a group survey was carried out to check the compliance of 35 frequently used social health websites with the Fair Information Practices (FIP) (Savla & Martino, 2012). An important finding was that the user's privacy is at substantial risk, if the privacy policy does not comply with the FIP principles, and the privacy policies of healthcare service providers do not support informed decisions.

"Agree or disagree" is the nature of acceptance of the privacy policies set by service providers when users avail services offered by them. Looking at the privacy policies of service providers, the users are unclear about personal information management (to what level user information is being gathered, used and made available to others for reuse). Considering the length and the complex language of privacy policy, the privacy policies are not effective in terms of reading, understanding, etc. A survey carried by Futuresight(Futuresight, 2011) confirms that at least 50% of mobile users blindly accept the terms and conditions without reading. Hence there is a need to raise privacy policy awareness by developing privacy management tools.

Many years have been invested in developing the technologies that adhere to the principles of privacy and data protection. Some of the principles include anonymization, pseudonymization, and data minimization given by the European Network and Information Security Agency (ENISA)(Mantelero, 2016), an agency of the EU, which aims to improve the security of network and information in EU. Privacy Enhancing Technologies are the outcome of these guidelines that cover a wide range of technologies. A study from Denmark denotes the analysis and classification of protection methods provided by PETs (Fritsch, 2012) and is as shown in Fig 2-4.

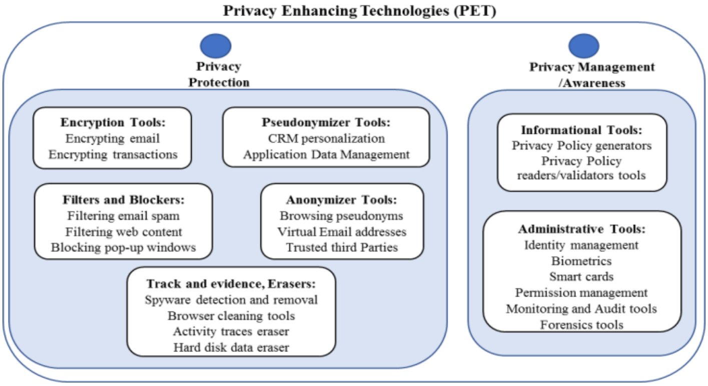


Figure 2-4 Classification of PET mechanisms (Fritsch, 2012)

According to the study as illustrated in Fig 2-4, the basic categories are protection and management/awareness of privacy. Looking at several tools, the awareness tool can be improvised to spread awareness among Internet users.

There are three categories suggested to Privacy Enhancing Technologies (PET) by Wang and Kobsa (Wang & Kobsa, 2008).

1. Protection of identity: In this category, the aim is to protect users' identity, replacing with the non-traceable (e.g. Who are you?). Anonymizer is one of the examples of PET.

2. Seclusion: In this category, the aim is to protect users from being concerned or troubled from the unwanted or the unknown (e.g., Spam or bulk emails). "Mailwasher", "PrivacyBird", "Thunderbird" are few examples of PET.

3. Control over data: In this category, the aim is to empower the user to control user information (e.g. user will decide what data to reveal to whom under what circumstances). OpenID is one of the examples of PET.

Along with the above categories, another category would be education and awareness that can also share their contributions in privacy protection. Let's consider an example of adult vs teenagers. In the context of privacy, privacy values, and its invasion, adults are more concerned about privacy as compared to teenagers. School and college students fascinated towards social media, give personal information without a thought to its consequences. Also, the working of the Internet or business model of companies is not known completely to the kids/college students (Khan & Hasan, 2016).

This definition includes essential elements like user choice and control, minimum disclosure, informed consent. Platform for Privacy Preferences (P3P)[29] [64] specification has extended privacy related elements/attributes that include machine readable and automatic evaluation of privacy policies. It means that the P3P enables the companies to express privacy policies in the standard way that is readable and interpreted by users. P3P defines a set of rules to service providers/companies that they should publish intended uses of users' information collected. It is expected that the rules are designed to enhance users' control over personal data management, and understand its purpose. Considering P3P guidelines to the websites to publish their personal data practices to the users in terms of the privacy policy for example. But, P3P does not include mechanisms for transferring data or for securing personal data

[29] http://www.w3.org/TR/2002/REC-P3P-20020416

in transit or storage. P3P may be built into tools designed to facilitate data transfer, these tools lack appropriate security safeguards.

The PrimeLife Privacy Dashboard[30] is a browser extension which helps the user to track what information is collected by the websites he visits. For this, it collects information about the website the user is currently visiting, such as whether it has a P3P policy, whether it collects cookies, and whether it is certified by trust seals. The dashboard then provides a visual ‘privacy quality’ rating of a website: the presence of a P3P version of the privacy policy increases the rating while the presence of external or flash cookies decreases it.

However, the low adoption of P3P is a major disadvantage of this approach: a website may have a good privacy policy, but may be rated low because of the lack of a P3P version. Also, the content of the privacy policy, if it does exist, is not considered.

Using K-anonymity and anonymization, a framework was developed to protect the privacy of healthcare data (Chen, Yang, Wang, & Niu, 2012). This framework mostly focused on elements of privacy policies (permitted users, methods of collection, and use) of health domain. This work needs to be extended to consider other important components to users like security, consent, third-party sharing and retention of data.

Based on Authentication mechanism, a verification framework is proposed by Liu et. al. (Liu et al., 2013) that helps to carry out Data Big Audit. “Closeness” (Venkatasubramanian, 2010) and “Slicing” (Li, Li, Zhang, & Molloy, 2012) are few examples of anonymization mechanisms used for protection of users’ privacy in Big Data.

Mowbray et al. (Mowbray, Pearson, & Shen, 2012) used obfuscation as an approach in data transmission. The level of obfuscation depends on the context of data transmission, leads to weak protection of information. There is a need to develop context-based partial identities that are dynamic in nature.

2.6.2. APPROACH BASED ON MACHINE LEARNING

Organizations do not have effective ways of linking their written privacy policies with the implementation (C. A. Brodie, Karat, & Karat, n.d.). SPARCLE is a privacy workbench which enables organizational users to enter policies in natural language, parse the policies to identify policy elements and then generate a machine readable (XML) version of the policy. In this paper, the researchers have presented the strategies employed in the design and implementation of the natural language parsing capabilities that are part of the functional version of the SPARCLE authoring utility. They have created a set of grammars which execute on a shallow parser that is

[30] http://www.w3.org/2011/D1.2.3/.

designed to identify the rule elements in privacy policy rules and they have presented empirical usability evaluation data from target organizational users of the SPARCLE system and highlighted the parsing accuracy of the system with the organizations' privacy policies.

The researchers of this paper have conducted a survey which investigated some of the most relevant approaches both in the areas of single-document and multiple document summarization, giving special emphasis to empirical methods and extractive techniques (Das & Martins, 2007).

A solution is proposed that automatically analyses privacy policy text and shows what personal information is collected (Costante, den Hartog, & Petković, 2013). This solution is based on the use of Information Extraction techniques and represents a step towards the more ambitious aim of automated grading of privacy policies. This thesis focuses on analyzing the contents of a policy, namely the part regarding data collection. The authors of this paper describe that the more accuracy can be achieved, by using ontologies and thesaurus, to enrich the gazetteer lists with synonyms and close lexical concepts. The authors believe that the system would not benefit from the use of syntactic parsing since its computational costs would not allow providing real-time responses to the users.

There is an increasing awareness of a fundamental need to address privacy concerns in information technology and that doing so will require an understanding of policies that govern information use as well as the development of technologies that can implement such policies (C. Brodie, Karat, Karat, & Feng, n.d.). This paper describes the work of identifying organizational privacy requirements, analyzing existing technology, ongoing research to identify approaches that address these requirements and authors' efforts to design a privacy management workbench which facilitates privacy policy authoring, implementation, and compliance monitoring.

A solution is presented to assist the user by providing a structured way to browse the policy content and by automatically assessing the completeness of a policy, i.e. The degree of coverage of privacy categories is important to the user (Costante, Sun, Petković, & den Hartog, 2012). The privacy categories are extracted from privacy regulations while text categorization and machine learning techniques are used to verify which categories are covered by a policy. The authors have used an automatic classifier to associate the right category to paragraphs of a policy with an accuracy approximating that obtainable by a human judge. However, the authors state that the effectiveness of the classifiers can be improved which ultimately will improve the overall accuracy of the system.

It is believed that the user privacy management needs to be both from the user's side and from the web application side and the Web applications must be compliant with privacy policies (W. Yu, Doddapaneni, & Murthy, 2006). The authors state that the

User awareness about information security is the factor for user side privacy management. In this paper, the authors have designed and implemented an approach for a privacy policy checker engine that automatically verifies and certifies a Web service application based on the levels of overall privacy principle compliance and privacy statement compliance

Privacy has gaining a high concern and hence, creating well documented and comprehensive organizational privacy policies still remain a challenge (W. D. Yu & Murthy, 2007). This paper presents results on a special Privacy Policy Modeling Language Processor (PPMLP) based on service-oriented architecture (SOA) for an organization to model the structure and contents of private policy they want through a meta-type of privacy policy specifications.

Social Networking Sites such as MySpace, Facebook, and LinkedIn have attracted millions of users and have become established places for keeping contact with old acquaintances and meeting new ones (Aïmeur, Gambs, & Ho, 2009). Nonetheless, due to lack of user awareness and proper privacy protection tools, huge quantities of user data, including personal information, pictures and videos are quickly falling into the hands of authorities, strangers, recruiters and even the public at large. By using Social Networking Sites and accepting their privacy policy, users have volunteered to relinquish their ownership of their own data. The authors of this paper present a User Privacy Policy (UPP) which provides users with an easy and flexible way to specify and communicate their privacy concerns to other users, third parties and with the Social Networking Site provider.

2.6.3. PRIVACY AWARENESS/VISUALIZATION TOOLS

“It is difficult to protect your privacy even if you know how,” said by Lorrie Cranor (Pedro G. Leon, Blase Ur, Rebecca Balebako, Lorrie Faith Cranor, Richard Shay, 2012) after analyzing the business model of the social network along with connected tracking entities. According to her team’s survey on privacy protection/ awareness tools, the participants were not aware of the tools and struggled to use them. Hence, there is a need to do a comprehensive analysis of privacy policies and develop a simple dashboard that will help to enhance users’ privacy knowledge.

Several tools have been developed to assist users in not being tracked online. The tools help to provide enhanced privacy awareness knowledge. As stated earlier, this thesis focuses on privacy awareness and not privacy protection. So, tools that are discussed here are more oriented towards awareness. Some of the tools are:

- Just Delete Me[31]:
  It is a list of the most popular web apps and services by providing links to delete your account from those services. When you click on a service, you're automatically taken to the page where you can delete your account so you don't have to go searching for it. Each one (app/service) is color coded representing the difficulty level of deletion. Green is easy, yellow is medium, red is hard, and black is impossible. For example, Amazon.com and the NewYorkTimes.com are rated "hard" to delete, while movie directory IMDB.com and PayPal are listed as "easy." Sites such as Pinterest and Netflix are "impossible". This Provides up to date information about whether an account is easy to delete before you sign up.

- Collusion[32]
  It is a Firefox extension that will show you in real time, which sites are tracking you, where you picked up their tracking cookies, and what they can see. Collusion looks to offer more transparency to users by creating a visualization of how your data is being spread to different companies as you navigate the web. Each time it detects data being sent to a behavioral tracker, it creates a red (advertisers), gray (websites) or blue dot on the visualization and shows the links between the sites you visit and the trackers they work with. Collusion does the effective job by visualizing how your data goes in far and wide to places without your knowledge.

- Free VPNs (Zenmate, DotVPN, TOR)[33]
  Services like ZenMate [82], DotVPN are simple VPN solutions, providing easy to use Security and Privacy on the Internet. These services create a tunnel like a Virtual Private Network (VPN) between user’s devices and service’s server network. This impenetrable tunnel prevents snoopers, hackers, governments and ISP’s from spying on the web browsing activities, downloads, credit card information or anything else the users send over the network. With these services, users can change his/her IP address to hide the real location, circumvent network restrictions and unblock Geo-restricted sites. These services can be chosen from currently offered different country locations. The User can install an add-on for these services in the web browser.

  It is a simple VPN solution, providing easy to use Security and Privacy on the Internet. ZenMate creates a tunnel similar to a Virtual Private Network (VPN) between your device and our server network. This impenetrable

[31] http://justdelete.me/

[32] http://collusion.toolness.org/

[33] https://zenmate.com

tunnel prevents snoopers, hackers, governments and ISP's from spying on your web browsing activities, downloads, credit card information or anything else you send over the network. With ZenMate you can change your IP address to hide your real location, circumvent network restrictions and unblock Geo-restricted sites. It is helpful when you are using a public Wi-Fi network.

- AdBlock Plus[34]
  AdBlock Plus is a free extension that allows you to block annoying ads, disable tracking and block domains known to spread malware. In order to block the ads, you need to add external filter lists. Filter lists are essentially an extensive set of rules that tell AdBlock Plus which elements of a website to block.

- Tor[35]
  The name is an acronym derived from the original software project name The Onion Router. It is free software, and an open network that helps you defend against traffic analysis, a form of network surveillance that threatens personal freedom and privacy, confidential business activities and relationships, and state security. Using Tor makes it more difficult for Internet activity to be traced back to the user: this includes "visits to Web sites, online posts, instant messages, and other communication forms".

- HTTPS Everywhere[36]
  HTTPS Everywhere is a Firefox, Chrome, and Opera extension that encrypts your communications with many major websites, making your browsing more secure. HTTPS Everywhere is produced as a collaboration between The Tor Project and the Electronic Frontier Foundation and is recently updated with thousands of more rules, ensuring HTTPS is enabled on as many sites as possible. It automatically connects you to the HTTPS version of thousands of websites. If you're doing-particularly security-conscious work, it could help ensure that all your information is protected for that session.

- A simple visualization tool, “Privacy Pal” (Tucker, Tucker, & Zheng, 2015) has been developed to make users aware of the security and privacy risks associated with third-party applications permission granting. The case study shows that Privacy Pal helps to understand the threats to privacy.

[34] https://adblockplus.org

[35] https://www.torproject.org

[36] https://www.eff.org/HTTPS-EVERYWHERE

- Privacy Badger[37]
  Privacy Badger is a browser plug-in developed by the Electronic Frontier Foundation (EFF), which can help users to block tracking from advertisers and third parties.

- Web of Trust[38]
  Web of Trust (WOT) is a website reputation and review service that helps people make informed decisions about whether to trust a website or not. WOT is based on a unique crowdsourcing approach that collects ratings and reviews from a global community of millions of users who rate and comment on websites based on their individual experiences. Web of Trust goes beyond simple vote-counting with an algorithm that incorporates user reputation, and it pulls in data from third-party blacklists as well. Web of Trust marks safe links with a green icon and dangerous ones with a red icon. For example, if your friend posts an article on Facebook, and the link might lead somewhere fishy, Web of Trust puts a red icon next to it.

- Lightbeam[39]
  Lightbeam is a Firefox add-on that shows the user the first and third-party websites interacted by the user using interactive visualizations. It displays a graph of visited and interacted websites by the user and tracking websites to which they provide information. After installing and enabling it will create a record of events for the websites visited by the user and every third-party site that is stored on the user's browser. It displays a graph to highlight the interactions between visited websites by the user and the third parties. It adds the website to the graph as soon as it is visited by the user. As visualizations grow, the user can observe the relationships between the various first and third-party websites stored in the user’s data. The user can not only reset or save data but also contribute his data to the Lightbeam database at any time.

- Dinconnect.me[40]
  Disconnect developed a user-friendly privacy and security software used by millions of people. Disconnect created the privacy and security tools that make it easy for the user to understand about online privacy and provides the ability to control access user's personal information. It makes the user aware of unsecured connections and hidden requests for user's personal info and allows the user to block trackers and hackers. Routes all the Internet activity

[37] https://www.eff.org/privacybadger

[38] https://www.mywot.com/

[39] https://www.mozilla.org/en-US/lightbeam/

[40] https://disconnect.me/about

of the user through an encrypted tunnel, which prevents wireless eavesdropping. Disconnect's VPN technology blocks more than 5000 malicious trackers, sources of malware identity and theft.

- Ghostery[41]
  It's owned by privacy technology and advertising company Ghostery Inc. It enables users to easily detect and control web bugs, which are objects embedded in a web page, invisible to the user, which allow the collection of the user's browsing habits. Ghostery also has a privacy team that creates profiles of page elements and companies for educational purposes. Ghostery blocks HTTP requests and redirects according to their source address in two ways: cookie blocking and cookie protection. When cookie protection is enabled, if a cookie is selected from Ghostery's list, it is not accessible to anyone but the user and thus cannot be read when called upon. Ghostery reports all tracking packages detected and whether Ghostery has blocked them or not, in a temporary purple overlay box at the bottom right of the screen.

- MyPermission[42]
  This is another online privacy protection tool from Online Permissions Technologies for application and browsers. This tool provides real-time alerts to users as soon as any application gets connected. The user is enabled to control over their data that are accessed by the applications. The single interface will show the list of all service permissions. This app will give rise to other functionality like revoke, trust when the user is online.

- Terms of Service; Didn't read (ToS; DR)[43]
  "Terms of Service; Didn't Read" (short: ToS; DR) is a project started in June 2012. ToS: DR rate and label website terms and privacy policies, from very good (Class A) to very bad (Class E). While creating a new account on any website the user doesn't read the privacy policy and license agreement statement because it is very long and not understandable to the user. But that policy is very important as user personal information is concerned. So ToS; DR makes the user aware of how the famous web services use or handle their personal information, along with the rating given to the website or web service.

[41] https://www.ghostery.com/

[42] https://mypermissions.com/whoweare

[43] https://tosdr.org/about.html

A comparison of these privacy-enhancing/awareness tools is shown in table 1. Most of the tools are blocking third parties and help the users to avoid being trapped by third/unauthorized/untrusted parties(Khajuria, Sørensen, & Skouby, 2017).

Tool Name	Functionality	Type of Tool
Privacy Badger	This tool helps users to block tracking from advertisers and third parties.	Blocking
Lightbeam	Using this tool, the user will be aware of how the first and third-party websites interact and their relationships.	Awareness
Disconnect	Unsecured connections and hidden requests for users' personal information are visualized by this tool. Also, this tool allows the user to block trackers and hackers.	Awareness and Blocking
Ghostery	Detecting and blocking of invisible trackers is the main objective of this tool	Blocking
MyPermission	This tool gives users complete control over those apps that access the users' data	Control and Blocking
Terms of Service; Didn't Read (ToSDR)	Using ToSDR, rating, and labeling of the terms and privacy policies of major websites can be seen, based on a user community. The ratings cover a range from very good (Class A) to very bad (Class E).	Privacy Awareness
Web of Trust (WOT)	Rating of websites based on user comments/community	Trust Awareness

Table 1 Comparision of Privacy-enhancing/awareness tools (Khajuria et al., 2017)

2.7. SUMMARY ON PRIVACY, PRIVACY AWARENESS, STATE OF THE ART

Looking at the various contexts of the user (health care, anonymity, shopping), user information is gathered at service providers' end at a high rate at large scale.

Privacy has many definitions but the emphasis is given on certain elements are like consent, choice, context, excessive information collection, disclosure without consent, etc. Privacy awareness is very useful to understand the requirements based on the changing definition of privacy. To empower the users, it is very important to spread awareness to Internet users.

Considering recent technologies in today modern digital world, the users have a lot of convenient services that make their daily tasks easier. However, the price of these services is the loss of privacy as the companies are accessing more than necessary information like personal attributes, device information, biometrics information, user choices, location, etc.

Certain privacy protection guidelines are being specified that help to have basic protection. Even though, the legislation encompasses privacy, the enforcement is an issue in several countries. But having guidelines, legislation and PET will help to control information processing happening now without users' explicit consent. This should be noted that the legislation like GDPR and the guidelines like PbD will help countries like India to improve their current legislation like IT Act.

There are still issues that need to be given more attention like too much of information collection and processing, ownership and user control, users' unawareness about sharing and re-using of their information in spite of having a privacy policy by service providers which are complicated and difficult to understand.

The researchers are also putting their contribution to developing protection or awareness mechanisms/tools. The tools developed so far help to track and block third-party entities. However, there is a need to enhance and protect users' privacy by developing privacy awareness/management tools.

2.8. REFERENCES

Abdullah, K., Conti, G., & Beyah, R. (2008). A Visualization Framework for Self-Monitoring of Web-Based Information Disclosure. In *2008 IEEE International Conference on Communications* (pp. 1700–1707). IEEE. http://doi.org/10.1109/ICC.2008.328

Aïmeur, E., Gambs, S., & Ho, A. (2009). UPP: User Privacy Policy for Social Networking Sites. In *2009 Fourth International Conference on Internet and Web Applications and Services* (pp. 267–272). IEEE. http://doi.org/10.1109/ICIW.2009.45

Alan Westin, Osborne M., & Jr. (1967). Privacy and Freedom. *New York*. American Bar Association. http://doi.org/10.2307/40708684

Alsagri, H. S., & Alaboodi, S. S. (2015). Privacy awareness of online social networking in Saudi Arabia. *2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, CyberSA 2015*. http://doi.org/10.1109/CyberSA.2015.7166111

Altman, I. (1975). *The environment and social behavior : privacy, personal space, territory, crowding*. Brooks/Cole Pub. Co. Retrieved from https://books.google.co.in/books/about/The_environment_and_social_behavior.html?id=GLBPAAAAMAAJ&redir_esc=y

Ann Cavoukian, P. . (2012). The 7 Foundational Principles. *Privacybydesign.ca*. Retrieved from www.privacybydesign.ca

Bergmann, M. (2009). Testing Privacy Awareness (pp. 237–253). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-03315-5_18

Bonneau, J., & Preibusch, S. (2010). The Privacy Jungle:On the Market for Data Protection in Social Networks. In *Economics of Information Security and Privacy* (pp. 121–167). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4419-6967-5_8

Brodie, C. A., Karat, C.-M., & Karat, J. (n.d.). An Empirical Study of Natural Language Parsing of Privacy Policy Rules Using the SPARCLE Policy Workbench. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.298.2880&rep=rep1&type=pdf

Brodie, C., Karat, C.-M., Karat, J., & Feng, J. (n.d.). Usable Security and Privacy: A Case Study of Developing Privacy Management Tools. Retrieved from https://cups.cs.cmu.edu/soups/2005/2005proceedings/p35-brodie.pdf

Bruce, S. (2006). The Eternal Value of Privacy - Schneier on Security. Retrieved February 27, 2017, from https://www.schneier.com/essays/archives/2006/05/the_eternal_value_of.html

Cavoukian, A. (2008). Privacy in the clouds. *Identity in the Information Society*, *1*(1), 89–108. http://doi.org/10.1007/s12394-008-0005-z

Cavoukian, A., & Jonas, J. (2012). Privacy by Design in the Age of Big Data. Retrieved June 28, 2014, from https://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf

Chen, L., Yang, J. J., Wang, Q., & Niu, Y. (2012). A framework for privacy-preserving healthcare data sharing. *2012 IEEE 14th International Conference* *on E-Health Networking, Applications and Services, Healthcom 2012*, 341–346. http://doi.org/10.1109/HealthCom.2012.6379433

Cochran, T. (2013). Personal Information is the Currency of the 21st Century. Retrieved August 10, 2016, from http://allthingsd.com/20130507/personal-information-is-the-currency-of-the-21st-century/

Conner, M., & Norman, P. (2005). *Predicting health behaviour : research and practice with social cognition models*. Open University Press. Retrieved from https://books.google.co.in/books/about/Predicting_Health_Behaviour.html?id=MZhzQgAACAAJ&redir_esc=y

Costante, E., den Hartog, J., & Petković, M. (2013). What Websites Know About You (pp. 146–159). Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-35890-6_11

Costante, E., Sun, Y., Petković, M., & den Hartog, J. (2012). A machine learning solution to assess privacy policy completeness. *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society - WPES '12*, 91. http://doi.org/10.1145/2381966.2381979

Das, D., & Martins, A. F. T. (2007). A Survey on Automatic Text Summarization. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.5100&rep=rep1&type=pdf

Dourish, P., & Anderson, K. (2006). Collective Information Practice: Exploring Privacy and Security as Social and Cultural Phenomena. *Human-Computer Interaction*, *21*(3), 319–342. http://doi.org/10.1207/s15327051hci2103_2

Economist. (n.d.). Getting to know you | The Economist. Retrieved May 18, 2015, from http://www.economist.com/news/special-report/21615871-everything-people-do-online-avidly-followed-advertisers-and-third-party

EU Agency for Fundamental Rights. (2014). Handbook on European data protection law. In *European data protection law*. Retrieved from http://fra.europa.eu/en/publication/2014/handbook-european-data-protection-law

EU Directive-GDPR. (2016). Retrieved May 18, 2015, from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=en

Fortes. (2016). Privacy concerns and online purchasing behaviour: Towards an integrated model. *European Research on Management and Business Economics*, *22*(3), 167–176. http://doi.org/10.1016/J.IEDEEN.2016.04.002

Fritsch, L. (2012). *State of the art of Technology (PET)*.

Futuresight. (2011). Futuresight:, "User perspectives on mobile privacy, Summary of research findings." Retrieved from http://www.gsma.com/publicpolicy/wp-content/uploads/2012/03/futuresightuserperspectivesonuserprivacy.pdf

Govt of India. (n.d.). Guidelines for securing Identity Information. Retrieved from http://www.dot.gov.in/sites/default/files/2017_05_26 Circulation Letter for Security of Information.pdf

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal Of Big Data*, *1*(1), 2. http://doi.org/10.1186/2196-1115-1-2

Huber, M., Mulazzani, M., Weippl, E., Kitzler, G., & Goluch, S. (n.d.). Exploiting Social Networking Sites for Spam Our FITM Attack FITM -Friend in the Middle Attack. Retrieved from https://www.sba-research.org/wp-content/uploads/publications/Poster_CCS_2010.pdf

Jyoti Panday. (2017). Aadhaar: Ushering in a Commercialized Era of Surveillance in India | Electronic Frontier Foundation. Retrieved July 12, 2017, from https://www.eff.org/deeplinks/2017/05/aadhaar-ushering-commercialized-era-surveillance-india

Karahasanovic, A., Brandtzaeg, P. B., Vanattenhoven, J., Lievens, B., Nielsen, K. T., & Pierson, J. (2009). Ensuring Trust, Privacy, and Etiquette in Web 2.0 Applications. *Computer*, *42*(6), 42–49. http://doi.org/10.1109/MC.2009.186

Khajuria, S., Sørensen, L., & Skouby, K. E. (2017). *Cybersecurity and Privacy - Bridging the Gap*. River Publishers. Retrieved from http://www.riverpublishers.com/book_details.php?book_id=434

Khan, R., & Hasan, R. (2016). The Story of Naive Alice: Behavioral Analysis of Susceptible Internet Users. *Proceedings - International Computer Software and Applications Conference*, *1*, 390–395. http://doi.org/10.1109/COMPSAC.2016.206

Kim Cameron. (2005). The Laws of Identity -. Retrieved June 21, 2014, from http://www.identityblog.com/stories/2005/05/13/TheLawsOfIdentity.pdf

Kumaraguru, P. (2012). Privacy in India : Attitudes and Awareness V 2 . 0.

Lee, R., & Janna, A. (2014). The Future of Privacy | Pew Research Center. Retrieved February 27, 2017, from http://www.pewinternet.org/2014/12/18/future-of-privacy/

Levchenko, K., Pitsillidis, A., Chachra, N., Enright, B., Félegyházi, M., Grier, C., … Savage, S. (n.d.). Click Trajectories: End-to-End Analysis of the Spam Value Chain. Retrieved from https://cseweb.ucsd.edu/~savage/papers/Oakland11.pdf

Li, T., Li, N., Zhang, J., & Molloy, I. (2012). Slicing: A New Approach for Privacy Preserving Data Publishing. *IEEE Transactions on Knowledge and Data Engineering*, *24*(3), 561–574. http://doi.org/10.1109/TKDE.2010.236

Liu, C., Ranjan, R., Zhang, X., Yang, C., Georgakopoulos, D., & Chen, J. (2013). Public Auditing for Big Data Storage in Cloud Computing -- A Survey. In *2013 IEEE 16th International Conference on Computational Science and Engineering* (pp. 1128–1135). IEEE. http://doi.org/10.1109/CSE.2013.164

Lou, W. L. W., & Ren, K. R. K. (2009). Security, privacy, and accountability in wireless access networks. *IEEE Wireless Communications*, *16*(August), 80–87. http://doi.org/10.1109/MWC.2009.5281259

Mantelero, A. (2016). Draft Guidelines on the Protection of Individuals With Regard To the Processing of Personal Data in a World of Big Data. *Council of Europe*. Retrieved from https://rm.coe.int/16806ebe7a

Mowbray, M., Pearson, S., & Shen, Y. (2012). Enhancing privacy in cloud computing via policy-based obfuscation. *Journal of Supercomputing*, *61*(2), 267–291. http://doi.org/10.1007/s11227-010-0425-z

Narayanan, A., & Shmatikov, V. (n.d.). Robust De-anonymization of Large Sparse Datasets. Retrieved from https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

Neill, E. (2001). *Rites of privacy and the privacy trade : on the limits of protection for the sacred self*. McGill-Queen’s University Press.

NIST. (2010). Personally identifiable information. National Institute of Standards and Technology. Retrieved from https://en.wikipedia.org/wiki/Personally_identifiable_information#cite_note-4

OECD. (2013). OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data - OECD. Retrieved September 23, 2017, from http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

Ortlieb, M. (2014). The Anthropologist ' s View on Privacy. *IEEE Security & Privacy*, (June), 85–87.

Pedro G. Leon, Blase Ur, Rebecca Balebako, Lorrie Faith Cranor, Richard Shay, and Y. W. (2012). *Why Johnny Can't Opt Out: A Usability Evaluation of Tools to Limit Online Behavioral*.

Rogers, R. W. (1975). A Protection Motivation Theory of Fear Appeals and Attitude Change1. *The Journal of Psychology*, *91*(1), 93–114. http://doi.org/10.1080/00223980.1975.9915803

Rule, J. B., & Greenleaf, G. W. (Graham W. (2010). *Global privacy protection : the first generation*. Edward Elgar.

Ryan, P. S., Merchant, R., & Falvey, S. (2011a). Regulation of the Cloud in India. *Journal of Internet Law*, *15*(4), 13.

Ryan, P. S., Merchant, R., & Falvey, S. (2011b, July 30). Regulation of the Cloud in India. Retrieved from https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1941494

Savla, P., & Martino, L. D. (2012). Content analysis of privacy policies for health social networks. *Proceedings - 2012 IEEE International Symposium on Policies for Distributed Systems and Networks, POLICY 2012*, 94–101. http://doi.org/10.1109/POLICY.2012.20

Solove, D. J. (n.d.). "I"ve Got Nothing to Hide' and Other Misunderstandings of Privacy. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=998565

The European Parliment and The Council of the EU - GDPR. (2016). GDPR. *Official Journal of the European Union*. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

Tucker, R., Tucker, C., & Zheng, J. (2015). Privacy pal: Improving permission safety awareness of third party applications in online social networks. *Proceedings - 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security and 2015 IEEE 12th International Conference on Embedded Software and Systems, H*, 1268–1273. http://doi.org/10.1109/HPCC-CSS-ICESS.2015.83

United Nations. (n.d.). Universal Declaration of Human Rights | United Nations. Retrieved September 23, 2017, from http://www.un.org/en/universal-declaration-human-rights/

Venkatasubramanian, S. (2010). Closeness: A New Privacy Measure for Data Publishing. *IEEE Transactions on Knowledge and Data Engineering*, *22*(7), 943–956. http://doi.org/10.1109/TKDE.2009.139

Wang, Y., & Kobsa, A. (2008). Privacy-Enhancing Technologies, 352–375. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.299.1335&rep=rep1&type=pdf

Warren, S. and Brandeis, L. D. (. (2008). The Right to Privacy. . Harvard Law Review, IV(5).

Young, J. (1978). (1978). *Introductions: A Look at Privacy, chapter Privacy*. New York, John Wiley & Sons.

Yu, W. D., & Murthy, S. (2007). PPMLP: A Special Modeling Language Processor for Privacy Policies. In *2007 IEEE Symposium on Computers and Communications* (pp. 851–858). IEEE. http://doi.org/10.1109/ISCC.2007.4381555

Yu, W., Doddapaneni, S., & Murthy, S. (2006). A Privacy Assessment Approach for Serviced Oriented Architecture Application. In *2006 Second IEEE International Symposium on Service-Oriented System Engineering (SOSE'06)* (pp. 67–75). IEEE. http://doi.org/10.1109/SOSE.2006.3

Zeadally, S., & Winkler, S. (2016). Privacy Policy Analysis of Popular Web Platforms. *IEEE TECHNOLOGY AND SOCIETY MAGAZINE*, (june), 75–85.

CHAPTER 3. SURVEY ON PRIVACY AWARENESS AND ONLINE PRACTICES

This chapter deals with the survey regarding privacy awareness in online activities conducted in India in 2014. The chapter explains the issues and challenges in privacy protection and awareness, objectives of this survey, methodology, analysis of responses received, and important findings that form the directions of this research. The extensive analysis of the responses has resulted in defining and deciding the requirements for the development of a privacy protection and privacy awareness system. This chapter concludes with the discussion on the issues and challenges faced to address the issues of privacy awareness.

The survey and its few results were already presented in a conference (Paper no: Section 1.7, B-2). As, the survey was executed once and the results cannot change, this chapter include few results from the paper. However, the difference between the paper and this chapter is the way of representation and describing the results of the same survey. The updates in this chapter include objectives (section 3.4), assumptions (Section 3.6), and results and discussions (section 3.9).

3.1. INTRODUCTION

Looking at the “mass-self communication” (Kuneva, 2009), Internet users are sending billions of texts and making billions of posts, every day. Several texts and posts are rising exponentially as we have an enormous growth in the number of Internet users across the globe. The rise of Internet users in the world is shown in fig 3-1.

6.5 billion, 6.9 billion and 7.3 billion users have used the Internet in the year 2005, 2010 and 2016 respectively (ITU, 2007) (US Census Bureau, 2016). Compared to the Internet users in developed countries, the number of internet users is less in the developing countries. In 2016, 81% of the population used the internet in developed countries as compared to 40% in the developing countries.

However, penetration of the Internet is at a rapid pace. The is a substantial rise in the use of the Internet in developing countries like India. India is considered to be in the top 3 countries where digital market is gearing for handling an increase of Internet users at the rate of 41% per annum. (Srinivasan, Prasad, & Shrisha, 2013).

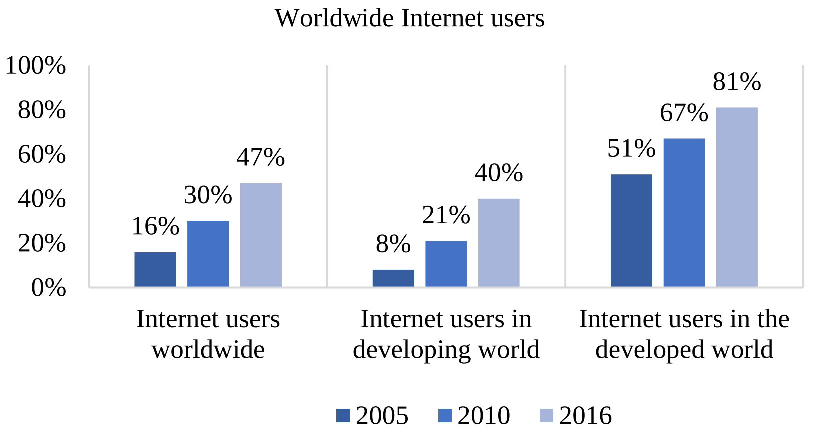


Figure 3-1 Internet users in the World (ITU, 2007) (US Census Bureau, 2016). The increase in Internet use is observed in developing countries.

It is estimated that the average age of Internet users in India will be 29 years in the year 2020 as compared to countries like Japan (48 years) and China (37 years) (Bureau of South and Central Asian Affairs, 2015). As the technological industries are being set up in India, there is a continuously increased use of the Internet in India and Internet users in the year 2021 is expected to be more than 635.8 million (The Statistics, 2017).

The demographic description of active Internet users in India as shown in fig 3-2.

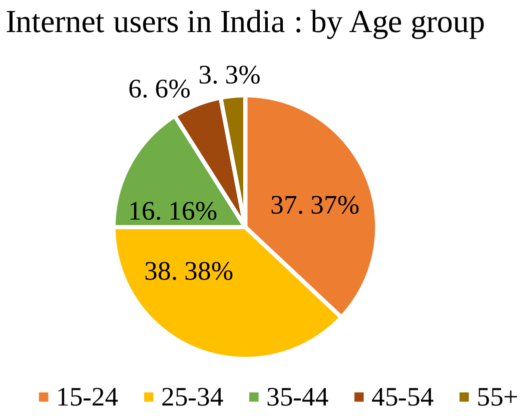


Figure 3-2 Internet user taxonomy in India (Statista, 2017).

Most of the Internet users are in the age group of 15 to 34 years who make a greater use of modern technologies for improving their daily life.

The Internet has helped in providing IT services to the businesses and the society on a large scale. This penetration has been very fast due to mobile and wireless technologies offered these days (Gnanapriya C, 2006). Importantly, the digital services are offered in several segments across the country. It mainly includes (Gnanapriya C, 2006):

1. Governments: Users access government portals to avail government schemes, view notifications, and alerts.
2. Agriculture: Users have the choice to access experts' advice and experience of harvesting, crop rotation, pest control, etc.
3. Healthcare: Users communicate with hospitals, doctors, and other health departments to know about health programs, and engage with experts on specific issues, etc.
4. Financial services: Mobile banking, transactions, knowledge about loan eligibility criteria, manage bank accounts, etc. are a few benefits of using Internet services.
5. Utilities: Users can pay bills (electricity, mobile, etc.) using the services.
6. Communications: Watching movies, and listening to music online, are a few examples of the digital services that users avail of.
7. Transportation: Users book the journey tickets online and check updates on departures and schedule.

Due to advancement in technologies, personal data of the Internet users have become a profitable commodity. Systems rather than users decide which information is displayed. Hence, in this paradigm shift, we are confronted with the privacy issues and challenges that need to be discussed.

3.2. MODERN TECHNOLOGIES AND THEIR IMPACT ON PRIVACY

Modern technology plays a vital role in developing services that will give decent living standards and create jobs in the future for citizens. India's capability in Information technology is making a significant impact at the global level (Kaka, Madgavkar, Manyika, Bughin, & Parameswaran, 2014). However, despite making steady progress towards eradicating poverty, the citizens need to be empowered

The challenges include education skills (500 million are without basic skills), lack of resources in the health sector (doctors, health workers), effective services in various domains (finance, agriculture). Also, the government schemes don't reach the indigent (Kaka et al., 2014).

In India, almost 33% mobile subscribers have a smartphone for performing daily tasks that make their life easier by utilizing online services (almost 30 applications per user) (Puru Naidu & Ranjeet Rane, 2017). There has been an exponential growth of the Smartphone market, leading to a proliferation of applications for users. In promoting the applications, the business model emphasizes on an extensive collection of user information.

The challenge is to know how many users are aware of the type of information collected, what happens once collected, for how long it will be retained, who has the access to it, etc.

A collection of users' information at the service providers' end is done in mainly two ways. One, it is voluntarily given by the user (by means of filling forms) or second, without the user's knowledge (by means of analysis of browser information, IP packets, search engine queries, online patterns, etc.) Sometimes embedding Flash Programs and JavaScript in web pages, useful information is gathered without informing the users. The latter way of the collection encompasses information about user devices, operating systems, browser details, geographical location, history of web pages visited, etc.

However, users believe that widespread use of the online shared information is used for service enhancement, advertising, etc. To go beyond this, such information is actually used for unspecified purposes like to target patient participation in disease awareness programs(Mantelero, 2016), personalization of user interest (Timothy MoreyTheodore "Theo" ForbathAllison Schoop, 2015), personalize searching (Hannak, Soeller, Lazer, Mislove, & Wilson, 2014), Identity Theft(Javelin, 2016), calculating financial credit score (Katie Lobosco, 2013), surveillance from government (Jonathan Mayer, 2013).

The user Information privacy management or awareness completely depends on the information management policy of service provider or organization. Ideally, the user data should be used and processed for the said purposes. However, in Big Data environment the data are extracted easily by violating the principles of privacy (Jayasingh, Patra, & Mahesh, 2016). Hence, there is need to understand the issues of privacy for the development of a privacy awareness method, or tool (Cardenas, Manadhata, & Rajan, 2013).

According to an analysis of MGI report, 12 technologies are promising to enhance users' decision ability and empower them by digitizing their life and work. This includes digital identity and cashless payments, mobile internet, automation of information, etc.

Privacy and privacy issues in India are thought-provoking and challenging too. Internet privacy is a critical area of concern that must be handled by the developing countries like India. As mentioned in chapter 2, the issues of digital identity (AADHAR card) in India are rising as Indian citizens' information is not being treated sensitively and the user has no control over it (Jyoti Panday, 2017) (India Today Tech, 2017). Similarly, the issues of unauthorized access to personal information are worsening. One of the examples include malpractices from online shopping websites dealing with enormous amounts of users' information[44].

Consider an incident of sharing and using WhatsApp users' data by Facebook, the German Commissioner ordered Facebook to delete collected data from German WhatsApp users (Mike Isaac & Mark Scott, 2016). However, for the same incident, in India, the users are asked to stop using WhatsApp's service to prevent data sharing to Facebook.

Besides having basic privacy rules mentioned in IT Act 2008 (The Gazette of India, 2009), these rules are not explained in an exhaustive manner and not firmly enforced (Horbach, 2017). The uncertainty in the enforcement of such rules will lead to demand for a new formation of privacy law like in the EU or US (Puru Naidu & Ranjeet Rane, 2017). The new legislation on the "Right to Privacy" is an amendment to the IT Act that embraces a provision for the protection of information privacy. This shield is found to be deficient in ensuring the protection of citizen's privacy.

The differences in the laws and judgments on the issues of personal data sharing, clearly indicate that there is a need for major changes in privacy principles and a revision in the existing privacy law. So, security and privacy concerns persist and need immediate attention to prevent loss/ misuse of users' personal information.

3.3. RELATED WORK

There have been several research contributions to understand users' perception about privacy. Some of the contributions are mentioned in this section.

To identify online practices of information handling and related privacy issues, an extensive survey of 116 complaints was carried by the Federal Trade Commission (FTC) ("CLIP- Center on Law and Information Policy," 2014). The survey presented the list of petitions filed by victims and their classification as shown in fig 3-3. It illustrates four main categories as Unauthorized disclosure of personal information, a Surreptitious collection of data, inadequate, and Wrongful retention of personal information.

The complaints consist of issues of personal information sharing, unauthorized access and sale, no notification, no consent, unnecessary extensive information collection,

[44] https://www.snapdeal.com/

inefficient information protection system, incomplete privacy policy, and retention of information collection for a longer time.

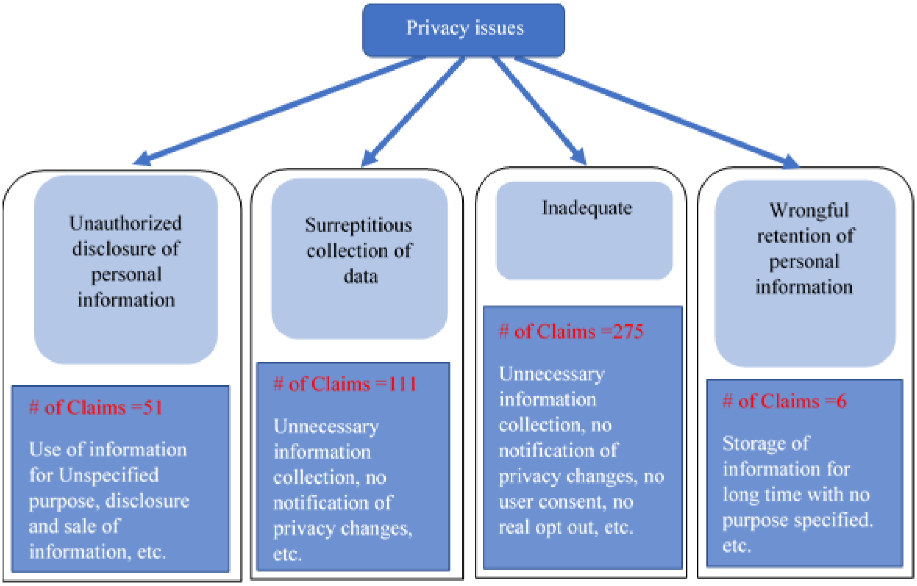

Privacy issues
Unauthorized disclosure of personal information
# of Claims =51
Use of information for Unspecified purpose, disclosure and sale of information, etc.
Surreptitious collection of data
# of Claims =111
Unnecessary information collection, no notification of privacy changes, etc.
Inadequate
# of Claims =275
Unnecessary information collection, no notification of privacy changes, no user consent, no real opt out, etc.
Wrongful retention of personal information
# of Claims =6
Storage of information for long time with no purpose specified. etc.

Figure 3-3 Summary of privacy issues at FTC Enforcement Actions ("CLIP- Center on Law and Information Policy," 2014). The complaints are separated into four categories.

In the United States of America, a group survey was carried out to check the compliance of 35 frequently used social health websites with the Fair Information Practices (FIP) (Savla & Martino, 2012). An important finding was, that the user's privacy is at a considerable risk if the privacy policy does not comply with the FIP principles. The privacy policies of healthcare service providers do not support informed decisions.

An aim of the survey carried out by Cheung (Cheung, Chiu, & Lee, 2011) was to know the reasons for sharing personal information by the students on Social websites. The reasons that force users to use social media are interpersonal connectivity, self-discovery, social enhancement, etc.

Another study was carried out to understand users' knowledge of the privacy policy. For this study, social websites were selected (Facebook, Twitter, LinkedIn, etc.). The important finding in this study was that the Internet users do not completely understand what they have agreed to while registration (Zeadally & Winkler, 2016).

Another survey exemplifies the high privacy concerns in countries like Saudi Arabia (Alsagri & Alaboodi, 2015), UK (Khan & Hasan, 2016). The study concluded that the users' online behavior (excessive use of Facebook, Snapchat, etc.) is not in proportion to the highest privacy concerns (collection and use of information), despite privacy concerns. This study needs to be extended to understand users' view on service providers, privacy policies, etc.

A comprehensive nationwide study was done on the privacy and security habits of Indian Internet users (Kumaraguru, 2012). 10,427 responses were collected from various cities in India. The Study showed that the participants were more concerned about their privacy. An important finding was that the password is considered most private information as compared to religion, mobile phone number, and health-related issues. Another finding was that about 40% of the participants would never save/share personal information in/ through emails. Privacy seems to be the primary reason for this behavior.

In a study, 400 users of social networking sites were analyzed to understand their awareness, behavior, and attitude based on three attributes namely- age, gender, and profession (Dhawan, Singh, & Goel, 2014). Important findings of this study include: Women are more serious towards privacy than men, teenagers spend more time on social websites, but the teenagers are less concerned about privacy than adults, etc. Such studies need to be extended to understand users' knowledge about privacy, their online practices, threats, and challenges whenever they are online.

After understanding the related work presented in this section, we conclude that most of the research (privacy issues, surveys on privacy protection and management) is focused on the developed countries (Zeadally & Winkler, 2016) (Savla & Martino, 2012) (Futuresight, 2011) etc. Also looking at India as a major country in various dimensions, and the background of privacy issues, we concluded that there is little work done as far as research is concerned about privacy issues in India (Kumaraguru, 2012) (Abdullah, Gregory, & Beyah, 2008) leading to insufficient observational data in India, making it vital and necessary to focus research on the users' privacy knowledge and awareness in India. Therefore, this chapter provides the complete details about the survey conducted in 2014 to understand the motivation and mitigation factors for improving personal information privacy.

The survey results were already presented in a conference (Dhotre & Olesen, 2015) as a part of knowledge dissemination. However, the detailed analysis and outcome of the survey are presented in this chapter.

3.4. OBJECTIVE AND DESIGN OF THE SURVEY

Quantitative analysis is a useful method to gain information and insights from Internet users about privacy awareness. Each user's opinion and feedback is utilized for intended purposes. This could be one of the ways to understand the requirements and what is it that satisfies the users in the context of privacy awareness.

Keeping the issues of privacy, we conducted a survey in India, which has several different objectives. Few of them are listed below:

- To learn more about Internet users: This survey is going to help us to know users' online practices and activities, knowledge about privacy and concerned law, etc. This will come in handy when understanding different segments of users and their online practices.

- To understand issues and challenges: To get better insights of privacy issues and challenges in privacy this survey is useful. This survey will give more chances to identify the possible threats to the users' information privacy.

- To receive suggestions regarding effective privacy awareness mechanism: This survey is indirectly asking Internet users to contribute their suggestions, opinions on privacy awareness and protection, to help to frame the requirements to form a system or tool for privacy awareness or protection. The analysis of suggestions and feedback help to identify the motivation and mitigating factors for the protection of personal information privacy.

Considering these issues and evaluation of state of the art presented, it is important to understand different users' knowledge of privacy, privacy law, novel issues, and level of awareness in the Indian subcontinent.

There are several purposes for conducting the survey in India. First, important thing is to bring privacy issues of India on the privacy canvas at world level. Therefore, the aim of this study is to know various concerns of Indian Internet users while they are online, users' perspective on privacy and the need for privacy awareness.

Furthermore, it is important to understand what is missing in the present privacy solutions that adhere to the principles of privacy laws in India as well as other countries. Last, but not the least, the purpose is to know the users' attentiveness on certain things like a privacy policy, and responsibilities of service providers.

Considering the viewpoints mentioned above, and to create more interest among the respondents in filling this survey, questions were designed and divided into various sections. So, the sections are:

- Current practices: The set of the questions in this section was focused to know the user's preferred online services, the location of access, etc.

- Awareness about privacy: This section contains the questions to assess the users' knowledge of privacy. The answers to the questions in this section would provide the possible threats or motivational factors for privacy protection or awareness. The questions were designed to understand users' knowledge of privacy risks, personal information management, government/state's duty, etc.

- Seriousness when online: This section describes the list of questions to know the how serious is the user whenever they are online.

- Behavior: The section focused on the users' online behavior in different the context of various scenarios like information sharing, responding to bulk messages, reactions to online tracking.

- Privacy policies and law: Considering the privacy policies and concerned law, the questions were listed in this section. The questions were asked to know policy reading frequency, view on privacy policy, privacy law

- Organization/ Service providers: While interaction with a service provider, it is important to know what the user thinks about them and their practices on the personal information management. This section lists the question about users view on their organizations or service providers considering privacy.

- A user: in the context of medical and Health information: It was also important to know the users' activities on a cell phone like security, apps management, location disclosure. Also, this section of the survey would like to know the users' view on medical information when they are online.

The questions in this survey have combinations of several types of questions (multiple choice, yes/no, five-point scaling, optional questions, etc.). This will help the participants to break monotonous nature while giving feedback on the survey, which is one of the principles of this survey.

3.5. DESIGN PRINCIPLES OF THE SURVEY

According to principles of the survey (Susan Farrell, 2016), the survey mentioned in this chapter is simple to use. This survey has questions that can be responded using radio buttons, checkboxes, etc.

The reason for selecting quantitative survey is that it counts results gathered from respondents (like how many respondents do this vs do that). The respondents were selected randomly using standard methods of quantitative survey. The Same set of questions were asked to the respondents which are another principle followed in this survey.

Once the responses were statistically analyzed, quantitative survey ensured that findings from the survey are significant. Another important characteristic is that the survey findings are representative of the entire population.

So, in short, the advantages of the quantitative survey over qualitative survey are it's cheaper, random selection of participants, counts the result, etc.) (Sherrie Mersdorf, 2016).

3.6. ASSUMPTIONS

Following are some assumptions made when the survey was conducted.

- The participants were informed about the purpose and given the required information about this survey.
- The participants have filled this information voluntarily and without any coercion.
- It is assumed that the participants were honest and have been eager to participate in the survey.
- Most of the participants could be contacted by email or through social media.
- The email addresses provided by the participants were accurate.
- A participant responded to the survey only once.

3.7. SURVEY METHODOLOGY

For the execution of this survey, the online method was used to send questionnaires and receive the responses. The questions were prepared and sent to the participants randomly by means of email, social media, etc. The survey included a few interviews given by those who preferred in-person interactions.

Online questionnaire and interviews were the two methods used in this survey. Instead of using the traditional offline method of survey, the questionnaires were administered online. Using Google Forms (Google service), the questions were prepared and link to questions was sent to all possible users in India via email. The link was available at http://goo.gl/forms/ctCImSfWHH. The contacts were selected from the researcher's contact list.

The aim of administering the questions online was to save time spent to reach users offline and handover the hardcopy of questions Since, it was important to contact as many Internet users as possible. The link had been posted on the social networking sites (Facebook, Google). Using the groups on Facebook and Google, the questions were posted in just one click. Similarly, the invited user had an opportunity to share the same link to his/her contacts to garner more responses from other users to whom the researcher may or may not know.

While conducting this survey, attention was given to both users from the public sector and from the private sector. The users from the public sector were identified from the social websites with whom this researcher is connected to. To get the responses from private sector users, the survey was carried out in private organizations like Education, Banking, IT companies, Owned organizations, Government offices, etc.

A national level survey mentioned in this thesis was conducted from October 10, 2014, to November 05, 2014 across various parts of India. A total of 950 Internet users were invited randomly to participate in the survey. The questionnaires were sent out using electronic mail and social media. Effective responses were received from 297 users (Male- 205, Female- 81 and non-specified- 11). Hence, the response rate of this survey is 31.26%.

To get actual and correct feedback, having responses of participants from diverse background is very important. Hence, this survey has taken care to ensure that the participation should come from various age groups, professions, and age.

3.8. DATA ANALYSIS TOOL

Using online survey the responses were received in a Google spreadsheet, a crucial step to do an analysis of responses. Very few responses were invalidated from the analysis because the questions were either partially answered or were left blank. Hence, the final analysis was done on 297 responses using a business analytics tool called Tableau[45]. 15 days' free version of Tableau was used followed by its unlimited public version.

There are a couple of reasons to select this tool. Firstly, this tool is easy to download and use too. Second, it doesn't require any type of programming to perform the analysis. Using drag and drop process on the spreadsheet, the results were obtained with ease. The best feature of this tool is the dashboard which allows the reader to see the results with an effective GUI.

3.9. RESULTS AND DISCUSSIONS

This is the most important section of this chapter. The responses were assessed and classified into various categories for easy understanding. While evaluation, the focus was given to certain things like user awareness, privacy knowledge, online practices, threats to user information privacy. Also, motivation and mitigating factors were identified that has helped to build a privacy protection or awareness mechanism.

[45] https://www.tableausoftware.com

3.9.1. GENERAL INFORMATION ABOUT RESPONDENTS

The survey received responses from the Indian users living in various parts of the country. The general observation is represented in table 2. It represents the distribution of respondents in terms of gender, skill level, profession, and age.

Total Responses: 297		
Gender	**%**	**Number of records**
Male	71.68 %	205
Female	28.32 %	81
Not specified	3.7 %	11
Skill level	**%**	**Number of records**
A beginner	15.03%	43
An Intermediate User	32.52%	93
A legitimately experienced user	41.61%	119
A very experienced user	11.54%	33
Not specified	3.03%	9
Profession	**%**	**Number of records**
Full Time employed	36.36%	104
Part-time employed	0.35%	1
Student	55%	156
Retired	0.35%	1
Student + Part time employed	0.35%	1
Self Employed	1.75%	5
Working Student	5.59%	16
Other (please specify)	0.35	1
Not specified	4.04%	12
Age	**%**	**Number of records**
Under 16	0%	0
Between 16 and 25	75.87%	217
Between 26 and 35	17.13%	49
Between 36 and 45	4.90%	14
Between 46 and 55	1.05%	3
Between 56 and 65	0.70%	2
Above 65	0%	0
Not specified	4.04%	12

Table 2 General analysis of the respondents (Gender, skill, profession, and age)

The survey received maximum responses from male participants as compared to female participants (72% vs 28%). 75% of the user's computer skill level was either of intermediate or had just attained little experience and only 27% of participants were either beginner.

Moreover, the participants of this survey found to be from students group (55%) or employee group (37%). As discussed earlier in this chapter (fig 3-2), most of the Internet users in India are students or employees, so the maximum response (93%) received whose age was between 16 to 35 as compared to other age groups (93% vs 7%). The maximum responses received were from Indian male citizens in the age group of 16-35, who are employed and had some experience.

3.9.2. FINDINGS ON ONLINE PRACTICES, KNOWLEDGE, THREATS

This section describes the analysis of users' feedback on various parameters like online practices, knowledge of privacy, and threats

- Online practices

Access to Internet services	Q5: From where do you access services/utilities provided by service providers, please choose:		Q6: Which of the internet facilities you use from computer/ laptop/ mobile?	
	Location		Type of services	
	213 records (71.71%)	84 records (28%)	221 Records (80.95%)	76 records (28%)
	home, School/University or office/work locations	Other locations	Social websites, online shopping/ reservation, search engines	Other services

Table 3 Accessing Internet services

Table 3 represents the users' online practices like the location and type of services. Almost 72% of total participants of this survey use the Internet from home, university or office location. Also, basic facilities on the Internet (social websites, shopping, reservation of travel tickets, searching, etc.) are used on a large scale (81%).

- Knowledge

Looking at the responses (fig 3-4) on the question related to information collection (Q8), the participants had various views on information being recorded when a user requested for a web page.

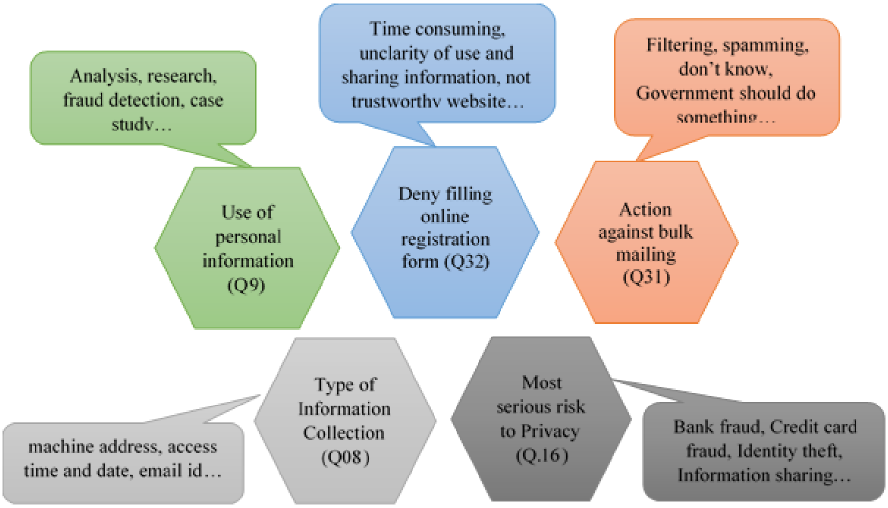

Analysis, research, fraud detection, case study…
Time consuming, unclarity of use and sharing information, not trustworthy website…
Filtering, spamming, don't know, Government should do something…
Use of personal information (Q9)
Deny filling online registration form (Q32)
Action against bulk mailing (Q31)
machine address, access time and date, email id…
Type of Information Collection (Q08)
Most serious risk to Privacy (Q.16)
Bank fraud, Credit card fraud, Identity theft, Information sharing…

Figure 3-4 Understanding users' knowledge. Important five parameters were identified to know users' knowledge.

The common answer was machine address, access time and date, email id and location as a set of information. However, a research represented EFF[46] reveals a huge amount of information is being captured when a user is utilizing online services. So, it's clear that there is a significant discrepancy in the amount of information revealed by the respondents and actual information being said to be collected for the research. Hence, this indicates that the users have limited knowledge about information collection.

In response to the questions on most important risk to privacy (Q16), it is observed that participants have limited knowledge on privacy risk. The participants are more worried about financial information risk as compared to personal information. This indicates that the users are not fully aware of "the value of user information" mentioned in earlier research (Cochran, 2013), (OECD 2013).

The difference in the level of understanding in the use of personal information is observed during analysis of a question (Q9). As per the analysis of the responses, the information is used not only for analysis of user information but also used for fraud detection, case study and in research too. However, the use of information is enormous for many purposes like personalizing searching (Hannak, Soeller, Lazer, Mislove, & Wilson, 2014), Identity Theft (Javelin, 2016), calculating financial credit score (Katie Lobosco, 2013), and surveillance by the government (Jonathan Mayer, 2013).].

[46] https://panopticlick.eff.org/

However, the users have an acceptable level of knowledge when it comes to refusing to fill registration forms or handle spam emails.

Knowledge	Q.25.Do you know that your behavior on Internet, like your browsing, what you search and online purchases, dealing with services, etc. can be monitored by…			Q.26.Do you take any steps/actions to limit tracking/ monitoring of Internet/online activities?			Q.10 Do you know any national institutions/Law that will help the respondents to deal with user's privacy along with their identifiable information protection from the wrong way of information collection, its utilization (use) and sharing options?		
	Yes	No	Don't Know	Yes	No	Don't know	Yes	No	Don't know
	85%	12%	03%	33%	57%	2%	57%	39%	4%

Table 4 Understanding knowledge of respondents

As shown in Table 4 and at the outset, 85% of participants were aware that their online activities are monitored and tracked by websites or other companies (Q.25). However, it would be interesting to know their action (s) to stop tracking or monitoring activities to protect their own privacy.

Despite the availability of many blocking tools (many of them are free), the participants (57%) do not take steps to limit or block users' online activities. Ignorance about privacy protection tools could be one of the reasons that the participants are not able to block monitoring activities.

In most countries, national laws provide rules and regulations for the protection of user privacy. Also, the guidelines are given in the case of privacy breach. Only 57% of the participants are aware of such institutions or laws that can help to resolve the issues related to the unintended use of personal information. Considering a rise in the issues of privacy infringement in a country like India, the knowledge of privacy laws among Indian users need to be enhanced.

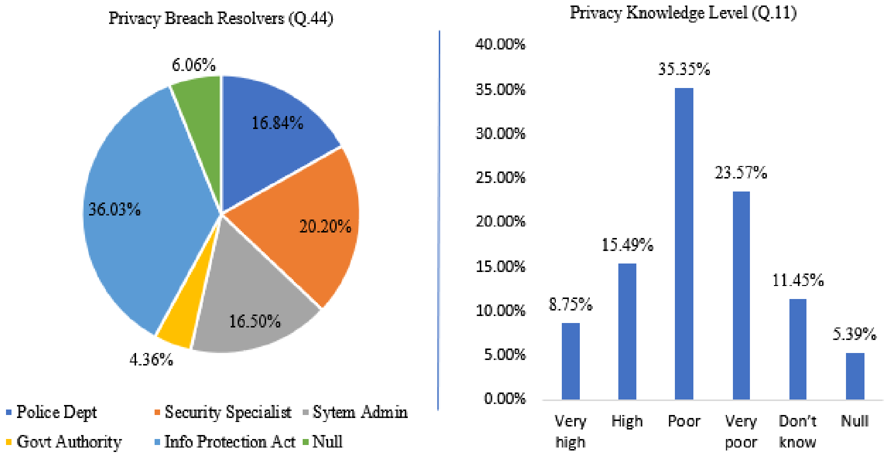


Figure 3-5 Users' knowledge – privacy breach resolvers and privacy right's level. The first part represents the various departments where people contact in case of privacy breach. The second part represents various levels of users' privacy knowledge.

Question number 44 was related to privacy breach resolvers (Q44). If there are privacy issues or privacy leaks, the victim should approach the right place/authority for filing a complaint. In this survey, the participants have different views and possibly do not know the right place to put forth their grouse. Most of the participants will refer a security specialist or read the protection act (20% and 36%, respectively) for further actions, rather than asking the system administrator or the police department. Fig 3-5 illustrates the participants' knowledge of privacy rights. Hardly 25% of participants said that they have (Very high or high) knowledge. But overall, it is abundantly clear that 2/3rd (75%) agreed their knowledge is inadequate.

- Threats

Threat	Q13.Protecting the personal information of the respondents will be the most critical issues facing our country in the next ten years.			Q49.Does your agency disclose/uses individual AADHAR Card Number (or Passport Number) or disclose user records that contain AADHAR Card Number (or Passport Number)?		
	Yes	No	Don't Know	Yes	No	Don't know
	91%	3%	05%	10%	46%	38%

Table 5 Understanding threats (1)

Looking at the responses (table 5), the participants do consider protection of personal information as the most critical issue to be faced in the next 10 years (Q13). Not only at present, but in the days to come too, 251 participants (91%) agreed on the importance of protection of personal information. It is a threat for the participants when it comes to the sharing of AADHAR (unique identity number) by the service providers. Because 38% participants are not sure about the practices of service providers about the records that contain AADHAR card details. This indicates that the service providers should clearly specify their policy in the easiest way possible.

Threats	Q20. Without a warrant, the law administration/national security agencies collects it for general surveillance/investigation purposes.			Q21. Marketing companies use your personally identifiable information to analyze your opinions (For example likes/Yes, dislikes/No).			Q22. Organizations/companies use the personal identifiable information to determine a possible list of a job for you.		
	Very nervous	Somewhat nervous	Not nervous	Very nervous	Somewhat nervous	Not nervous	Very nervous	Somewhat nervous	Not nervous
	24%	55%	20%	30%	38%	30%	14%	44%	40%

Table 6 Understanding threats (2)

From table 6, we observe that the respondents are not much worried (almost 75% are less worried) about the collection of personal information by law administration/national security agencies for intended purposes like investigation or surveillance. Therefore, there is the need to spread awareness about information collection and its use.

The participants do not realize the urgent need to understand the strategies of companies on the analysis of users' opinion. Certainly, users are interested in sharing their views (like or dislike) whenever they are online. The participants are less nervous (68%) about the companies' analysis of users' opinions, their likes and dislikes to build marketing strategies based on personal information. Consequently, this raises questions about transparency and notification from service providers about their business models. Interestingly, if the companies clearly share the purpose of data collection to the users, then the users will share their information without much worry. As per the analysis of the survey, only 14% participants are worried to share information in the context of job offerings by companies. It means that the rest of the respondents (almost 84%) are not worried/nervous to share information if the company offers or tailors a suitable job for them.

3.9.3. FINDINGS ON PRIVACY POLICY

In this section, the analysis of responses and the findings on privacy policy issues are presented in fig 3-6.

Due to limited privacy knowledge, hardly 8% participants of the survey read a privacy policy on a regular basis (Q39). So, it would be interesting to know the reasons for neglecting to look into privacy policies which are an important channel of communication between users and service providers. At the outset, the obvious reasons for this neglect to read and understand the privacy policies include, it is too descriptive, difficult to understand and lengthy (according to 88% participant). Hence, it is not only challenging to understand these obstacles but also there is a need to come up with solutions that help to understand and do not discourage the users to read the privacy policy.

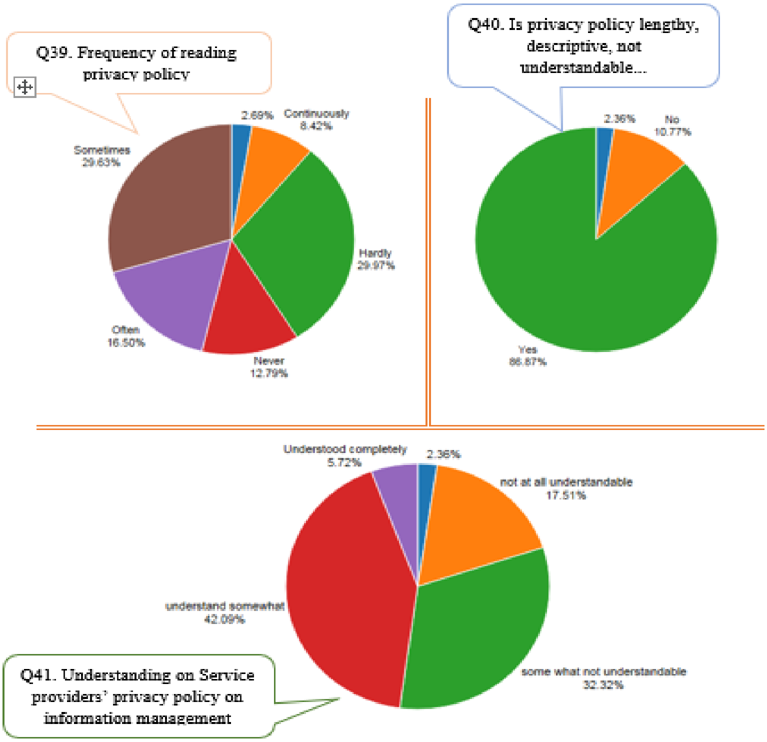


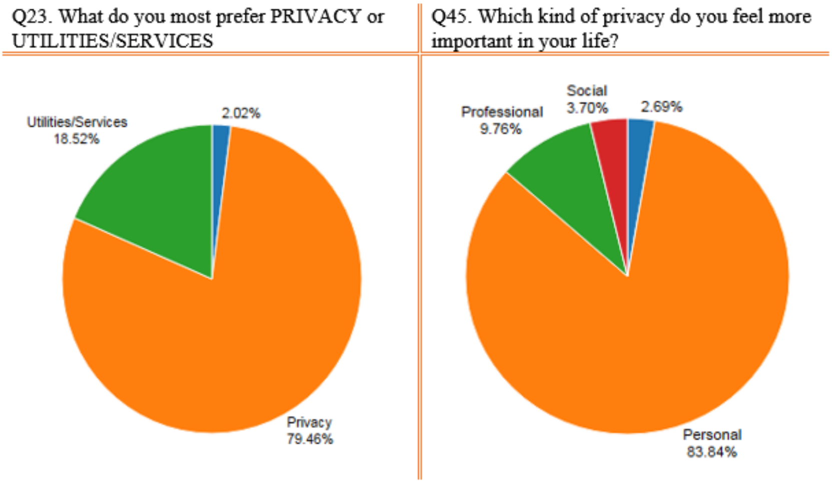
Figure 3-6 Issues on privacy policy. Important issues are the frequency of reading the privacy policy, users' view and its understanding.

Similarly, only 6% of the respondents understand personal information management (use, share, etc.) mentioned in the privacy policies. Therefore, it is imperative that research addresses this critical issue. The contribution mentioned in this thesis is an attempt to resolve this issue.

3.9.4. FINDINGS OF MOTIVATIONS FOR PRIVACY PROTECTION SYSTEM

This section discusses motivational factors identified after analysis of the responses. As shown in Fig 3-7, 79% of the total respondents realize the importance of privacy over the convenience offered by utilities or services provided by the service providers/companies. Keeping the risks in mind, 19% of the respondents favored the convenience of utilities/services. The rest of the participants are not interested in using the utilities/services if their personal information privacy is breached. It denotes a strong tradeoff between privacy and utilities and therefore the necessity to address the privacy issues.

This indicates the anxiety to protect the privacy of the personal life of the respondents as compared to professional. Almost 84% of the respondents feel that the importance of personal privacy far outweighs the importance of professional or social life privacy.



Q23. What do you most prefer PRIVACY or UTILITIES/SERVICES

Q45. Which kind of privacy do you feel more important in your life?

Figure 3-7 Motivation factors for privacy protection. Users believe that the privacy is preferable over services and personal privacy is more important.

97% of the total respondents want changes/major revision in the existing laws that protect user information which will protect their information privacy so that they can access online services hassle-free (Q33). This also implies the risks involved in today's internet, or the issues of internet security that the users feel need to be addressed.

Motivation	Q33. There should be new laws to protect user privacy on the Internet.		35. Content/service providers have the right to share/resell information about their users to other agencies/companies.		Q52. Would you like to disable the location disclosure option on your cell phone	
	Agree	Disagree	Agree	Disagree	Agree	Disagree
	97%	03%	27%	71%	68%	28%

Table 7 Understanding motivation factors (I)

Considering limited knowledge of the respondents on privacy (table 7), the respondents strongly disagree (71%) on the rights of service providers for sharing or reselling their information to other agencies. 71% of the respondents disagree with the information loss and sharing with either known or unknown company.

68% of the respondents are concerned about disclosing their location information and hence turned off the location tracking options from their cell phones considering the access to their other information (Q52).

The respondents are completely dissatisfied with existing/present privacy system. They believe that there is a need for an improvement in the privacy systems of the service providers. 82% of the respondents have rated their experience with the present privacy system as “neither good nor bad”, or “bad”, or “very bad”.

3.9.5. FINDINGS OF MITIGATIONS FOR PRIVACY PROTECTION SYSTEM

In this section, the responses that suggest privacy protection are discussed.

The users are keen to share their information if, and only if, the service provider informs them about the collection methodology and use of collected information. 197 out of 297 (66%) participants strongly believe in sharing or providing their critical information to a service provider, if and only if, they (website/service provider) reveal what sort of information will be gathered and how the information will be utilized for the users’ benefit. The rest of the participants (34%) are not able to share their critical information to website/service providers for some reason.

Mitigation	…establishment of a personal trust manager … (Q34)		…A user should have complete control over information … (Q36)		…anonymous when visiting sites on the Internet… (Q37)		…service provider should notify you when they deal personal information? (Q47)		…mechanism to rank the service provider according to your experience … (Q48)	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
	88%	10%	92%	06%	74%	23%	87%	10%	88%	8%

Table 8 Understanding mitigation factors (II)

The users are looking for the establishment of coinciding between the service provider and themselves. From table 8, we observe that majority (88%) of the respondents wants a "personal trust manager" in future. While the user is online, the preferences and the experience of the user must be observed and monitored by the personal trust manager.

Respondents believe that they should have the control over their personal information which has been collected by the service provider. A strong assertion is made by 92% of the respondents for providing user control of the critical and essential information that a service provider possesses. It implies a need for a mechanism where the user can see how the service provider collects, uses and shares personal information in detail.

From the security (privacy) point of view, 74% of the respondents want to anonymously visit and browse the website/the Internet. Indians feel that substantial risk is involved when they are online.

68% of the respondents say that communication over the internet must not be accessible to any third party. It indicates that the respondents are increasingly concerned about privacy and aware of the privacy concerned issues.

The respondents show an inclination towards more openness between the service providers and themselves. 87% of the respondents believe that the service provider must notify the users when they access, sell or share users' information.

The respondent users expect an establishment of a strong trust between the service provider and themselves. 88% of Respondents expect a mechanism that should tell them the rank/level of the service provider based on the experience of other users. They believe such mechanism will help them to know how secure a specific site is.

3.10. REFERENCES

Abdullah, K., Gregory, C., & Beyah, R. (2008). A visualization framework for self-monitoring of web-based information disclosure. *IEEE International Conference on Communications,* 1700–1707. http://doi.org/10.1109/ICC.2008.328

Alsagri, H. S., & Alaboodi, S. S. (2015). Privacy awareness of online social networking in Saudi Arabia. *2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, CyberSA 2015*. http://doi.org/10.1109/CyberSA.2015.7166111

Bureau of South and Central Asian Affairs. (2015). U.S. Relations With India. Retrieved May 29, 2017, from https://www.state.gov/r/pa/ei/bgn/3454.htm

Cardenas, A. a., Manadhata, P. K., & Rajan, S. P. (2013). Big Data Analytics for Security. *IEEE Security & Privacy*, *11*(6), 74–76. http://doi.org/10.1109/MSP.2013.138

Cheung, C. M. K., Chiu, P.-Y., & Lee, M. K. O. (2011). Online social networks: Why do students use facebook? *Computers in Human Behavior*, *27*(4), 1337–1343. http://doi.org/10.1016/j.chb.2010.07.028

CLIP- Center on Law and Information Policy. (2014). Retrieved from https://www.fordham.edu/download/downloads/id/1867/privacy_enforcement_actions.pdf

Dhawan, S., Singh, K., & Goel, S. (2014). Impact of privacy attitude, concern and awareness on use of online social networking. *Proceedings of the 5th International Conference on Confluence 2014: The Next Generation Information Technology Summit*, 14–17. http://doi.org/10.1109/CONFLUENCE.2014.6949226

Dhotre, P. S., & Olesen, H. (2015). A Survey of Privacy Awareness and Current Online Practices of Indian Users. In *Proceedings of WWRF Meeting 34, Santa Clara, CA, USA, Apr. 2015* (p. 10). WWRF. Retrieved from http://vbn.aau.dk/en/publications/a-survey-of-privacy-awareness-and-current-online-practices-of-indian-users(92c00b4f-a720-45b8-b3cb-9cfbffc7d4bc).html

Futuresight. (2011). Futuresight:, “User perspectives on mobile privacy, Summary of research findings.” Retrieved from http://www.gsma.com/publicpolicy/wp-content/uploads/2012/03/futuresightuserperspectivesonuserprivacy.pdf

Gnanapriya C, P. G. (2006). Need and Relevance of Mobile Based Information Services in Emerging Markets - India. Retrieved July 23, 2017, from https://www.w3.org/2006/07/MWI-EC/PC/paper_infosys.html

Hannak, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring Price Discrimination and Steering on E-commerce Web Sites. In *IMC '14 Proceedings of the 2014 Conference on Internet Measurement Conference* (pp. 305–318). Vancouver, BC, Canada: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2663744

Horbach, L. (2017). Privacy and Data Protection in Medicine. *Cryptography*, (2012), 228–232. http://doi.org/10.1007/3-540-39466-4_16

India Today Tech. (2017). Aadhaar data of 130 millions leaked from govt websites: Report : News, News - India Today. Retrieved July 12, 2017, from http://indiatoday.intoday.in/technology/story/aadhaar-data-of-130-millions-bank-account-details-leaked-from-govt-websites-report/1/943632.html

Javelin. (2016). Identity Fraud Hits Record High with 15.4 Million U.S. Victims in 2016, Up 16 Percent According to New Javelin Strategy &amp; Research Study | Javelin. Retrieved May 29, 2017, from https://www.javelinstrategy.com/press-release/identity-fraud-hits-record-high-154-million-us-victims-2016-16-percent-according-new

Jayasingh, B. B., Patra, M. R., & Mahesh, D. B. (2016). Security Issues and Challenges of Big Data Analytics, 204–208.

Jonathan Mayer. (2013). How the NSA Piggy-Backs on Third-Party Trackers | Center for Internet and Society. Retrieved May 29, 2017, from http://cyberlaw.stanford.edu/publications/how-nsa-piggy-backs-third-party-trackers

Jyoti Panday. (2017). Aadhaar: Ushering in a Commercialized Era of Surveillance in India | Electronic Frontier Foundation. Retrieved July 12, 2017, from https://www.eff.org/deeplinks/2017/05/aadhaar-ushering-commercialized-era-surveillance-india

Kaka, N., Madgavkar, A., Manyika, J., Bughin, J., & Parameswaran, P. (2014). *India ' s technology opportunity : Transforming work , empowering people*.

Katie Lobosco. (2013). Facebook friends could change your credit score - Aug. 26, 2013. Retrieved May 29, 2017, from http://money.cnn.com/2013/08/26/technology/social/facebook-credit-score/

Khan, R., & Hasan, R. (2016). The Story of Naive Alice: Behavioral Analysis of Susceptible Internet Users. *Proceedings - International Computer Software and Applications Conference*, *1*, 390–395. http://doi.org/10.1109/COMPSAC.2016.206

Kumaraguru, P. (2012). Privacy in India : Attitudes and Awareness V 2 . 0.

Kuneva, M. (2009). Keynote Speech on Roundtable on Online Data Collection, Targeting and Profiling. *European Consumer Commissioner*, (March).

Mantelero, A. (2016). Draft Guidelines on the Protection of Individuals With Regard To the Processing of Personal Data in a World of Big Data. *Council of Europe*. Retrieved from https://rm.coe.int/16806ebe7a

Mike Isaac, & Mark Scott. (2016). Relaxing Privacy Vow, WhatsApp Will Share Some Data With Facebook - The New York Times. Retrieved July 23, 2017, from https://www.nytimes.com/2016/08/26/technology/relaxing-privacy-vow-whatsapp-to-share-some-data-with-facebook.html

Puru Naidu, & Ranjeet Rane. (2017). To share personal data or not? It’s time India came clean on its privacy laws | tech | Hindustan Times. Retrieved May 29, 2017, from http://www.hindustantimes.com/tech/to-share-personal-data-or-not-it-s-time-india-came-clean-on-its-privacy-laws/story-DpQ3mqdTi8unTC9wDHYSoM.html

Savla, P., & Martino, L. D. (2012). Content analysis of privacy policies for health social networks. *Proceedings - 2012 IEEE International Symposium on Policies for Distributed Systems and Networks, POLICY 2012*, 94–101. http://doi.org/10.1109/POLICY.2012.20

Sherrie Mersdorf. (2016). Qualitative vs. Quantitative Research Methods - Blog. Retrieved July 12, 2017, from https://blog.cvent.com/events/feedback-surveys/qualitative-vs-quantitative-research-methods/

Srinivasan, R., Prasad, V. A., & Shrisha, S. (2013). The impact of technology on the buying behaviour in an online retail environment in India. *2013 Proceedings of PICMET 2013: Technology Management in the IT-Driven Services*, 569–575.

Statista. (2017). • *Age distribution of internet users in India 2013 | Statistic*. India. Retrieved from https://www.statista.com/statistics/272394/age-distribution-of-internet-users-in-india/

Susan Farrell. (2016). 28 Tips for Creating Great Qualitative Surveys. Retrieved May 29, 2017, from https://www.nngroup.com/articles/qualitative-surveys/

The Gazette of India. (2009). *Information Technology Act 2008,*. Retrieved from http://meity.gov.in/sites/upload_files/dit/files/downloads/itact2000/it_amendment_act2008.pdf

The Statistics. (2017). India: number of internet users 2021. Retrieved May 29, 2017, from https://www.statista.com/statistics/255146/number-of-internet-users-in-india/

Timothy MoreyTheodore "Theo" ForbathAllison Schoop. (2015). Customer Data: Designing for Transparency and Trust. Retrieved May 29, 2017, from https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust

US Census Bureau, D. I. S. (2016). International Programs, International Data Base. Retrieved July 23, 2017, from https://www.census.gov/population/international/data/idb/worldpoptotal.php

Zeadally, S., & Winkler, S. (2016). Privacy Policy Analysis of Popular Web Platforms. *IEEE TECHNOLOGY AND SOCIETY MAGAZINE*, (june), 75–85.

CHAPTER 4. EXTENSIVE MANUAL ANALYSIS OF PRIVACY POLICIES

This chapter summarizes the important outcomes from the survey mentioned in chapter 3. Also, important guidelines are suggested from the survey. This chapter focuses on one of the important guidelines, - Privacy awareness, the challenges in spreading awareness are discussed in this chapter. Privacy policy analysis is one of the challenges in privacy awareness. Hence, this chapter discusses the extensive privacy policy analysis carried using the manual as well as automated methods The chapter concludes with important findings from the analysis like data collection, sharing, security measures from the perceptive of privacy policies.

The work discussed in this chaper is already presented in a conference (Paper no: Section 1.7, B3). The updates in this chapter include the outcome of survey (section 4.1), followed by objective of doing manual analysis of privacy policies (section 4.2), methodology of manual analysis (section 4.3), extensive analysis of literature work (section 4.4 and 4.5). Though some figures are from the paper, in this chapter, the focus has been given on the interpretation and writing the results in a different way (e.g. Table 10, figure 4.8 etc). The additions include screen shots of privacy policies, examples of service providers, sentences from the privacy policies, etc.

4.1. SURVEY OUTCOME AND SUGGESTIONS

Our previous research work mentioned in Chapter 3 is the discussion on the survey carried out in the India along with its detailed analysis. The analysis has revealed many findings on users' online practices and threats to privacy. Also, the survey has helped to understand the requirements for privacy protection and privacy management. The mitigating factors from the survey, act as the suggestions or recommendations to develop a mechanism to shield users' information and enhance user empowerment.

Privacy Protection of users' information is accomplished by various solutions. An abiding faith is important for users during effective communication with service providers. A personal trust manager is an entity who can be responsible for supporting users by communicating how safe the website is to communicate, or what is the trustworthiness of the website.

The participants of the survey prefer to be anonymous whenever they are online and communicating with other entities. Anonymization techniques can help to hide the real identity of the users. Also, users should have control over their personal information being collected by service providers. The user now believes that explicit consent should be necessary for service providers before dealing with users' information. There should be limited tracking and monitoring activities by the service providers. This will lead to a solution for service providers that disclose the name of tracking and monitoring entities.

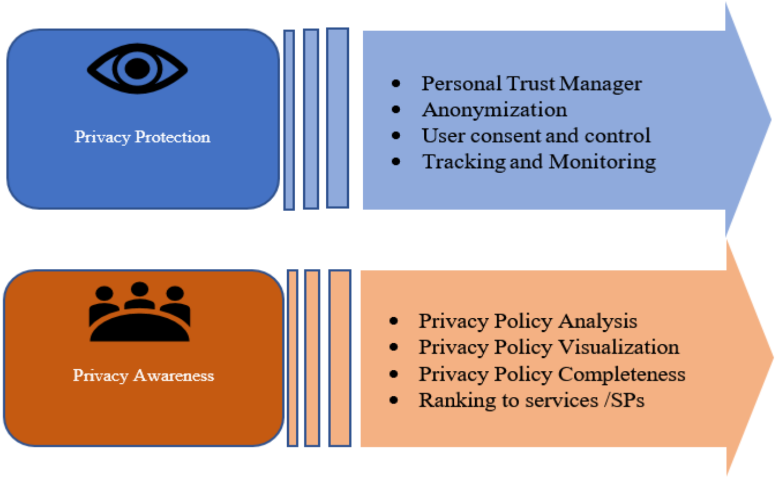


Figure 4-1 Recommendations to handle privacy issues

The participants of the survey also suggested solutions to enhance their privacy awareness. It is important to convey the information gathered, used and shared by service providers. The privacy policy is an important document that describes the user's information management, but its issues motivate to consider further solutions.

4.2. AIM OF THIS CHAPTER

The survey presented in chapter 3 revealed users' privacy concerns – perceptions – issues. One of the issues was related to the privacy policy contents. As the privacy policies are difficult to read and understand, there is need to understand the contents of the privacy policy. Hence, the aim of executing manual analysis of privacy policies is presented below:

- **To know data practices of service providers**: A major concern for users it to understand what are the elements/sections of privacy policies. The analysis also wants to know the data practices of service providers and to know various indicators/elements that will help to enhance users' awareness. The indicators could be service providers' policies over data collection and sharing, or security measures adopted, ways of data collection, and so on.

- **To recognize the contents of the privacy policy to be visualized**: The results from the analysis are valuable and those results need to be shown to the users. So, the visualization is necessary. The analysis will help to develop tools that will visualize the contents of privacy policies. Visualization will be very convenient to see the contents of the entire or particular section that user would like to read. The visualization helps them to make a decision on acceptance or rejection of privacy policies.

So, looking at fig 4-1, we decided to work on designing a privacy solution that will help to spread privacy awareness by analyzing privacy policies.

4.3. METHODOLOGY FOR MANUAL ANALYSIS

The manual analysis of privacy policies (content analysis) is carried out that contains following steps

- Initially, the work started to understand current state of art for Users, privacy policy, and challenges.

- Second, next task is to identify privacy policies from different domains including Bank, Business, Social networking, E-Commerce, News, Government, Wikipedia, Entertainment, Sports, Travelling, etc.

- In the third phase, the extensive manual analysis is performed to understand:
    - Availability of privacy policy link on the website.
    - The structure, language, and readability of privacy policy
    - Aim of privacy policy
    - The set of attributes collected by service providers
    - The possible methods of data collection
    - Data retention strategy
    - Use of cookies
    - Security measures
    - Sharing and selling of data
    - Other aspects like policy change notification to the user in case of a change in policy, children policy on children data, grievance officers, etc.

The manual analysis of privacy policies is presented in a conference. The difference of the paper and this chapter is that this chapter discusses the in-depth analysis of each attribute and other important concerns.

4.4. USER, PRIVACY POLICY AND ISSUES

The internet user has a right to know about personal information management (collection, use, sharing, etc.). One aspect of information privacy is that the user should know what, when, how and to how much the user information, is being gathered and subsequently the purposes it is used for (Wang, Na, Bo Zhang, Bin Liu, 2009). This will help to enhance awareness about privacy. Users privacy awareness affects their online practices/behavior (Wang, Na, Bo Zhang, Bin Liu, 2015), (Kelley, Cranor, & Sadeh, 2013). The privacy awareness level of the user determines whether the online practices of the users is appropriate. So, it is imperative for the user to understand privacy and its implications to make correct and informed decisions.

The privacy policy of service providers is a legal document, acts as a communication channel, and informs the users (who owns the data) about personal information management. With the help of privacy policy, the user is notified that their information is protected and user privacy is protected in accordance with laws or guidelines.

Let’s consider an example. The Gamestation (Video Game Company in the UK) had played a prank on users in the year 2010 (Bosker, 2011). The mischievous act was to secretly add a condition about “stealing the soul of internet users” in their privacy policy. The unexpected outcome of that prank was the company exposed that they have “soul possession” of almost 7,500 users. The crucial point to be noted here is that the vital document of privacy policy made by the service provider with the objective of amusement was accepted by almost 90% users. This comic example leads to serious privacy issues that should be given more attention.

It’s not only really difficult for an end user to make informed decisions if a privacy policy does not offer clear and sufficient information on personal information management in a reader-friendly way but also, hard for service providers to receive users informed decisions that they are looking for (S. Prechal, 2008). Ideally, privacy policies should be completely clear, concise and comprehensive to help the users to make informed decisions (Organisation for Economic Co-operation and Development OECD, 2002). On the contrary, in actual practice the online privacy policies are lengthy, information is incomplete, vague and complicated sentences of the legalese. The length of privacy policy is discouraging due to reading fatigue and information is not clear and appears to be “buried” in technical jargon (S. Talib, S. M. Abdul Razak, A. Olowolayemo, M. Salependi, N. F. Ahmad, S. Kunhamoo, 2014),(Eckert, Claudia, Katsikas, Sokratis K., Pernul, 2014), (Jensen & Potts, n.d.), (P. Dhotre & Olesen, 2015). Not only the length of the privacy policy but also other critical issues are what is the type of privacy policy (service provider), its semantic complexity, notification, etc.

Hence, there is a need to design a new way of analyzing and representing the contents of a privacy policy that would support the users in grasping the contents of privacy policies for making informed decisions.

4.5. PRIVACY POLICY AND NEW CHALLENGES

Analysis, visualization, and evaluation of the privacy policy are rather a new field for researchers in the recent years. Common concerns raised by the researchers are the clarity and length of the privacy policy (Massey, Eisenstein, Anton, & Swire, 2013).

A survey was carried out with more than 800 participants. The study reveals that 80% of the total respondents would be surprised if an app has collected the data that which could be used for other purposes than its intended use (McDonald, Aleecia M., 2013), (Miller, Buck, & Berkeley, 2012). This suggests a fact that the users do not actively consider reading and understanding the privacy policy. The service providers offer users very little information to control the mechanism that is difficult and unacceptable. The EU has recently mentioned in GDPR that the service provider should be obliged to follow new privacy rules. Every organization should consider Privacy by Design and Privacy by Default.

An evaluation matrix was developed to check the approach of policies, their implementation, and usability. It is beneficial to the organizations for maintaining web services. This matrix contains 53 questions applied to 10 websites to improve the websites. The study reveals about checking the implementation aspects by checking them. However, there is a need to produce important privacy policy findings to end user (Miller, Buck, & Berkeley, 2012).

The UML based approach was discussed to evaluate privacy aware systems. This has been implemented for social websites to understand privacy policy requirements for integrating it with third-party systems. (Caramujo & Silva, 2015). A distinct contribution was made in the health sector to assess the privacy policies and its compliance with FIP. The accessibility and readability were evaluated and they were found to be weak (Savla & Martino, 2012). On a large scale, 2,061 privacy policies were analyzed to assess requirements for readability (Massey et al., 2013). Also, the privacy protection or vulnerabilities were investigated using text mining techniques. The challenge is to read and understand privacy policy documents. This method supported the generalizability of multiple domains.

A privacy policy completeness evaluator has been developed using machine learning algorithms (Massey et al., 2013). The algorithm determines the existence of 8 different sections (access, choice, collect, cookies, purpose, retention, security, and share). The proposed approach may facilitate the users to calculate privacy policy completeness score. However, in the proposed method, the method of score calculation is the main concern and needs to be investigated carefully to give a guarantee of transparency.

There is a need to extend such type of work that emphasizes various and vital parameters like readability, unambiguity, content extraction and effective visualization of the privacy policy. Similar tools have been developed for checking the completeness of privacy policy (E. Costante, Y. Sun, M. Petkovi, 2012). In such tools, machine learning algorithms have been used to identify the contents as well completeness of privacy policy the limitations of such tools are a limited number of privacy policies available for analysis, time-consuming and no visualization of privacy policy contents.

Several tools have made a vital contribution to the display of the privacy notice (Wang, Na, Bo Zhang, Bin Liu, 2015),(Fisher, 2013). The limitation of such tools is that the user must navigate in forward and backward tabs/alternatives to get the contents of a privacy policy.

So, in the view of existing research and the outcome of the survey from chapter 3, it is important to focus on privacy policy issues to help users to enhance their privacy awareness. Hence, the most important question is: How could we perform the extensive analysis of privacy policies which follow privacy guidelines and regulations, especially in the collection of personal information attributes, collection methods, cookies used, sharing entities, security methods and other important aspects?

4.6. IDENTIFICATION OF PRIVACY POLICIES FOR ANALYSIS

Our approach to privacy policy analysis is quite different from other researchers' approach (McDonald, Aleecia M., 2013),(Miller, Buck, & Tygar, 2012),(Massey et al., 2013). The primary focus of our study is to identify and analyze the privacy policies of various service providers that are popular in India. The aim of this study is to reveal service provider's policy on collection of personal attributes, it's processing and sharing. Looking at the privacy principles and regulations on privacy (OECD, GDPR), it is important to know the sections of a privacy policy. Analyzing the sections like information collection, sharing, third-party involvement, and security measures; are important and need to bring this on a dashboard where users can read and understand easily.

Even though the Indian law (Law & Affairs, 2008) has directives for the service providers to have their privacy policies, it is important to note that the law doesn't efficiently and precisely stipulate what the contents of privacy policies should be so as to protect users' privacy. Hence, we identified and manually analyzed over 50 privacy policy of the most popular websites listed by Alexa (an Amazon Company)[47] The list of websites is shown in fig 4-2.

[47] http://www.alexa.com/topsites/countries/IN

Fig 4-2 shows the names of the websites sector wise. As the users interact with various service providers to avail of the services, this study considers various sectors to understand the practices of different service providers on personal information management. In this study, it is vital to know and compare the list of attributes collected by each website of various sectors. This study covers the most vital sectors from banking to E-commerce, and from business to entertainment, etc. Also, close attention has been paid while selecting private and government websites. From each sector, an adequate number of websites are selected for manual analysis.

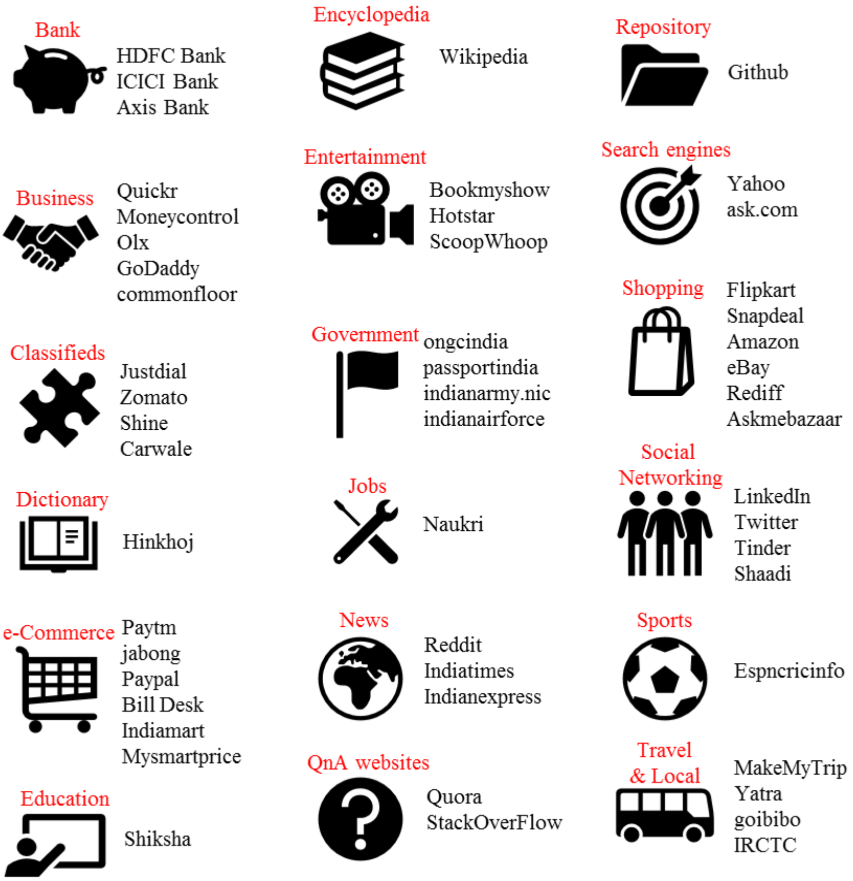


Figure 4-2 Websites names selected for analysis

The 52 websites are selected from several sectors. Each website is visited and as a next step, the privacy policies are downloaded for performing the manual analysis.

Before conducting manual analysis, an analytics tool (Tableau)[48] was used to perform statistical analysis (e.g. Analysis of personally identifiable information gathered by service providers). So, this study is a combination of content analysis using both manual and automated methods.

Further, the sectors and the number of websites chosen for manual analysis of privacy policies is illustrated in fig 4-3. There are 52 websites covered under 18 sectors. Considering the frequency of use, maximum (6) and minimum (1) websites were selected from E-commerce/shopping sectors and dictionary/sports respectively.

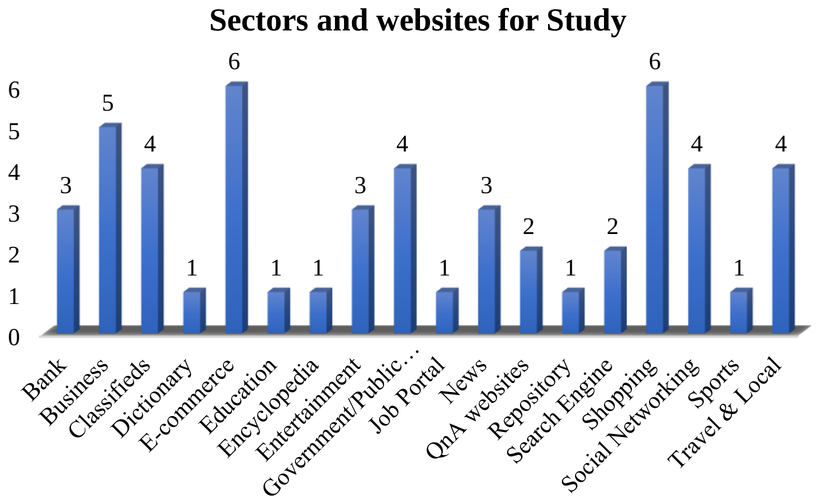


Figure 4-3 Sectors and Website size considered for study (P. S. Dhotre, Olesen, & Khajuria, 2016)

4.7. RESULTS AND DISCUSSIONS

This section of this chapter describes the important findings obtained from manual analysis of privacy policies. The important findings include accessibility to the privacy policy, structure and language, objectives of privacy, readability, personal attribute collection, collection methods, cookies, sharing of information, security, and other findings. Each finding is discussed in detail.

[48] https://www.tableau.com/

4.7.1. ACCESSIBILITY OF PRIVACY POLICY

After identifying websites, the study visited each chosen website and immediately noted the link to its privacy policy. The active privacy policy (whose link is active) is downloaded for manual analysis. However, there are few cases where the link is found to be inactive or dead (e.g. www.irctc.co.in). For such websites, we tried to consider other web pages of the websites or on the developer's website. Sometimes we took the help of search engines to locate privacy policy link.

Most of the websites have kept a hyperlink to their privacy policies on the home page. It is observed that the link is found at the bottom of the homepage. In very few cases it was observed that the link is found at the top left part of the webpage. An important observation for such websites is that the websites contain homepages with the scroll.

Terms of Service (ToS) contain sets of rules that a user should agree to before utilizing the services offered by service providers. Even though there is a difference between privacy policy and ToS, some of the websites merge ToS and privacy policy.

But, it is very safe to say that websites followed the same practice for privacy policy link location. Very few websites like *MoneyControl*[49], the link to privacy policy was not found on the home page. However, the link is accessible from other pages of the website.

4.7.2. STRUCTURE, LANGUAGE, AND READABILITY

One of the important and noticeable results starts from this point of the section. During the analysis, each privacy policy is checked for its format of privacy policies, and the language being used. Also, the analysis also included the readability of the privacy policy to know how much is the privacy policy readable?

- Structure:

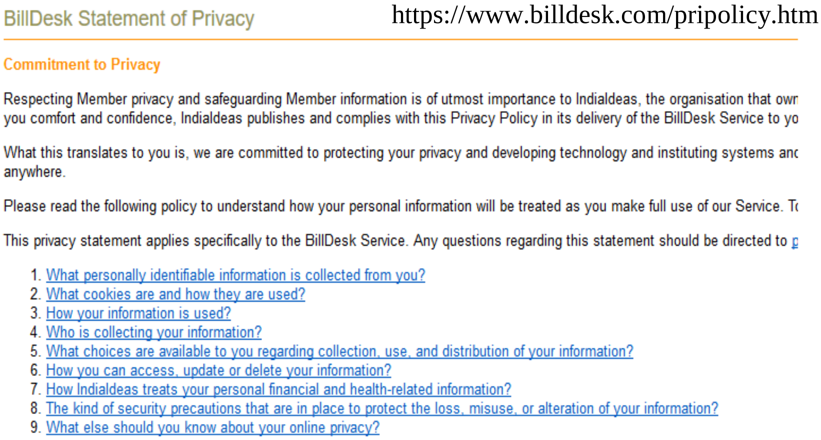
The selected privacy policies have different structures when compared to each other. Several ways are adopted by the service providers in structuring the privacy policy contents.

One method of the arranging the contents of the privacy policy is in the tabular format. The contents are divided into rows and column combinations. The column represents the heading of the sections and relevant content is represented in the row. Sometimes, for a few websites, the tables are not visible to Internet users. Once the language

[49] *www.moneycontrol.com*

(XACML, HTML, EPAL) is identified and the source code is seen, then it's easy to conclude that the contents are represented in the tabular method.

The second method is in the form of question and answer format presented from websites like *BillDesk*, *Stack Overflow*, etc. The complete privacy policy is represented by answering the questions that are common. For example, the questions would be based on what kind of information is collected, cookies details, etc. For example, the privacy policy of *BillDesk*[50] has listed 9 questions at the beginning of their privacy policy.



https://www.billdesk.com/pripolicy.htm

BillDesk Statement of Privacy

Commitment to Privacy

Respecting Member privacy and safeguarding Member information is of utmost importance to Indialdeas, the organisation that owr you comfort and confidence, Indialdeas publishes and complies with this Privacy Policy in its delivery of the BillDesk Service to yo

What this translates to you is, we are committed to protecting your privacy and developing technology and instituting systems anc anywhere.

Please read the following policy to understand how your personal information will be treated as you make full use of our Service. Tı

This privacy statement applies specifically to the BillDesk Service. Any questions regarding this statement should be directed to p

1. What personally identifiable information is collected from you?
2. What cookies are and how they are used?
3. How your information is used?
4. Who is collecting your information?
5. What choices are available to you regarding collection, use, and distribution of your information?
6. How you can access, update or delete your information?
7. How Indialdeas treats your personal financial and health-related information?
8. The kind of security precautions that are in place to protect the loss, misuse, or alteration of your information?
9. What else should you know about your online privacy?

Figure 4-4 Screenshot of BillDesk's[51] *privacy policy.*

Here is a question 2 mentioned in the privacy policy of *www.billdesk.com*[5]*:*

> *"What are cookies and how does IndiaIdeas use them?*
>
> *Cookies are pieces of information that a Web site transfers to your computer's hard disk for record-keeping purposes. Cookies can make the Web more useful by storing information about your preferences on a particular site. The use of cookies is an industry standard, and many major Web sites use them to provide useful features for their customers. Cookies in and of themselves do not personally identify users, although they do identify a user's computer. Most browsers are initially set up to accept cookies. If*

[50] *www. billdesk.com*

[51] *https://www.billdesk.com/pripolicy.htm*

> *you'd prefer, you can set yours to refuse cookies. However, you may not be able to take full advantage of a Web site if you do so.*
>
> Currently, *BillDesk does not use cookies but may do so in the future uses cookies to track user traffic patterns and to help us assist you with questions about site navigation, functionality or performance.*
>
> *The third category observed is in terms of bullet points.*
>
> *So, it is safe to say the websites are not having common structure, used complex language which creates a confusion among users. Also, the readability of privacy policy is the main concern for almost every website…"*

The third method is in the form of bullet points. Each bullet point describes one section of the privacy policy. For example, a website like *Snapdeal*[52] has followed this structure:

> *"Snapdeal has provided this Privacy Policy to familiarize You with:*
>
> - *The type of data or information that You share with or provide to Snapdeal and that Snapdeal collects from You;*
> - *The purpose for collection of such data or information from You;*
> - *Snapdeal's information security practices and policies; and*
> - *Snapdeal's policy on sharing or transferring Your data or information with third parties. This Privacy Policy may be amended/updated from time to time. Upon amending/ updating the Privacy Policy, we will accordingly amend the date above. We suggest that you regularly check this Privacy Policy to apprise yourself of any updates. Your continued use of Website or provision of data or information thereafter will imply Your unconditional acceptance of such updates to this Privacy Policy…"*

There are some other methods that follow a format where the complete privacy policy is divided into various sections. Each section has a title followed by the related policy contents. One example from a travel company *MakeMyTrip*[53] is mentioned as:

[52] *https://www.snapdeal.com/page/privacy-policy*

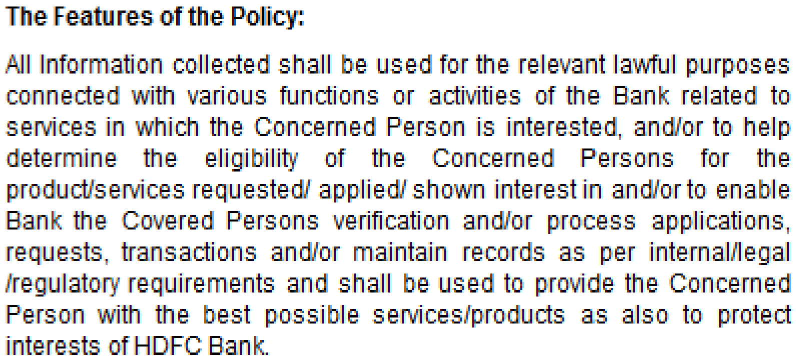
[53] *https://us.makemytrip.com/various/privacy-policy.htm*

> *"What Personal Information we collect from you and how we use it?*
>
> *'Personal Information' means and includes all information that can be linked to a specific individual or to identify any individual, such as name, address, mailing address, telephone number, email address, credit card number, cardholder name, expiration date, information about the travel, bookings, passengers, frequent traveller / flier numbers, and any and all details that may be necessary from the customer…."*

So, looking at the three ways described above of describing privacy policy, we conclude that the contents of the privacy policy are disorganized. The monolithic structure of privacy policy will create confusion among Internet users. The difficulty level of accessing particular information from privacy policy will be high. Hence, there is a need for a standard structure for displaying the contents of the privacy policy.

- Language

Additionally, the language used in the privacy policy is observed carefully and it is found that the language is also an issue of concern for the users. Most words or terms in the privacy policy are convoluted that becomes a tedious job for users in India. For example, privacy policy HDFC[54] (a well-known and frequently visited bank in India) has used complicated terms that become difficult to read. For example, complicated information is presented in the privacy policy and is mentioned in fig 4-5.



The Features of the Policy:

All Information collected shall be used for the relevant lawful purposes connected with various functions or activities of the Bank related to services in which the Concerned Person is interested, and/or to help determine the eligibility of the Concerned Persons for the product/services requested/ applied/ shown interest in and/or to enable Bank the Covered Persons verification and/or process applications, requests, transactions and/or maintain records as per internal/legal /regulatory requirements and shall be used to provide the Concerned Person with the best possible services/products as also to protect interests of HDFC Bank.

Figure 4-5 Screenshot of HDFC bank's privacy policy[55]

[54] *https://www.hdfclife.com/privacy-policy*

[55] *https://www.hdfcbank.com/*

As a complicated example, Fig 4-5 illustrates a portion of the privacy policy of HDFC bank. It is a single sentence that spans over 10 lines and has more than 95 words. Despite the length of the sentence, the language (a set of words) is very difficult to understand.

Hence, the content of the privacy policies should be simple to understand and must be short in the length.

- Readability

The readability analysis has been performed on each chosen privacy policy. The analysis is based on the readability test performed by an online tool (ReadabilityScore.com, 2016). This tool uses a formulation provided by the Flesch-Kincaid. This formula helps to find the grade level of a text, called as “score”. This formula analyzes an average number of words per sentence followed by the average number of syllables per word. Along with these calculations, a constant is combined to get the final score of the text. It is a simple test that rates readability with increasing scores i.e. “Greater the score the better is the readability”.

The readability level and its mapping with various school level ages are mentioned in Table 9. Here the school level considered is in the context of US, not India.

Score	School Level	Remark/Note
90.0–100.0	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
80.0–90.0	6th grade	Easy to read. Conversational English for consumers.
70.0–80.0	7th grade	Fairly easy to read.
60.0–70.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
50.0–60.0	10th to 12th grade	Fairly difficult to read.
30.0–50.0	College	Difficult to read.
0.0–30.0	college graduate	Very difficult to read. Best understood by university graduates.

Table 9 Readability score (ReadabilityScore.com, 2016)(P. S. Dhotre et al., 2016)

So, the readability score of each privacy policy document is calculated and the result is represented in fig 4-6. The Y-axis represents the readability score range and the X-axis represents the number of privacy policies based on their score.

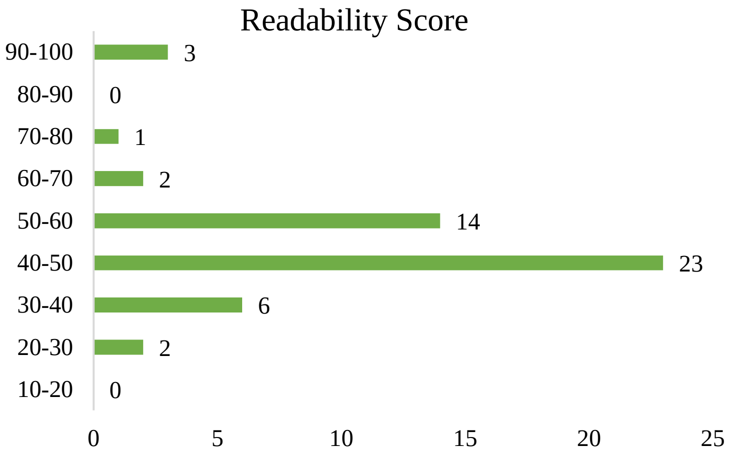


Figure 4-6 Readability score of each privacy policy (P. Dhotre, Olesen, & Samant, 2016)

This analysis reveals the fact that most of the privacy policies (i.e. 37 out of 52) score in the range of 40 to 60. So, the vital finding is that there are very few policies (6) whose score is 60 or more than 60 which is required to enhance readability. Hence, this shows that very few websites' privacy policy is easy to read and users or an 11-year-old student will understand it easily. Comparing the score from fig 4-6 and the table 4.1, we observed that more than 30 (out of 52) privacy policies are difficult to read and be understood by graduate students at university level.

In short, the outcome of this study shows that privacy policies are very difficult to read and understand. Therefore, there is a need to simplify the privacy policy without using complicated language irrespective of the education level of users.

4.7.3. OBJECTIVE OF PRIVACY POLICY

This section focuses on the objective or aim of the privacy policy. Ideally, every website/company should comment on the purpose or aim of privacy policy at the beginning of privacy policy document. The aim should describe the purpose of the privacy policy that it adheres to in the area of privacy protection principles.

During the analysis of the various privacy policies, we identified different objectives as the collection and management of data, security, and privacy, etc. As shown in fig 4-7, there are only 31% websites who have stated that their aim is to provide security and privacy to users/customer data in the beginning of privacy document. This shows that the websites do have the right intention to inform the purpose of collection, sharing of users' information. However, their count is very less.

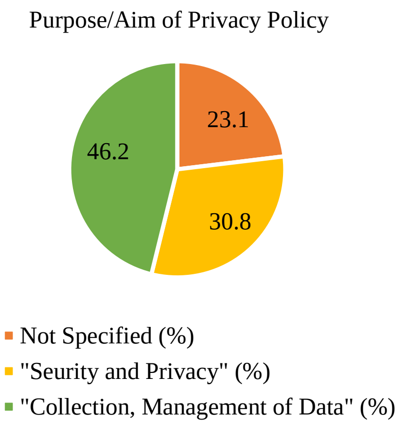


Figure 4-7 Purpose of privacy policies (P. S. Dhotre et al., 2016)

Ideally, a privacy policy should focus on protection of users' information. Some of the examples from LinkedIn, Flipkart are given below:

- The privacy policy of Linkedin.com[56] has a statement that says:

> "*Our aim is for you—our members—to always feel informed and empowered with respect to your privacy on LinkedIn.*"

Another example:

- Flipkart.com[57] also states that their primary goal is:

> "*Our primary goal in doing so is to provide you a safe, efficient, smooth and customized experience.*"

The two examples given above have clearly mentioned about various terms like informed, empowerment, respect, safety, user experience, etc. This shows that the service providers intend to provide security to users' information along with protecting users' information privacy.

On the contrary, the purpose of remaining websites (46%) is to describe data collection, how they collect, what happens to user data, and so on. The websites like twitter.com, quiker.com have similar objectives as eBay.in.

[56] *https://www.linkedin.com/legal/privacy-policy*

[57] *https://www.flipkart.com/pages/privacypolicy*

The privacy policy of eBay.in focused on the description of personal information management like the collection, disclosure, retention of information and protection.

- An example from the privacy policy of eBay.in[58] states that:

> *"This Privacy Notice describes our collection, use, disclosure, retention, and protection of your personal information."*

The most surprising finding is that 23% of the total privacy policies do not specify any objective in their privacy policy. This could be lead to privacy risk to users' information. Thus, not all web service providers have clearly mentioned the aim of their privacy policies. This will not only raise an issue about privacy risk but also will lead to lack of transparency in the handling of users' information by the service providers. Transparency can be one of the parameters to decide the trustworthiness of the website.

4.7.4. PERSONAL ATTRIBUTE COLLECTION

This section discusses the analysis and comparison of personal information collected as per the privacy policies of several websites in this study. During the analysis, we identified important personal and other attributes. These attributes are checked in each privacy policy. The analysis was based on whether the websites collect these attributes or not. Some privacy policies have mentioned about the collection or use of attributes.

So, depending on the statements on attributes like (first name, last name, user name, and password), we have defined 1,0 and NS. Here, 1 represents that the privacy policy clearly stated that they will collect that user attribute. If they do not collect attributes as per the privacy policy, then 0 is marked. However, the NS- no sentences- is used for those privacy policies where there are no sentences on a collection of user attributes. The attributes collection in each privacy policy is represented in table 10.

Service Provider	**First Name**	**Last Name**	**User name**	**Password**
Amazon	1	1	1	1
ask.com	1	1	1	1
Askmebazaar	1	1	NS	NS
Axis Bank	1	1	NS	1
Bill Desk	1	1	NS	1
Bookmyshow	1	1	NS	NS
CartoonNetworkIndia	1	0	1	1

[58] *https://pages.ebay.com/help/policies/privacy-policy.html*

Commonfloor	1	1	NS	1
Coursera	1	1	NS	NS
eBay	1	1	NS	NS
Espncricinfo	1	1	1	1
Flipkart	1	1	NS	1
GitHub	1	1	NS	NS
Glassdoor	1	1	NS	1
GoDaddy	1	1	NS	NS
goibibo	NS	NS	1	1
HDFC Bank	NS	NS	NS	1
hinkhoj	NS	NS	NS	NS
Hotstar	1	1	1	1
ICICI Bank	NS	NS	NS	NS
Indiamart(IIL)	1	1	NS	NS
Indianairforce	1	1	NS	NS
Indianarmy.nic	NS	NS	NS	NS
Indianexpress	1	1	NS	NS
Indiatimes	1	1	NS	NS
jabong	1	1	NS	NS
Justdial	1	1	NS	NS
LinkedIn	1	1	NS	1
MakeMyTrip	1	1	1	1
Moneycontrol	1	1	NS	NS
Mysmartprice	NS	NS	NS	NS
Naukri	1	1	NS	NS
Olx	1	1	1	NS
ongcindia	1	1	NS	NS
passportindia	NS	NS	NS	NS
Paypal	1	1	NS	NS
Paytm	1	1	NS	NS
Quickr	NS	NS	NS	NS
Quora	1	1	1	1
Reddit	1	1	1	1
Rediff	1	1	NS	NS
ScoopWhoop	1	1	NS	NS
Shaadi	1	1	NS	1

Shine	1	1	1	1
Snapdeal	1	1	NS	1
StackOverFlow	NS	NS	1	NS
tinder	1	1	NS	NS
Twitter	1	1	1	1
Wikipedia	0	0	1	1
Yahoo	1	1	NS	NS
Yatra	1	1	NS	NS
Zomato	1	1	NS	1

Table 10 Attribute collection policy of service providers on first name, last name, user name, and password

The above table illustrates the list of websites and the four attributes. Based on the contents of privacy policies, the values (1 or 0 or NS) are marked. The summary of above table is presented in fig 4-8.

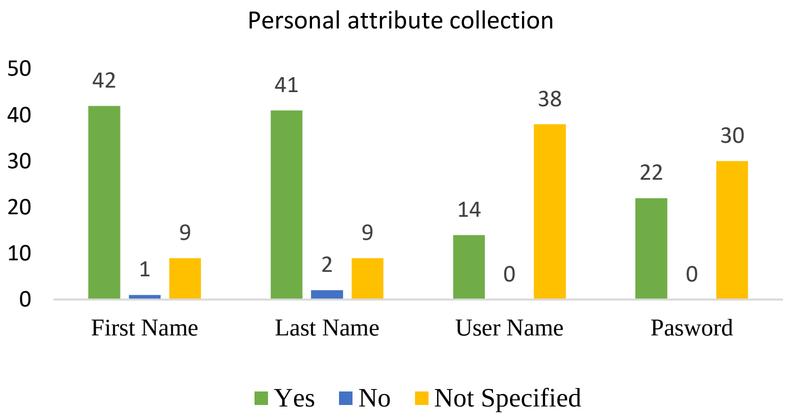


Figure 4-8 Personal attribute collection specification given in the privacy policies

It is observed that 80% of the websites (42 out of 52) collect the first name from users. Only one website (Wikipedia.com) does not collect the first name as well as the last name. Unexpectedly, 9% of service providers have nothing to say about the collection of the first name. Similar statistics can be observed for the last name too. So, the statistics denote that the first and last name are collected on a large scale and with a possible intention to be used.

A similar analysis of username and password are done, which also are vital information of users. From the observations in this study, 14 privacy policies clearly stated that they collect username and 22 policies stated that they collect the password from the users. However, it is not clear from 38 service providers about their strategy on a collection of the username. Similarly, 30 privacy policies have no sentences on the collection of users' password in their privacy document.

Hence, the service provider's approach is open to suspicion in the collection of users' attributes. So, it is necessary to understand the information collection types, methods, and its use.

4.7.5. INFORMATION COLLECTION METHODS

Moreover, with attribute collection, it was observed that around 80% of the policies studied had agreed that more user attributes are collected and assessed. The attributes were collected in different contexts. This includes basic attributes like the first name, last names, contact number, email id, etc. Specific attributes like financial information were collected while performing banking operations.

The banking websites are interested in the collection of the income range of users. Few privacy policies stated that they handled income values. Half of the total privacy policies stated that billing information and credit/debit card/bank account details are collected and stored by them. It is observed that the use of each attribute is not clearly stated in the privacy policy. So, the question is what is the utilization/purpose of each user characteristics by websites?

In this section, the study talks about the identification of several ways of information collection adopted by the service providers. While reading the privacy policies of service providers, we observed vital information collection methods represented prominently under four titles as registration, browsing, cookies, and forms. The detailed analysis is represented in Fig.4-9

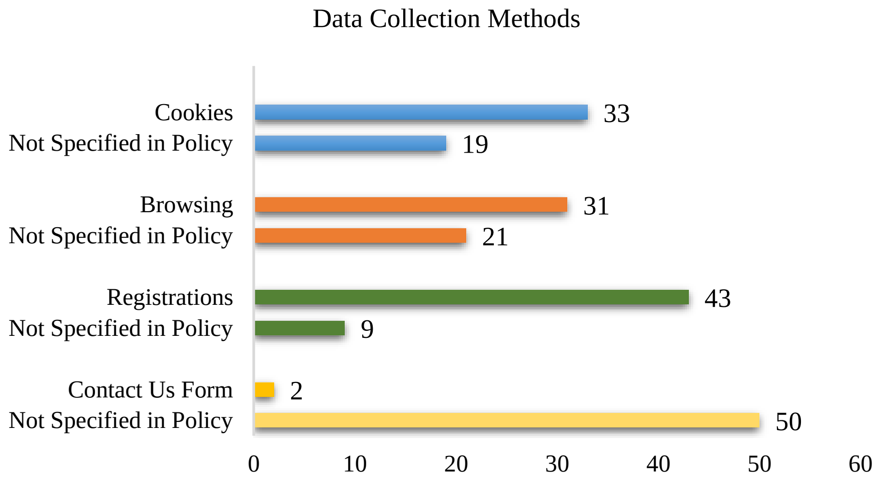


Figure 4-9 Data collection ways as per privacy policy analysis (P. S. Dhotre et al., 2016)

The four vital methods of data collection are represented on Y-axis. The X-axis represents the number of service providers. Collecting users' information from registration is the major source of collection. Almost 83% (43 out of 52) of service providers gather user information when the users fill the registration forms. The subsequent information collection methods include cookies, browsing, and contact us form.

During the analysis, it is found that almost 63% of privacy policies collect the information from cookies. Information generated at the time of browsing is considered vital for almost 60% service providers.

Hardly 2 service providers stated in their privacy policy that they will collect user information when users wish to contact service providers.37% and 41% of total service providers studied did not mention about information collected by means of either cookies or browsing. This high value is a threat to users' privacy.

Apart from the four methods just discussed, there are other ways of information collection like information from user device or location (toc.io, 2015). The details about these methods are not found in the study or discussed in this chapter.

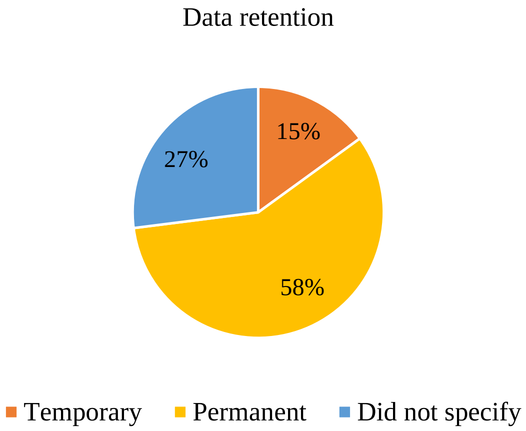


Figure 4-10 Data retention by service providers (P. S. Dhotre et al., 2016)

Once the data is collected, it is important to know the location of data storage and its retention. The deep study of duration of data storage by service providers is represented in the fig 4-10. This reveals that there are only 15% service providers who stated that they store some or every information temporarily. Permanent data storage is mentioned by 58% service providers in their respective policy. But, 35% of the policies did not clearly specify the duration during which they will retain data, it reads something like ‘as long as necessary’.

Each user who is an owner of the information submitted to the service providers has the rights to access their information. The study brings insights on access, modification of users' data. Out of the total privacy policies studied in this part of the research, we identified that 67% privacy policies have mentioned about the right to access or modify user data. However, there are still 23% privacy policies who have not said anything about user rights to access or modify data.

But, overall the study brings to light the fact that the users are not notified about information collection, their collection methods, and do not have ways for effective management of their data. Hence, there is a need to design a 'dashboard' showing how service provider collects users' information and use it.

4.7.6. COOKIES

From the previous section, we observed that the cookies are an important way of information collection. Hence, there is need to do an in-depth study on cookies. This part of the section describes the types of cookies being used by service providers and the information gathered by each type of cookie.

The study says that 88% (46 out of 52) of service providers stated that they are prepared to use cookies to collect and store users' information. The collected information consists of non-personal information like type of OS, browser details, clickstream information, ISP, etc. But, cookies also collect profiling information like gender, location, age, income etc. However, there is a lot of information collected from the browser as mentioned in a study(toc.io, 2015).

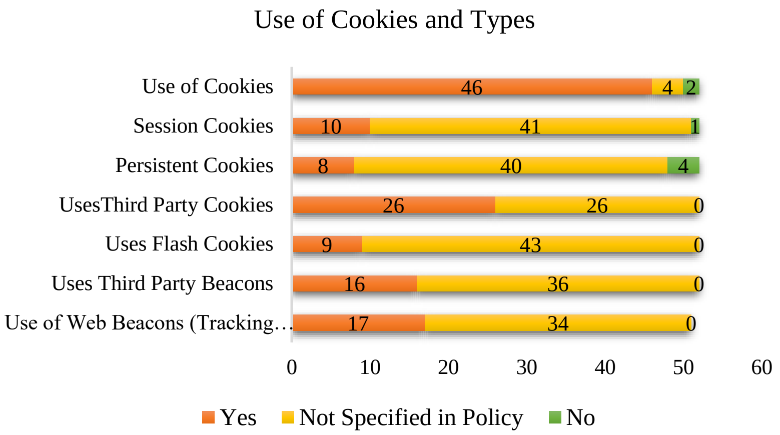


Figure 4-11 Categories of cookies identified in the privacy policy analysis (P. S. Dhotre et al., 2016)

Fig 4-11 represents diverse types of cookies and their use by service providers during interaction with users. The type of cookies found in the analysis includes session cookies, persistent cookies, third-party cookies, flash cookies, etc.

Almost every service provider (46 out of 52) uses cookies; however, the analysis was further carried to know the cookie types and its specification in the privacy policy.

Session cookies and persistent cookies use are mentioned by 10 and 8 service providers respectively (for example Olx.in). This clearly indicates that the other service providers (almost 43 out of 52) have kept users in the dark on cookies. Hence, there is a need to specify the complete details about cookies (types, set of information they collect, purpose, etc.).

The other aspect of analysis on cookies revealed the huge use of cookies by service providers. But, the privacy policies have not mentioned the type of cookies they use. If the service providers provide the complete information on cookies, then hopefully, the users will be interested in sharing information. Hence, there is a need for transparency and clarity in the privacy policy document.

Another key observation is that there are many websites that use unencrypted cookies.

- For example, hdfc.com[59] stated one sentence on cookies as:

  "*We may in the future implement encryption of the cookies*".

This clearly shows that at present the cookies are unencrypted and therefore vulnerable. Also, if such cookies get into the hands of unauthorized entities, then it will surely lead to many privacy issues.

4.7.7. SECURITY

Regarding the protection of users' privacy, the security of personal information is vital.

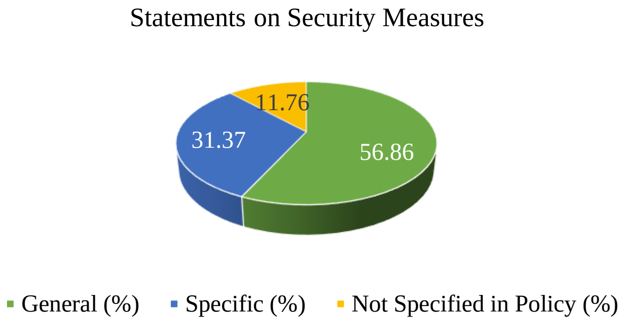


Figure 4-12 Categorization of statements on security measures from analysis (P. S. Dhotre et al., 2016)

The analysis further extended to understand the security measures adopted by service providers. From fig 4-12, we can say that 56% of total privacy policies have given general statements on use encryption considering users' data security.

To highlight few sentences on security, here are few examples of privacy policies:

- Axis bank's privacy policy contains a statement on security as:

  "*The security of personal information is a priority and is protected by maintaining physical, electronic, and procedural*

[59] *https://www.hdfc.com/privacy-policy*

> *safeguards that meet applicable laws. Employees are trained in the proper handling of personal information.*"

- The reddit.com has a general statement in the privacy policy on security as:

  "*We take reasonable measures to help protect information about you from loss, theft, misuse and unauthorized access, disclosure, alteration, and destruction.*"

Looking at the above examples, the service providers are providing security for users' data security without specifying proper names of security measures. There is no guarantee of data security in spite of having strong security standards nowadays.

However, only 31 % of total service providers have precisely declared that collected users' data would be secured by means of encryption techniques like SSL. Also, a clear statement was observed that specify security standard like PCI-DSS (Wikipedia, n.d.), US-EU-Swiss Safe Harbor Principles (European Commission, 2013). However, the companies should make sure that their business partners should abide to protect user privacy. Another finding is, there are some policies that stated nothing about security measures.

For example:

- In the privacy policy of snapdeal.com, the statement is:

  "*...recommended data security standard for security of financial information such as the Payment Card Industry Data Security Standard (PCI-DSS).*"

- The privacy policy of github.com included a statement on security as:

  "*GitHub adheres to the US-EU and US-Swiss Safe Harbor Privacy Principles of Notice, Choice, Onward Transfer, Security, Data Integrity, Access and Enforcement, and is registered with the U.S. Department of Commerce's Safe Harbor Program.*"

The study also checked service providers who have earned the TRUSTe seal. There are very few service providers (only 6 websites (11%)) who have tried to gain the customers trust by adhering the rules specified by TRUSTe. So, the users or customers, hopefully, will not trust or not give respect to the services offered by the websites they visit. This is a clear sign of less commitment from the service providers to transparency, choice, and accountability in the process of handling users' valuable information.

Since in this study, the policies that clearly promise encryption are used this indicates that very few take practical and authorized measures to ensure complete protection of user information.

4.7.8. SHARING AND SELLING OF DATA

an analysis of cookies and security aspect was followed by the analysis of the sharing of data and its use. As per the analysis, websites do not reveal complete information on the use of users' data despite the various collections mentioned in section 4.4.6.

At the outset, the analysis reveals the fact that most have general statements on the purpose of information collection like to improve services, to enhance the user experience, to measure consumer interest, to advertise, to troubleshoot and resolve disputes, to do research, to name some. However, there is no clear understanding of each purpose of every information collected.

Third party advertisement is claimed by 75% of the websites. However, 65% of the policies stated or indicated that they would share the user's data with other corporate entities and affiliates. The sharing entities are classified and are shown in Fig. 4-13.

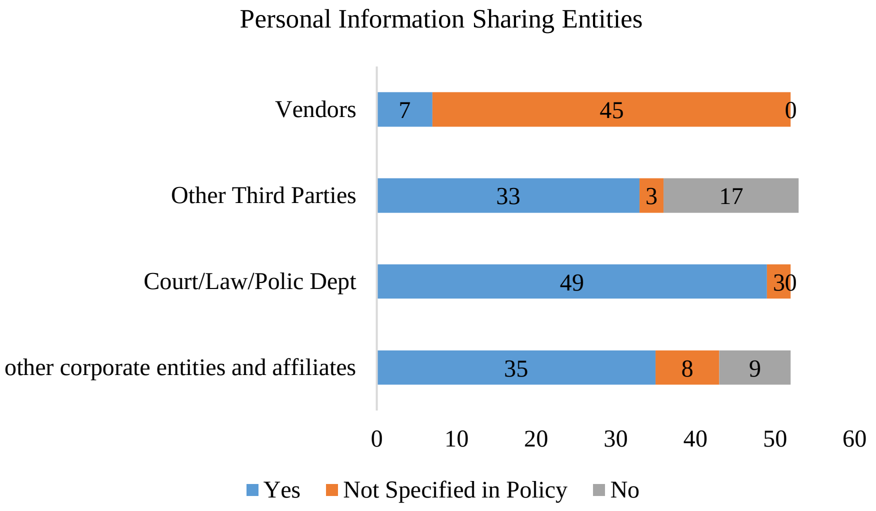


Figure 4-13 List of personal information sharing entities (P. S. Dhotre et al., 2016)

There are only 7 websites (13%) who said that they will share information with their vendors. It means that other service providers did not mention anything about data sharing with vendors.

33 websites are sharing users' information to third-party which is more as compared to the sharing with vendors This value is really high and leads to privacy risks as the privacy policy doesn't specify the details about third parties like name, location, purpose, etc.

Considering court/law cases, almost every (49/52) service providers agree to share users' information. As far as the selling of personal information is concerned, only 26 out of 52 service providers clearly mention in their privacy policies that they will not share/sell the data with third parties without notifying the user. While this is a key factor in the privacy policy, 48% of service providers did not mention whether they are going to share users' data with/without informing the user.

Do Not Track (DNT) is an experimental feature available in popular web-browsers (it is a proposed header field in HTTP Protocol), however, since the responses to the signal are yet to be standardized by W3C different websites have different behaviors towards the signal. 6 websites claimed that they would not respond or would not respond to the DNT signal sent by the browser. Only 2 websites (Twitter, Reddit) mentioned that they would change their behavior and stop serving targeted advertisements to such accounts that have do not track enabled

4.7.9. OTHER FINDINGS

The analysis continued to find other results that might be useful to notify the users. The other findings include the location of data storage, policy on children's data, notification of privacy policy change, etc.

Sometimes it is necessary to notify the users about the location of their data. But, unfortunately,85% of total privacy policies have not mentioned about data storage location. The user is unaware of their data location in terms of city/country. Only 15% websites like Amazon (Luxembourg), Wikipedia (San Francisco, California), Quikr (Mumbai, India) have mentioned their user's data location.

The children's data is equally important and needs to be treated well. The surprising finding from the analysis is that only 27% service providers had special policies to handle children's data and/or explicitly stated whether children are allowed to use the web-service. This means that the rest of all service providers/websites don't specify statements on children's data handling in the privacy policies.

Also, it is necessary to specify the name of grievance officers, in case of a privacy breach or users desire to communicate with regards to issues of the privacy policy. However, only a few have mentioned grievance officers and their details.

Last but not least, the finding of this study is about notification of privacy policy change. It is obvious that the privacy policies change over time. Hence, change in Privacy policies is inevitable. The analysis shows that a maximum number of service providers change their policies without notifying the user. Presently, half of total websites prefer to show notification of change on the same page. It is a tedious job to see the latest changes to the privacy policy by visiting the page from time to time. But, 47% of privacy policies do not contain any statements on privacy policy change. However, it will be interesting to know the privacy policy changes over a period of time with the help of effective communication like email.

4.8. REFERENCES

Bosker, B. (2011). 7,500 Online Shoppers Accidentally Sold Their Souls To Gamestation. *Times of India*.

Caramujo, J., & Silva, A. M. R. da. (2015). Analyzing Privacy Policies Based on a Privacy-Aware Profile: The Facebook and LinkedIn Case Studies. *2015 IEEE 17th Conference on Business Informatics*, 77–84. http://doi.org/10.1109/CBI.2015.44

Dhotre, P., & Olesen, H. (2015). A Survey of Privacy Awareness and Current

Online Practices of Indian Users. In *Proceedings of WWRF Meeting 34, Santa Clara, CA, USA, Apr. 2015* (p. 10). WWRF. Retrieved from http://vbn.aau.dk/en/publications/a-survey-of-privacy-awareness-and-current-online-practices-of-indian-users(92c00b4f-a720-45b8-b3cb-9cfbffc7d4bc).html

Dhotre, P. S., Olesen, H., & Khajuria, S. (2016). Interpretation and Analysis of Privacy Policies of Websites in India. *Proceedings of WWRF Meeting 36, Beijing, China, June 2016*.

E. Costante, Y. Sun, M. Petkovi, J. H. (2012). A machine learning solution to assess privacy policy completeness: (short paper). In *2012 ACM workshop on Privacy in the electronic society (WPES '12). ACM, New York* (pp. 91–96).

Eckert, Claudia, Katsikas, Sokratis K., Pernul, G. (2014). Trust, Privacy, and Security in Digital Business. In *11th International Conference, TrustBus 2014*.

European Commission. (2013). Functioning of the Safe Harbour from the Perspective of EU Citizens and Companies Established in the EU. Retrieved May 18, 2016, from http://ec.europa.eu/justice/data-protection/files/com_2013_847_en.pdf

Fisher, E. K. C. J. L. (2013). Nudging People Away from Privacy-Invasive Mobile Apps through Visual Framing. In *IFIP Conference on Human-Computer Interaction INTERACT 2013* (pp. 74–91).

Jensen, C., & Potts, C. (n.d.). Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=21B25342C369AFCB78E1B4B46B03BAED?doi=10.1.1.629.1633&rep=rep1&type=pdf

Kelley, P. G., Cranor, L. F., & Sadeh, N. (2013). Privacy as Part of the App Decision-Making Process (CMU-CyLab-13-003). Retrieved from https://pdfs.semanticscholar.org/0694/a1b6958be9c5e49fdd94430a285e839cd740.pdf

Law, M., & Affairs, C. (2008). *The Information Technology ACT , 2008* (Vol. 1922). Retrieved from https://cc.tifrh.res.in/webdata/documents/events/facilities/IT_act_2008.pdf

Massey, A. K., Eisenstein, J., Anton, A. I., & Swire, P. P. (2013). Automated text mining for requirements analysis of policy documents. *2013 21st IEEE International Requirements Engineering Conference, RE 2013 - Proceedings*, 4–13. http://doi.org/10.1109/RE.2013.6636700

McDonald, Aleecia M., and T. L. (2013). Nano-notice: Privacy disclosure at a mobile scale. *Journal of Information Policy 3, Vol. 3*, 331–354.

Miller, B., Buck, K., & Berkeley, U. C. (2012). Systematic Analysis and

Evaluation of Web Privacy Policies and Implementations, (December), 534–540.

Miller, B., Buck, K., & Tygar, J. D. (2012). Systematic Analysis and Evaluation of Web Privacy Policies and Implementations. *International Conference for Internet Technology and Secured Transactions (ICITST-2012)*. Retrieved from https://people.eecs.berkeley.edu/~tygar/papers/Buck.pdf

Organisation for Economic Co-operation and Development OECD. (2002). OECD guidelines on the protection of privacy and transborder flows of personal data.

ReadabilityScore.com. (2016). Measure the Readability of Text! - Improve your writing and your website marketing with Readability-Score.com.

S. Prechal, M. E. L. (2008). General Principles of EC Law in a Process of Development 2008, Austin, TX: Wolters Kluwer (pp. 201–242).

S. Talib, S. M. Abdul Razak, A. Olowolayemo, M. Salependi, N. F. Ahmad, S. Kunhamoo, and S. K. B. (2014). Perception analysis of social networks' privacy policy: Instagram as a case study. In *2014 5th Int. Conf. Inf. Commun. Technol. Muslim World, ICT4M 2014* (pp. 3–7).

Savla, P., & Martino, L. D. (2012). Content analysis of privacy policies for health social networks. *Proceedings - 2012 IEEE International Symposium on Policies for Distributed Systems and Networks, POLICY 2012*, 94–101. http://doi.org/10.1109/POLICY.2012.20

toc.io. (2015). Device Fingerprint. Retrieved May 18, 2016, from http://noc.to/#!Help

Wang, Na, Bo Zhang, Bin Liu, and H. J. (2009). Information transparency and digital privacy protection: are they mutually exclusive in the provision of e-services? In *Journal of Services Marketing* (Vol. 23, pp. 154–164). http://doi.org/10.1108/08876040910955161

Wang, Na, Bo Zhang, Bin Liu, and H. J. (2015). Investigating Effects of Control and Ads Awareness on Android Users' Privacy Behaviors and PerceptionsNo Title. In *MobileHCI '15 Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*.

Wikipedia, the free encyclopedia. (n.d.). Payment Card Industry Data Security Standard. Retrieved from https://en.wikipedia.org/wiki/Payment_Card_Industry_Data_Security_Standard

CHAPTER 5. VISUALIZATION OF PRIVACY POLICIES USING PRIVACY POLICY ELUCIDATOR TOOL (PPET)

The manual analysis of privacy policies has uncovered vital results, as dealt with in Chapter 4. The summary of the analysis includes various facts and issues related to the privacy policy. Those facts need to be visualized using graphical and pictorial format. privacy policies are opaque, complicated, verbose and inscrutable 'and clearly written for lawyers and not consumers. The visualization helps to grasp difficult contents of the privacy policy. Visualization of privacy policy content helps the users to see visual analytics before accepting or rejecting the privacy policy of service providers. This chapter discusses a visualization tool called Privacy Policy Elucidator Tool (PPET) which is developed as part of research presented in this thesis.

As a part of publication, the PPET and related work has been published as a book chapter (Section 1.7, A-1). This chapter is the core work of the thesis. The motivation to design the PPET tool, its architecture, and results have been published in the book chapter. However, the difference between the book chapter and this chapter include the section on motivation to design the tool (section 5.2), understanding the key sections of privacy policies based on privacy design principles (section 5.3), importance of the corpus (section 5.4- which was less explained in the book chapter). Also, this chapter includes the new sections like 5.7, 5.8, 5.9, 5.10.3. This chapter also include the discussion on testing and checking the performance of the PPET, which was not included in the book chapter.

5.1. INTRODUCTION

The information and its permission are listed before any app gets installed. The way of listing information and permission is not appropriate as it is not accessible, and precise. Hence, the users cannot understand the purpose and behavior of the app (Nauman, Khan, & Zhang, 2010),(Rosen, Qian, & Mao, 2013).

As per the country's laws (The European Parliment and The Council of the EU - GDPR, 2016), (The Gazette of India, 2009) and government agency like FTC (Julie & (FTC-US), 2015), every service provider should inform Internet users how users' information is handled after its collection. There exist several methods/ways to convey users about information handling strategies, like privacy policies are published on websites or hard copies are available of the privacy policy. However, certain challenges need to be discussed in detail.

There have been many cases of violating privacy laws and guidelines which resulted in a fine to service providers. The privacy policy of "Path" (a social networking app) didn't mention about their behavior in spite of collecting and storing users' information. As a result, in 2013, a fine of $800,000 was imposed by FTC (Julie & (FTC-US), 2015). Several such cases have been observed against companies (Federal Trade Commission, 2013)("CLIP- Center on Law and Information Policy," 2014) (C. Meyer, E. Broeker, A. Pierce, 2015).

5.2. CHALLENGES- NEED FOR A VISUALIZATION TOOL

Before assessing the usability of privacy policy, it is important to examine the contents of privacy policies and visualize to give the initial impression to Internet users on the strategies adopted by service providers. In this context, there is a need for a tool that will make the users aware about these growing privacy concerns to help to protect their privacy. This chapter will help to make the end user aware of the data handling policies of various websites/service providers.

From the state-of-the-art and as per the finding from our research (Dhotre, Olesen, & Khajuria, 2017),(Dhotre & Olesen, 2015), we identified the issues and challenges related to users' privacy. The privacy issues are serious concerns of Internet users. The summary of few issues and challenges are as follows (Dhotre et al., 2017):

- Users understanding and awareness about privacy:
  Users are unaware about the strategies of users' information management adopted by service providers. Users don't realize the value of privacy or they understand it only in its breach. Hence, the consequences of personal information analytics are not known clearly and completely to the users. So, the challenge is how to enhance the users' privacy and awareness.

- Understanding of privacy policies:
  The convoluted privacy policy contains jargons which become tough for users to understand. There is an increasing concern on the blind acceptance of privacy policies considering the strategies of companies on personal information management mentioned in it. Hence, the challenge is to represent the content of privacy policies in a simplified way.

- User control:
  The users cannot submit select data, thereby relinquishing complete control over their personal information. So, to avail the services, it is mandatory for users to share all the information solicited by service providers. Hence the challenge is to avail the service by disclosing only the minimum information necessary.

Considering the interaction between service providers and the users, there is a requirement for improving and enhancing usability. There are many privacy policy usability models available to evaluate the usability of privacy policies (Ghazinour, Majedi, & Barker, 2009). The formal, informal, and empirical are a few methods considered during evaluation of privacy policies. These methods compute usability measures based on thumb rules. One of the models called as Privacy Policy Visualization Model [PPVM] is presented, which is based on the evaluation measures such as clarity, completeness, transparency, etc.

After performing an extensive analysis of privacy policies as discussed in chapter 4, we deliberated on the contents of privacy policies, what and how the personal information is being handled, etc. Hence, the primary goal of this chapter is visualizing the contents of privacy policies in the easiest and simplest way.

The visualization can help the users to read the privacy policy in detail. The visualization of privacy policies' content will assist the users in deciding the transparency of the service providers towards users' personal information. This type of work will help to enhance users' privacy awareness and may act as a base for building trust between service providers and users.

The approach presented to develop a tool for classification and visualization in this chapter differs from previous tools in the following way:

1. A larger data set is studied containing over 50 privacy policies.
2. A detailed evaluation of information on personal identification.
3. Detailed categories and keywords to classify the privacy policy into sections.
4. The creation of the massive corpus (43,544 sentences) used for developing classifiers.
5. Use of effective visualization methods to represent key sections of privacy policies.
6. Summarized texts of the key section of privacy policies.
7. Sharing ratings of websites from Web of Trust (WoT).
8. This tool is a privacy awareness tool, not a privacy protection tool.

Our initial concern was to identify the privacy policies. Looking at the Indian users' privacy knowledge, awareness and privacy concerns (Dhotre & Olesen, 2015), we decided to consider the service providers in the Indian context.

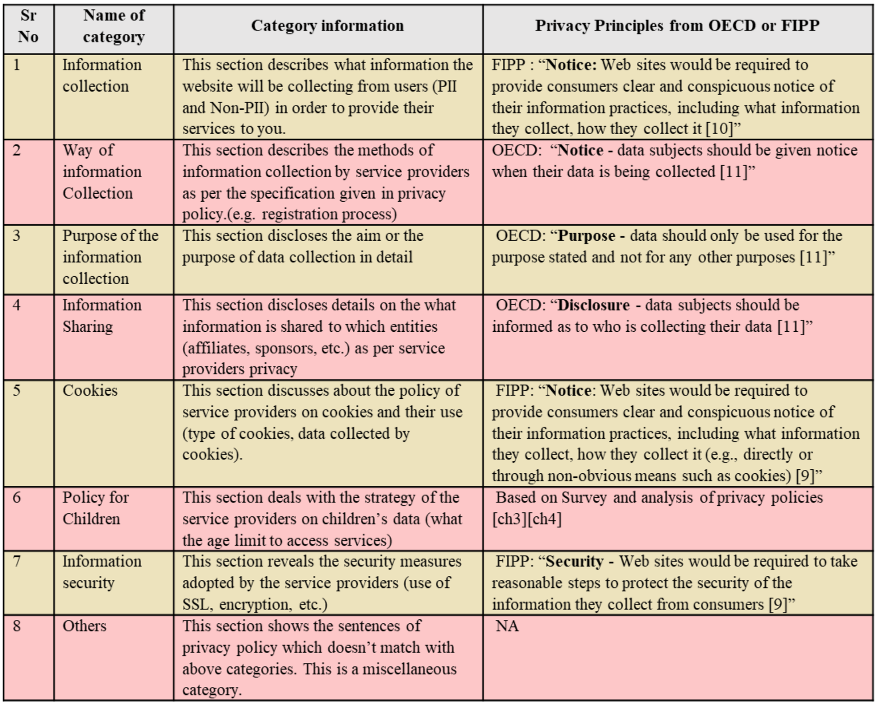
The OECD principles (OECD, 2013) or the GDPR in EU (The European Parliment and The Council of the EU - GDPR, 2016) aimed to provide best practices will serve as benchmarks to ensure that the practices of service providers are properly aligned to these principles and practices. It is important to understand the type of information collection, sharing, processing, etc. Hence, the goal is to visualize the contents of privacy policies by classifying the privacy policy into various and important sections.

It is also important to summarize the contents of the privacy policy of any website/service provider and represent it to the end user in an easily understandable form. The goals can be achieved by developing a tool as a browser extension or add-on. Therefore, we have developed a tool, PPET, which will classify and summarize the privacy policy contents automatically. Based on the analysis of privacy policies discussed in chapter 4 and machine learning, PPET reveals principal elements of privacy policies. So, classification and summarization of privacy policies are based on the contents of the privacy policy.

5.3. IDENTIFICATION AND DEFINITION OF PRIVACY POLICY CATEGORIES

The aim of PPET is to classify the privacy policy into important categories. The findings of our research mentioned in chapter 4 include the ambiguity of service provider's strategy on information collection and methods, cookies and sharing entities, security measures, readability, etc. This analysis has helped us to identify the key requirements to classify the content of privacy policy. Also, looking at the GDPR(The European Parliment and The Council of the EU - GDPR, 2016), Fair Information Privacy Information Principles [PFIPIP] from FTC in US(Julie & (FTC-US), 2015), Guidelines from OECD (OECD, 2013) , the privacy policy should have key categories which are, purpose, collection, cookies, share, security, retention, choice, access, and children policy.

So, PPET classifies the contents of privacy policies into one or more than one categories mentioned in fig 5-1.



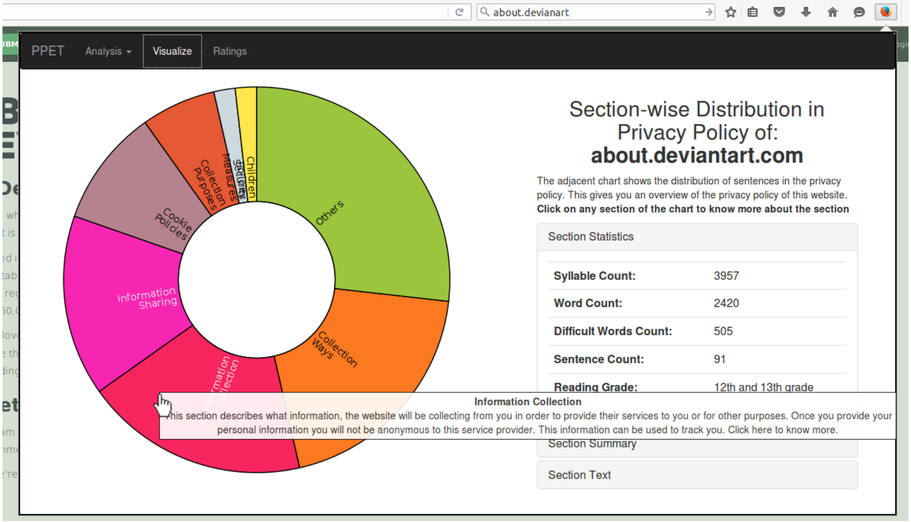
Sr No	Name of category	Category information	Privacy Principles from OECD or FIPP
1	Information collection	This section describes what information the website will be collecting from users (PII and Non-PII) in order to provide their services to you.	FIPP : "**Notice:** Web sites would be required to provide consumers clear and conspicuous notice of their information practices, including what information they collect, how they collect it [10]"
2	Way of information Collection	This section describes the methods of information collection by service providers as per the specification given in privacy policy.(e.g. registration process)	OECD: "**Notice** - data subjects should be given notice when their data is being collected [11]"
3	Purpose of the information collection	This section discloses the aim or the purpose of data collection in detail	OECD: "**Purpose** - data should only be used for the purpose stated and not for any other purposes [11]"
4	Information Sharing	This section discloses details on the what information is shared to which entities (affiliates, sponsors, etc.) as per service providers privacy	OECD: "**Disclosure** - data subjects should be informed as to who is collecting their data [11]"
5	Cookies	This section discusses about the policy of service providers on cookies and their use (type of cookies, data collected by cookies).	FIPP: "**Notice**: Web sites would be required to provide consumers clear and conspicuous notice of their information practices, including what information they collect, how they collect it (e.g., directly or through non-obvious means such as cookies) [9]"
6	Policy for Children	This section deals with the strategy of the service providers on children's data (what the age limit to access services)	Based on Survey and analysis of privacy policies [ch3][ch4]
7	Information security	This section reveals the security measures adopted by the service providers (use of SSL, encryption, etc.)	FIPP: "**Security** - Web sites would be required to take reasonable steps to protect the security of the information they collect from consumers [9]"
8	Others	This section shows the sentences of privacy policy which doesn't match with above categories. This is a miscellaneous category.	NA

Figure 5-1 Privacy policy key categories and principles of privacy (OECD, 2013)

This semi-automatic tool is intended to address the issues of privacy policies to assist the user in reading the privacy policy in the easiest way. This tool is developed so that it will work at user end without any setting or special arrangement.

5.4. GENERAL DESCRIPTION ON PPET

PPET is developed as a Mozilla Firefox extension and when it runs, the website is identified and the privacy policy contents are fetched, categorized as per fig 5-1 and effectively visualized for simpler reading. The output of PPET is represented in a dashboard. The dashboard contains information about the users PII collection in privacy policies considered during manual analysis. Also, the distribution of privacy policy text as per the defined categories (fig 5-1) is also represented in the form of donut chart. Last but not the least, this tool also shares the ratings of the websites obtained from WoT. This tool will assist the users to enhance the privacy by understanding the privacy policies. Hence, PPET will act as a facilitating tool for users to rate the websites based on the classification and summarization of the privacy policy text generated by PPET.



Section-wise Distribution in Privacy Policy of: **about.deviantart.com**

The adjacent chart shows the distribution of sentences in the privacy policy. This gives you an overview of the privacy policy of this website. **Click on any section of the chart to know more about the section**

Section Statistics

Syllable Count:	3957
Word Count:	2420
Difficult Words Count:	505
Sentence Count:	91
Reading Grade:	12th and 13th grade

Information Collection

his section describes what information, the website will be collecting from you in order to provide their services to you or for other purposes. Once you provide your personal information you will not be anonymous to this service provider. This information can be used to track you. Click here to know more.

Section Summary

Section Text

Figure 5-2 Dashboard of PPET (Khajuria, Sørensen, & Skouby, 2017)

Fig 5-2 represents the overall front end of PPET. There are mainly three tabs. The first label is ‘analysis’, where the users can view the list of personal attributes and the information on collection by service providers. For example, the number of websites that collect personal information like first name, last name, location, etc.

The second tab is ‘Visualize’, where a donut represents the way how the contents of the privacy policy are separated and distributed among categories. Each section of the donut represents the related text from the privacy policy of the current website. The last tab is ‘Ratings’, where the trust values (reputation) of the website are represented. The WoT[60] is a reputation and review service that helps the users to understand trustworthiness of the website based on users’ feedback.

To the right of donut chart is a set of tables. Statistics about the privacy policy document is represented followed by ‘Section Summary’. Here the contents of the key sections are summarized and displaced for quick reading.

The last part of the table is ‘section text’, where the actual contents of the privacy policy are displayed as the key section is clicked by the user.

The text of privacy policies is extracted and parsing is performed on it. Further, the parsed text is classified into key sections (8 sections). Machine learning and text

[60] *Mywot.com*

classification have been used to achieve a classification of the parsed text. In the development of PPET, a linear algorithm is used for building various classifiers.

5.5. NEED FOR CORPUS: PPET PREREQUISITE

Considering classification and summarization as two major tasks of PPET, it was important to use recent advancements of Natural Language Processing (NLP) to handle service provider’s website's existing natural language privacy policy. An NLP based tool like PPET will help users to empower themselves without any additional cooperation from service providers.

To develop PPET, it is vital to understand the tasks and related requirements. Initially, for the classification, the most important requirement is to design a machine learning model that will decide whether a document is a privacy policy document or not. To do this in the machine model, there is a need for strong, uniform corpus of privacy policies.

The are several structures of privacy policies (Dhotre, Olesen, & Samant, 2016). Some of them are presented in the form of bullet points, some are discussed in the form of Q&A form. Few websites have divided their privacy policy contents into various sections, other privacy policies are just a colossal text. Furthermore, each website provides different services to the users whose different data are involved in accessing those services. Hence, there is a basic need to have uniform structural data. Also, there is a need for a labeled corpus of privacy policy (sentences on information collection, sharing, security, etc.)

A corpus of over 100 privacy policies is available for research and other uses (Wilson et al., 2016) However, this corpus will help in identification of privacy policy, but not for the classification of privacy policy text into key sections. Also, summarization would be quite difficult using this corpus.

Hence, in this research, we identified over 60 privacy policies that are popular in India[61]. To create a corpus, each line of the privacy policy is extracted and labeled according to 8 key sections described earlier in this chapter. In this way, we have designed our own corpus of 43,544 sentences inclusive of all categories. This huge set of sentences will help to develop 8 classifiers to classify the privacy policy text into 8 key sections.

Tab-Separated Values (TSV) format was used to store the sentences of each key category. Each TSV file contains two columns; the first column encompasses all the sentences and the second column is the label (category) of those sentences. Apart from 8 classifiers, an additional classifier is designed to clean the input text given to each

[61] http://www.alexa.com/topsites/countries/IN

classifier. The unwanted text from the web pages is removed like the header and footer parts.

5.6. BUILDING PPET

Once the user visits a website, the privacy policy is fetched and its contents are visualized using PPET. Fig 5-3 illustrates the architecture of PPET where the analysis, visualization components of PPET are shown in detail. Fig 5-3 shows the architectural design of PPET. It mainly consists of three parts:

1. A user: The architecture consists of a user who interacts with the service provider, as well as the PPET. The user can visit any website like Twitter, Flipkart, LinkedIn, etc. If the user wishes to see the privacy policy in a simplified way, the user interacts with the developer too.

2. A service provider: Here, the second component is a service provider. Having said before, the PPET is for users, hence the job of the service provider is limited.

3. The PPET: Last but not the least, the tool interacts with the user as well as service providers. The tool will fetch the privacy policy document from the service provider. Once that document is pre-processed, then PPET uses the machine learning algorithm to classify and summarize the contents of the privacy policy and visualize it at the client side.

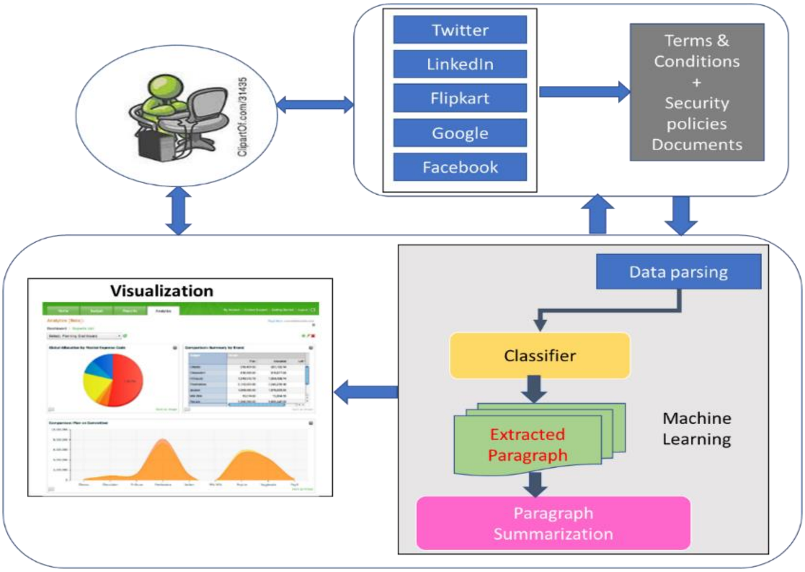


Figure 5-3 PPET Architecture.(Khajuria et al., 2017)

The job of a crawler in the PPET is to search and fetch privacy policies of several websites from various sectors. The policies are parsed and given to the classifiers. The trained classifiers will classify the sentences into the category as per its type. Once the classification is done, then the paragraph summarization will summarize the contents and store in the database. Then the summarized contents will be fetched from the database and visualization at PPET interface will take place at the user side.

The behavioral and structural view of the PPET can help the reader to understand the developed tool. Using Unified Modelling Language, the boundary, structural view and behavior of the PPET can be easily seen.

The following section discusses swim lane diagram and component diagram.

5.7. THE FLOW DIAGRAM OF PPET

It is useful to know what kind of information will be input and output from the PPET. The information about the timing of processes or information about processes that operate in sequence or in parallel is also available.

A swim lane (or swim-lane diagram) is a visual element used in the process flow diagram or flowcharts, that visually distinguishes task sharing and responsibilities of the process of PPET. Fig 5-4 represents swim lane of PPET.

The lanes are the columns that depict the actions separated from entities/elements. The entities are the column names that perform a set of actions. The elements are a user, Firefox extension, database, classifiers, etc. Here, the responsibilities of each element of PPET are much clearer than a flowchart. The start and end of PPET are illustrated.

Initially, the user visits websites and requests PPET tool (Firefox extension) to show classified and summarized privacy policy of the visited site. The PPET tool identifies the website and fetches the privacy policy from the database. This policy document is classified (Paragraph Classifier) and summarized (Paragraph Summarizer). This process is carried out during the machine learning model development phase. Further, the classified and summarized paragraphs from the database are shared with the user through PPET tool (Firefox extension).

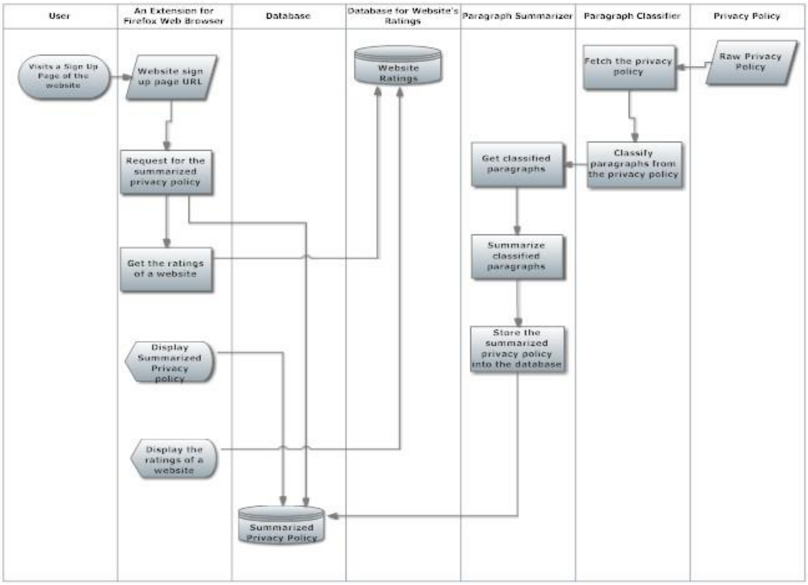

User
An Extension for Firefox Web Browser
Database
Database for Website's Ratings
Paragraph Summarizer
Paragraph Classifier
Privacy Policy
Visits a Sign Up Page of the website
Website sign up page URL
Website Ratings
Fetch the privacy policy
Raw Privacy Policy
Request for the summarized privacy policy
Get classified paragraphs
Classify paragraphs from the privacy policy
Get the ratings of a website
Summarize classified paragraphs
Store the summarized privacy policy into the database
Display Summarized Privacy policy
Display the ratings of a website
Summarized Privacy Policy

Figure 5-4 Swim lane diagram of PPET

Component diagram of UML represents the components, their organization, and connections between them. Fig 5-5 illustrates the component diagram of PPET.

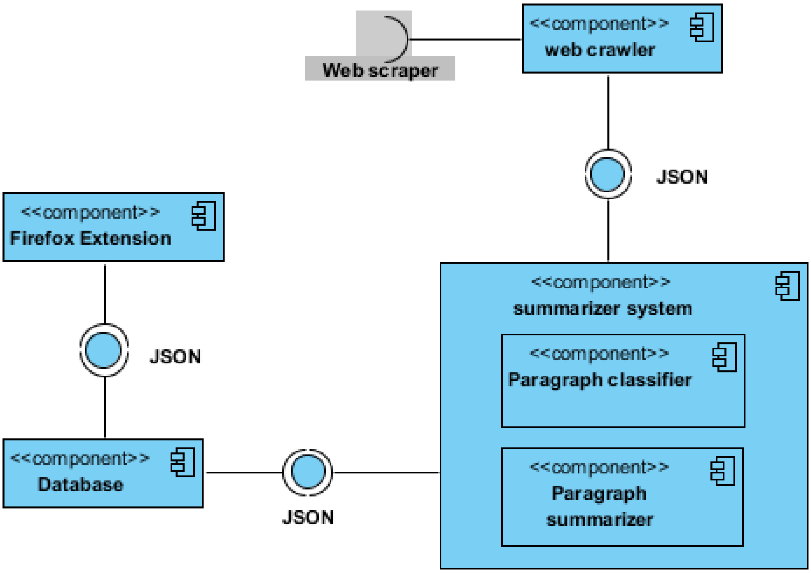


Figure 5-5 Component diagram of PPET

It consists of main components as:

- Web Scraper: The web scraper will be responsible for returning the privacy policy's text.

- Web crawler: Web crawler will be responsible for crawling the web, looking for privacy policies of different websites.

- JSON: It is JavaScript Object Notation, data-interface format, an industry standard machine-readable data exchange format. It is based on the JavaScript language.

- Summarizer system: Summarizer system is the main component of this research which will be responsible for generating the summarized privacy policy using machine learning algorithms and natural language processing.

  It will consist of two components:
  Paragraph classifier: Paragraph classifier will decompose the privacy policy into the different paragraphs.
  Paragraph summarizer: Paragraph summarizer will summarize the main ideas of paragraphs.
- Database: MySQL database has been used to store the classified and summarized privacy policy contents. The extracted paragraphs will be stored in a document in a collection. Each document will contain a key-value pair. The key in the document will be the key components of the privacy policy based on which the paragraphs were classified. The value will be the summarized paragraph.

- Firefox Extension: Firefox extension will be the main component responsible for interaction with the end-user.

5.8. THE DESIGN CONSTRAINTS OF PPET

The PPET was developed assuming certain limitations on the conditions under which the PPET is developed. The design constraint includes:

- Platform: The PPET is developed for the Mozilla Firefox web browser primarily. Hence, the user must have Mozilla Firefox web browser installed on their system.

- A number of websites covered: The PPET is developed considering the websites to which Indian users visit most frequently, it will work on limited websites initially. The scope of this project can be expanded further.

- Scalability: The number of users supported by the system is directly proportional to the frequency of requests to the database server. An increase in the number of requests within a unit of time over a threshold may severely mar the performance of the system.

5.9. THE DATA DESIGN OF PPET

The major data objects that the PPET is handling and manipulating is the textual data. The data is scraped from the websites' privacy policy. This data is classified and condensed via automatic summarization algorithms. The condensed data that will be obtained is further stored in the database.

The relation between several data objects observed during the design of PPET. Data objects and their major attributes and relationships among data objects are described in fig.5-6. It consists of a user, website, privacy policy, database server, Firefox extension, etc.

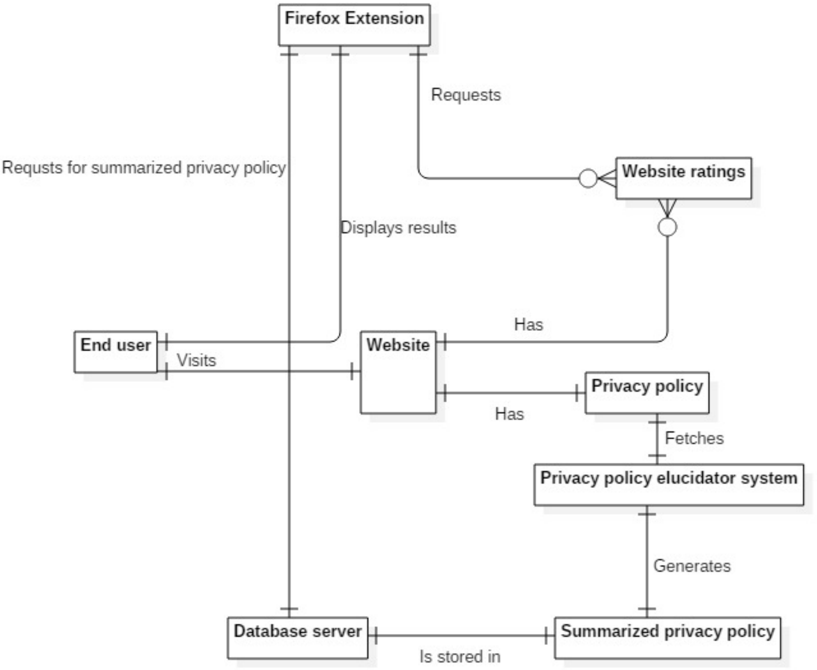


Figure 5-6 Data design of PPET

A user visits a website that has a privacy policy. The text of privacy policy is the major data object of PPET. The database server is updated once the PPET generates classified and summarized contents of the privacy policy. Furthermore, Firefox extensions request the rating of the websites and the summarized privacy policy is displayed on the users' side as a part of the results.

5.10. THE MACHINE LEARNING MODEL OF PPET

In the architecture of PPET, the pre-processing of the privacy policy is needed for several purposes. One of the purposes is to remove the needless text of the privacy policy. Hence, it is crucial to convert a raw text from the corpus to well-conditioned data. This will help the model to work optimally. Unnecessary features/word take up space, time and increase complexity.

In this research, the preprocessing techniques contain sentence tokenization, removal of stop words and proper nouns, stemming, transformation (achieving a set of useful words), and lemmatization. This bunch of techniques helps to trim the contents of privacy policies into meaningful parts. Based on the meaning of the parts of the privacy policies, the feature selection method will be simplified and effective.

To understand the relation of preprocessing, feature selection and classifier in the process of classification and summarization, a learning model is developed. Fig 5-7 represents two main parts of learning model i.e. Training and testing.

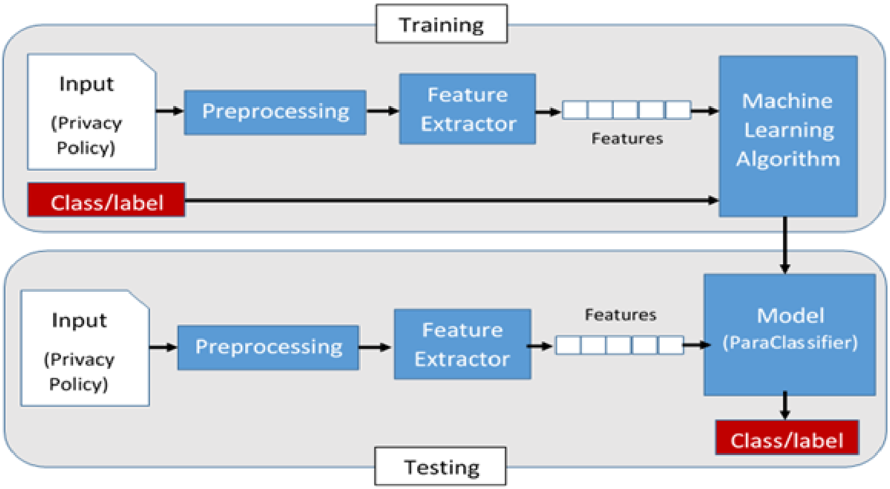


Figure 5-7 Machine learning model of PPET (Khajuria et al., 2017)

In the training phase, the privacy policy is preprocessed (stemming, lemmatization, tokenization) to extract the best possible features. For the training, the pre-processed text of privacy policy a corpus of 43,544 sentences is used to create a learning model. In the testing phase, the unclassified text (new privacy policy) is preprocessed to get some features. Using a probabilistic classification model, the privacy policy text is labeled.

5.10.1. CLASSIFICATION

Data mining (DM), Information retrieval (IR) and Machine learning (ML) are research fields which are considered highly sought-after areas of research for developers, academics, and researchers. However, there is a difference between data mining, information retrieval, and machine learning.

Massive amounts of information are gathered nowadays due to wireless and mobile communication. The aim is to extract the required data and convert it into an understandable and readable format. The process of extracting user information or identifying a useful pattern in the large set of information is called as DM[62]. Data mining involves concepts from various fields like machine learning, database systems, and statistics. Sometimes DM is also termed as Knowledge Discovery in Databases (KDD).

The task of DM encompasses automatic or semi-automatic analysis of vast amounts of data to find useful, interesting, and formerly unknown patterns. Cluster analysis (a group of records), anomaly detection (uncommon records), classification (categorization of data), and pattern mining (dependencies) are a few examples of tasks of DM (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The useful findings (like patterns) are a summary of huge data and may be useful in doing further analysis like predictive analytics. Data collection, preprocessing, result interpretation and visualization are the parts of DM.

As mentioned earlier, DM is the discovery of non-disclosed patterns. Hence, the result of DM is not the subset of documents or information sources. The goal of data mining is a prediction.

Information retrieval is the process of searching for some information from a collection of information sources. The process of IR is activated, once the user inputs a query. The result of the query is an object or set of objects depending on the level of relevancy. The result is the part of the documents or information sources. The goal of IR is to retrieve information accurately.

ML provides the ability for computers to learn themselves without explicit programming (Munoz, n.d.). In ML, the algorithms learn from given data and make predictions. Email filtering, data breach, optical character recognition, computer vision are a few applications of ML.

One of the tasks under ML and IR is text classification (TC). It is defined as automatic sorting of texts into a predefined set of categories (Marsland, 2009). There are several text classification approaches. Following are a few algorithms based on Statistical approach.

[62] https://en.wikipedia.org/wiki/Data_mining, 2017

1. Naïve Bayes Classifier:
   This technique of classification is based on Bayes' Theorem, with the strong assumption of independence among the features. In simple words, it assumes that the presence or absence of any feature is independent of the presence or absence of any other feature. The name is 'naïve' because of the features or properties are strongly independent of other features[63].The Naïve Bayes classifier is a probabilistic classifier and is a model based on a conditional probability.

   If a feature set is represented as[64]:

   $$x = \{x1, x2, x3, \ldots\ldots xn\}$$

   Then, conditional probability (instance probability) is represented as:

   $$P(C_k\,|x) = \frac{P(C_k)P(x|\,C_k)}{P(x)}$$

   The same equation in a simplified way is rewritten as:

   $$\text{posterior}= \frac{\text{(prior x knowledge)}}{\text{evidence}}$$

   Looking at the above equations, we can say that the denominator is independent of C, and the value of features. The Naïve Bayes classifier is simple and fast to classify the test data set. This is also clear that the training data required is less as it holds no dependency.

2. Decision Tree:
   This is a predictive model that helps to expand decisions along with its consequences. The combination of features that lead to classification at leaf level. Unique classification is represented by a leaf. Hence, each feature is an internal node of a decision tree. There are several types of decision trees (Bhatia & Jaiswal, 2016).

3. Support Vector Machine (SVM):
   This is a supervised algorithm that identifies the hyperplanes that help to divide given data sets into possible classes. Here the aim is to select hyperplane with the best possible distance between any data point and the hyperplane itself. This will provide a good chance of testing data being categorized correctly (Marsland, 2009).

[63] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[64] https://en.wikipedia.org/wiki/Naive_Bayes_classifier#cite_note-7

4. K Nearest Neighbors (KNN):
   This is the most accepted algorithm for classifying the patterns. Here, the algorithm starts with the selection of random data. During the learning phase, each data set is assigned to a cluster based on the Euclidean distance. The center of clusters is adjusted to the average of acquired data points. This algorithm is intended to find k nearest neighbor. The performance of KNN depends on the value of K and distance matrix. The comparison of statistical approaches is represented in table 5.2 (Brindha, Prabha, & Sukumaran, 2016).

Algorithm	Basic techniques	Strengths	Weaknesses	Highest accuracy
Naïve Bayes	This is a probabilistic based Supervised algorithm	Simple, independence of assumptions, widely accepted	Easily affected	99.99
Decision Tree	This is graph based supervised algorithm	Simple, handles missing data	Expensive computability	91.17
SVM	This is a Hyperplane based supervised algorithm	No need of large dataset calculation	Needs long training time	98.10
KNN	This is a neuron based unsupervised algorithm	Easy to implement	Difficult to handle noisy data	99.65

Table 11 Comparisons of statistical approach for classification (Brindha et al., 2016)

Based on the research mentioned in this section and table 11, we decided to select Naïve Bayes classifier to classify the contents of privacy policies.

Pre-processed privacy policy (bag of words) is converted into the feature set using feature extractor available by default in the Naïve Bayes classifier. The feature vector is a set of tokens. Basic information about the privacy policy is included in the feature sets. The classification of privacy policy text is carried out in 8 key sections using feature sets. The feature sets and labels are given as input to the classifier (ParaClassifier Model).
Mathematically, the ParaClassifier model is represented as:

Initially, the system (S) mathematically consists of three parameters:

$$S = \{I, O, P\} \qquad (1)$$

Where ‘I’ represents the input, ’O’ represents the output, and ’P’ consists of a set of operations needed to convert the input into output. In the PPET, the input is the privacy policy (a text file). Each sentence of the privacy policy is a labeled data. Using labeled data, the texts are classified into distinctive features.

Mathematically, the features (F) are represented as:

$F = \{F_1, F_2, F_3, \dots \dots, F_N\}$ (2)

These features are further classified into 8 key sections or classes (C):

$C = \{C_1, C_2, C_3, \dots \dots, C_k\}\,; \text{whereK} = 8$ (3)

The set *C* represents the classes where texts of the entire privacy policy are classified. Hence, the mapping function can be visualized as:

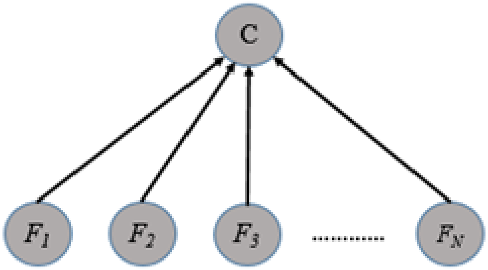


Figure 5-8 Mapping of features into classes using Naive Network (Khajuria et al., 2017)

The goal can be achieved using the Naïve Bayes classifier. This probabilistic based model considers that any feature $F_i \in F$ is independent of other features $F_k \in F$. This will help mapping the value to a class C as mentioned in Fig 5.8.

There might be a possibility that the features may belong to more than one class. For example, a naïve classifier for two features $\{F_1, F_2\}$ can be defined as following set of structures like

$$\{F_1 \rightarrow C_1\}, \{F_2 \rightarrow C_2\}, \{F_2 \rightarrow C_5\}, \{F_3 \rightarrow C_4\}, \{F6 \rightarrow \phi\}, \dots \dots \{F_1 \rightarrow C_8\}$$

In short, the classification model (C_M) is represented as:

$$C_M : F \rightarrow C$$

Once the contents are classified, the next job of PPET is summarization.

5.10.2. SUMMARIZATION

The process of shortening the text from the given document is called as text summarization. Even though there exist two main methods for text summarization (manual and automatic), the aim is common i.e.to generate a summary of the main points that describe the complete document. Automatic summarization generates a subset of data that contains all essential information of the entire document.

Automatic summarization comes under the fields of data mining and machine learning. The purpose of summarization is to create a short document from the original document and use it in a variety of contexts like news headlines, biographies, reviews, etc. (Bhatia & Jaiswal, 2016). There are several algorithms available to summarize the contents of the privacy policy. The list includes Term frequency based method, Graph-based method, Time-based method, a Clustering method, Semantic dependency based method, Topic-based approaches, Discover based approach, Latent semantic analysis (LSA), etc.

Term Frequency inverse document frequency (TF-IDF) was introduced in 1989. This is a numeric static method which helps to ascertain how important a word is to a document (Bhatia & Jaiswal, 2016). The weight of the word increases as the number of times the same word appears in the document. This is a ratio of a number of times the word appears in the document to the total number of words in the document. It is represented in mathematical form as:

$$\text{TF(w)}=\frac{number\ of\ times\ the\ word\ w\ appears\ in\ a\ document}{Total\ number\ of\ words\ in\ a\ document}$$

This equation gives the score of a word in the document. The sentence score is illustrated by calculating the related words in the sentences. This method is easy to calculate basic metric containing words and their weights. Considering the research mentioned in this thesis, for summarization, the TF-IDF is a set of words and hence, it doesn’t consider semantics.

In the PPET, the summarization method is based on a frequency count of a word/token that appears in a sentence. The sentence is more important and ranked on top if a token appears frequently in the sentence. The tokenization method takes all possible tokens generated for each sentence. The tuple contains a class/label and a token. Hence, the tuple of tuples will be created and is represented as:

$$T[i][j] = \begin{bmatrix} S_1W_1, C_1 & S_1W_2, C_2 \cdots & S_1W_m, C_3 \\ S_2W_1, C_3 & S_2W_2, C_8 \cdots & S_2W_n, C_4 \\ \vdots & \vdots & \vdots \end{bmatrix} \qquad (5)$$

Looking at the 2D matrix above, the values shown are a combination of token and its concerned class/label. For example, a pair, (S_1W_1, C_1) denotes the word W_1 from sentence S_1 which fit in class C_1. The size of T is not fixed and it depends on the total sentences and tokens in each sentence.

In the PPET, the paragraph summarizer is based on frequency summarizer which tokenizes the input into sentences and then calculates what is called the frequency map of the words.

$$freq[w] = freq[w] + 1\ ;\text{for all words } w \text{ in } S \qquad (6)$$

When the frequency dictionary is completed, the maximum frequency is identified. The ranking of the sentence is done using following formula for each word *w* of sentence *i* in text and is mathematically expressed as:

$$ranking[i] = ranking\ [i] + freq\ [w] \qquad (7)$$

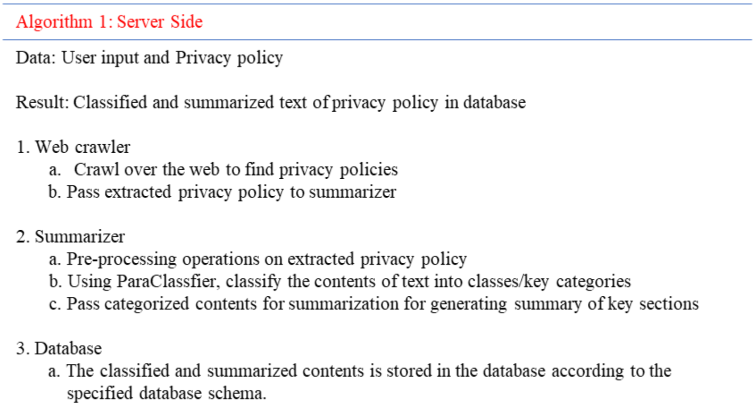
Depending on the word frequency, the sentences are ranked and final summary is generated that contains top sentences. For summarization, the input is the words from a tokenized sentence. Frequency dictionary is the output of the summarizer. i.e. The word, its frequency, and ranking of sentences. Therefore, summarization model S_M is mathematically expressed as:

$$S_M : [words\ (w), class\ (c)] \rightarrow ranking \qquad (8)$$

It is to be noted that the integration of paragraph summarizer is done and then stored in the database. This is privacy policy summarization.

5.10.3. ALGORITHMS OF PPET (CLIENT AND SERVER)

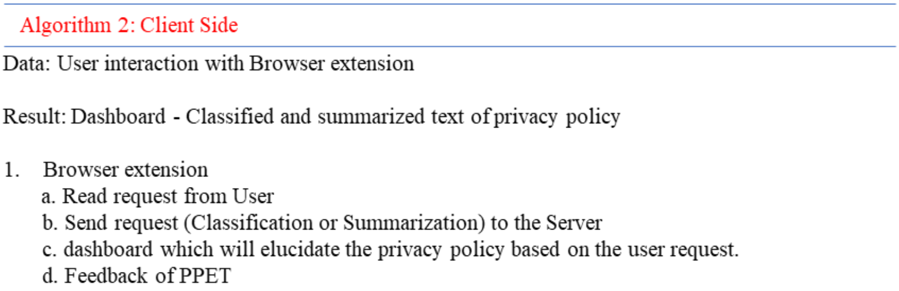
Based on the discussion of PPET, architecture, machine modeling, the algorithms of PPET of the client as well as of the server side is represented as below. Fig 5.9 represents server algorithm which consists of three main elements; the crawler, summarizer and database.



```
Algorithm 1: Server Side

Data: User input and Privacy policy

Result: Classified and summarized text of privacy policy in database

1. Web crawler
     a.  Crawl over the web to find privacy policies
     b. Pass extracted privacy policy to summarizer

2. Summarizer
     a. Pre-processing operations on extracted privacy policy
     b. Using ParaClassfier, classify the contents of text into classes/key categories
     c. Pass categorized contents for summarization for generating summary of key sections

3. Database
     a. The classified and summarized contents is stored in the database according to the
        specified database schema.
```

Figure 5-9 Server-side algorithm

Similarly, the client algorithm is represented in fig 5-10. The client algorithm consists of request and response during the interaction with PPET.

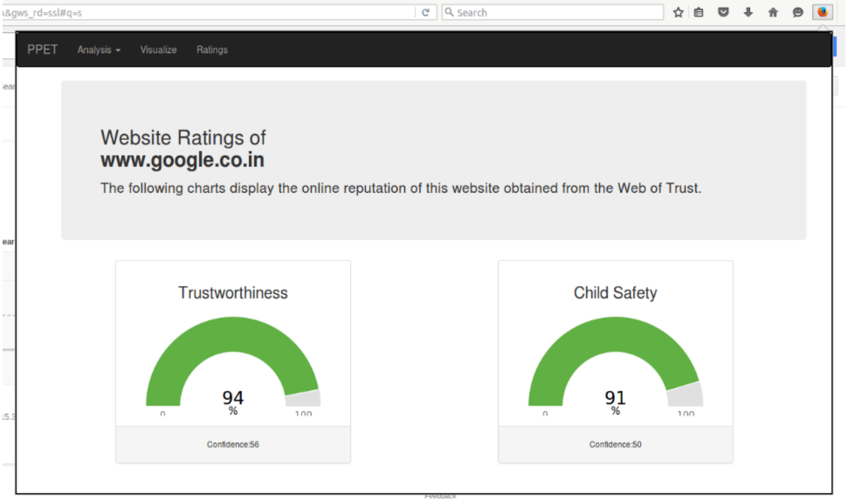


```
Algorithm 2: Client Side
Data: User interaction with Browser extension

Result: Dashboard - Classified and summarized text of privacy policy

1.  Browser extension
    a. Read request from User
    b. Send request (Classification or Summarization) to the Server
    c. dashboard which will elucidate the privacy policy based on the user request.
    d. Feedback of PPET
```

Figure 5-10 Client-side algorithm

5.11. RESULTS OF PPET

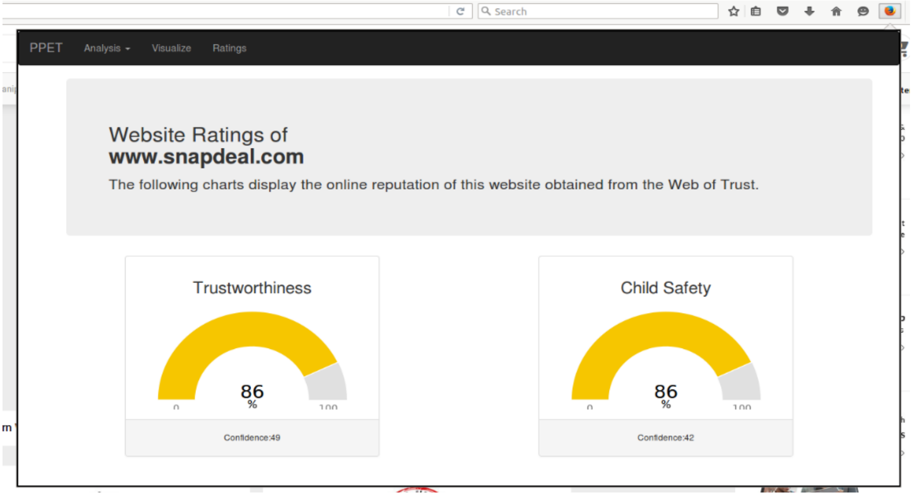
During the implementation of PPET, we have targeted over 60 policies of well-known and widely used websites in India. The PPET identifies the links of these policies by accessing the database. Policies are analyzed by employing machine learning approaches to understand various aspects of policies, including the language used, readability, structures and aim of privacy policies, information collection, use of cookies, the method of data collection and its purpose, personal information sharing parties, sharing of personal information, and security measures. After analysis of policy, PPET collects qualitative and descriptive information about the policy. PPET visualizes collected data as an output using the interactive donut chart for a better understanding of the user. The PPET uses Web of Trust (WoT)-API- C3 Gauge widget to retrieve ratings and trustworthiness of any website.



Website Ratings of
www.google.co.in

The following charts display the online reputation of this website obtained from the Web of Trust.

Figure 5-11 Add-on in action: Panel showing the ratings of www.google.co.in.(Khajuria et al., 2017)



Website Ratings of
www.snapdeal.com

The following charts display the online reputation of this website obtained from the Web of Trust.

Figure 5-12 Add-on in action: Panel showing the ratings of www.snapdeal.com (Khajuria et al., 2017)

Illustrative output for policies of Google (highlighted by green color) and Snapdeal (highlighted by yellow color) are discussed here. Fig 5-11 visualizes rating of the Google website in India (www.google.co.in) for trustworthiness and child safety rated more than 90%. whereas Fig 5-12 visualizes rating of Snapdeal websites in India (www.snapdeal.com) for trustworthiness and child safety which is moderate i.e. in between 70% to 90 %. This visualization clearly demonstrates that users trust Google more and believe that Google data practices and policies are safer and more inclined to be concerned about children than Snapdeal.

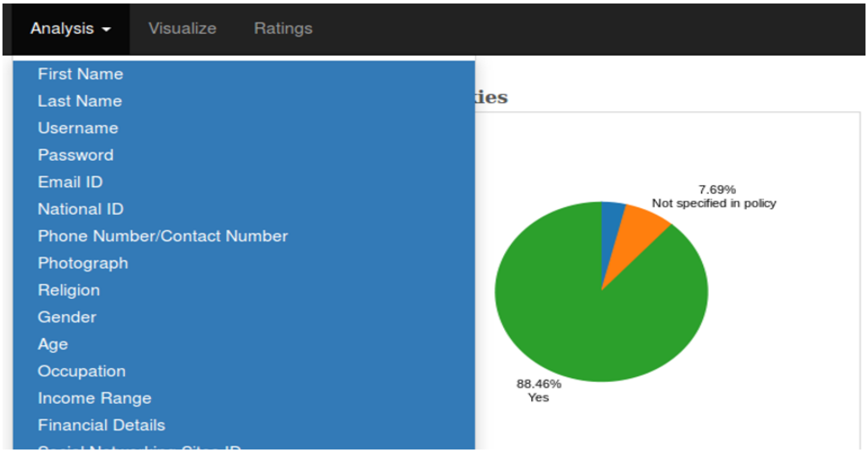


Figure 5-13 Add-on in action: Panel showing the list of attributes (Khajuria et al., 2017)

PPET is able to provide additional analysis and interpretations by accessing our previous work as Firefox extension which is aimed to alert the user about what user attribute data is being collected and (mis) used by web services. Taken altogether such collected information is conveyed to users by considering 78 attributes. As shown in fig 5-13, the various attributes include attributes for personal identification as the first name, last name, address, etc. and for non-personal attributes like the operating system, browser information, etc.

In this research study, the list of personal attributes and its analysis can be easily seen. Out of 72 information of the user, fig 5-14 represents the service provider's strategy on the use of cookies. Almost 88% of the total service providers in this study clearly stated that they use cookies. Further, there are only a few websites (4%) who do not use cookies. But, some websites (8%) have not stated about cookies in their privacy policy. This implies that cookies are a prominent way to collect users' information. Hence, user awareness will be enhanced by reading the privacy policy section on cookies to know what type of cookies, and what information is being gathered by service providers.

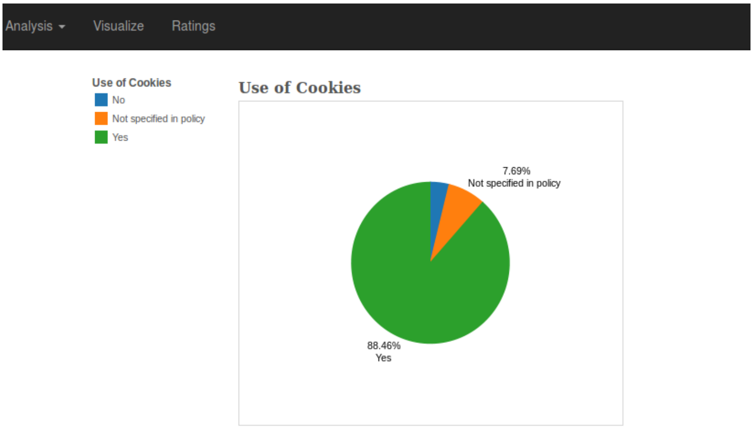
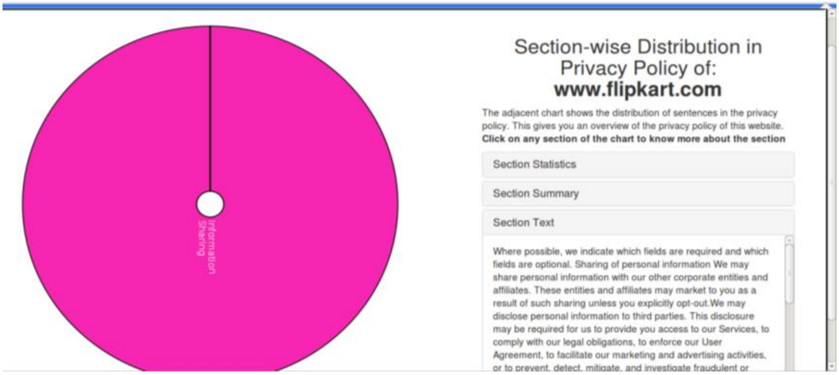


Figure 5-14 Add-on in action: Panel showing the use of cookies among the privacy policies analyzed in this study.(Khajuria et al., 2017)

The PPET provides information on such vital attributes through the dashboard to know the standard practices for information collected by service providers. The contribution of PPET is further depicted in Fig 5-15. The key 8 sections of the privacy policy are represented in the form of a donut chart. Use of sections makes it possible for the users to choose to read a particular section of their interest. The user can select one section of donut chart to see the information about that section. Once clicked, the right side of the donut will provide the requested information in a table. For example, Fig 5-15 represents an information sharing section of the privacy policy of Flipkart.com, a frequently used e-commerce website.



Section-wise Distribution in Privacy Policy of: **www.flipkart.com**

The adjacent chart shows the distribution of sentences in the privacy policy. This gives you an overview of the privacy policy of this website. **Click on any section of the chart to know more about the section**

Section Statistics

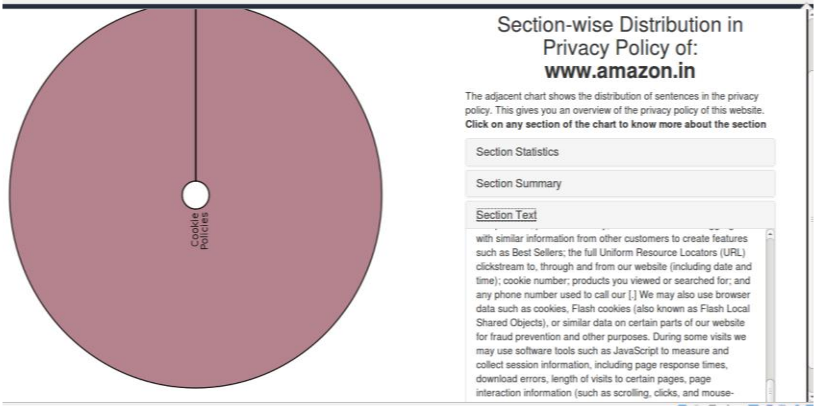
Section Summary

Section Text

Where possible, we indicate which fields are required and which fields are optional. Sharing of personal information We may share personal information with our other corporate entities and affiliates. These entities and affiliates may market to you as a result of such sharing unless you explicitly opt-out.We may disclose personal information to third parties. This disclosure may be required for us to provide you access to our Services, to comply with our legal obligations, to enforce our User Agreement, to facilitate our marketing and advertising activities, or to prevent, detect, mitigate, and investigate fraudulent or

Figure 5-15 Add-on in action: Panel shows the user interested section of the privacy policy (Information Sharing from Flipkart.com) (Khajuria et al., 2017).

The size of sectors (or sections) in the donut chart discloses more information about the key sections. The size of sectors represents the number of sentences identified from the privacy policy. If the size of a sector is bigger than another, then it indicates that there are more number of sentences related to that section in the privacy document compared to other section. For instance, if the size of third-party sharing section is bigger than the security, then it shows that the number of sentences on sharing information is more than security. This gives additional information that the service providers have given less information on users' information security in the privacy policy. Therefore, the PPET hopefully will help the users to know the strategy of service providers towards users' privacy and subsequently make a decision.



Section-wise Distribution in Privacy Policy of: **www.amazon.in**

The adjacent chart shows the distribution of sentences in the privacy policy. This gives you an overview of the privacy policy of this website. **Click on any section of the chart to know more about the section**

Section Statistics

Section Summary

Section Text

with similar information from other customers to create features such as Best Sellers; the full Uniform Resource Locators (URL) clickstream to, through and from our website (including date and time); cookie number; products you viewed or searched for; and any phone number used to call our [.] We may also use browser data such as cookies, Flash cookies (also known as Flash Local Shared Objects), or similar data on certain parts of our website for fraud prevention and other purposes. During some visits we may use software tools such as JavaScript to measure and collect session information, including page response times, download errors, length of visits to certain pages, page interaction information (such as scrolling, clicks, and mouse-

Figure 5-16 Add-on in action: Panel shows the summary of a section of privacy policy (cookies Policy of Amazon.in)(Khajuria et al., 2017)

Last but not the least, the contribution of PPET is summarization. Once the user clicks on any section, the related contents are fetched from the privacy policy. Considering the verbosity of privacy policy, the summarized text of privacy policy will help the users to read and understand quickly. Fig 5-16 represents the summarized text of cookies section. The advantage of the text summarization that it saves users' reading time. During summarization, the frequency count method helps find the prioritized sentences.

In addition, to the above contribution of PPET, various but interesting statistics are drawn regarding the text for each key category, for example, the sentence count, word count, the count of the difficult words and the reading ease.

5.12. TEST AND PERFORMANCE

For the evaluation of classifiers of PPET, the simplest matrix is used. If a privacy policy contains 'n' number of sentences, then a classifier should be accurately classifying those sentences according to their class label. An accuracy of a classifier is the percentage of sentences from the policy that the classifier has correctly labeled[65].

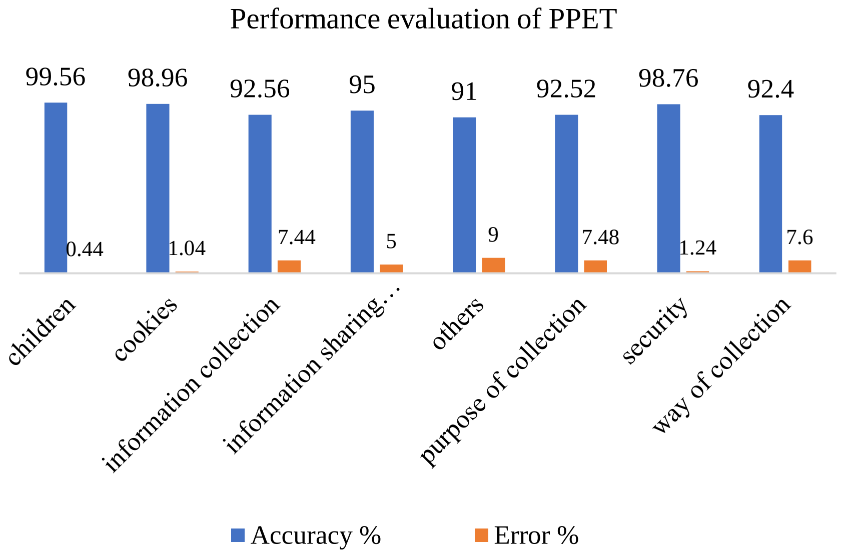


Figure 5-17 Performance evaluation of PPET

If there are 20 sentences on information collection in a privacy policy and a classifier predicts 15 sentences correctly then an accuracy is 15/20= 75%. Using this straightforward way of finding accuracy is used in this research work. We have used

[65] http://www.nltk.org/book/ch06.html#code-gender-features-overfitting

accuracy method from the *nltk.classify* package[66]. Having said before, a big corpus contains 43,544 sentences. This includes the sentences on information collection, sharing, security, cookies, etc.

The performance evaluation of PPET is tested on the big corpus which is able to correctly classify the sentences from the privacy policy. The fig 5-17 represents the variety of classes on the x-axis and the percentage of correct classification on the y-axis. On average the accuracy of PPET is around 95.09 %. Cookies and security are two top classes whose accuracy is close to 99%. The average accuracy considering all classes is 95%.

The error rate of each class is below 10%. Hence, the PPET is effective and has achieved performance at a high level.

5.13. REFERENCES

Bhatia, N., & Jaiswal, A. (2016). Automatic text summarization and it's methods - a review. In *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)* (pp. 65–72). IEEE. http://doi.org/10.1109/CONFLUENCE.2016.7508049

Brindha, S., Prabha, K., & Sukumaran, S. (2016). A survey on classification techniques for text mining. In *2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS)* (pp. 1–5). IEEE. http://doi.org/10.1109/ICACCS.2016.7586371

C. Meyer, E. Broeker, A. Pierce, and J. G. (2015). FTC Issues New Guidance for Mobile App Developers that Collect Location Data. *FTC.*

CLIP- Center on Law and Information Policy. (2014). Retrieved from https://www.fordham.edu/download/downloads/id/1867/privacy_enforcement_actions.pdf

Dhotre, P., & Olesen, H. (2015). A Survey of Privacy Awareness and Current Online Practices of Indian Users. In *Proceedings of WWRF Meeting 34, Santa Clara, CA, USA, Apr. 2015* (p. 10). WWRF. Retrieved from http://vbn.aau.dk/en/publications/a-survey-of-privacy-awareness-and-current-online-practices-of-indian-users(92c00b4f-a720-45b8-b3cb-9cfbffc7d4bc).html

Dhotre, P., Olesen, H., & Khajuria, S. (2017). User Privacy and Empowerment :

[66] [http://www.nltk.org/api/nltk.classify.html?highlight=classify#nltk.classify.util.accuracy

Trends , Challenges , and Opportunities. In *Springer*.

Dhotre, P., Olesen, H., & Samant, K. (2016). Interpretation and Analysis of Privacy Policies of Websites in India. In *WWRF Meeting 36*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. Retrieved from https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

Federal Trade Commission. (2013). Path social networking app settles ftc charges it deceived consumers and improperly collected personal information from users, mobile address books. *Federal Trade Commission, Tech. Rep. Case3:13-Cv-00448-RS*.

Ghazinour, K., Majedi, M., & Barker, K. (2009). A model for privacy policy visualization. *Proceedings - International Computer Software and Applications Conference*, *2*, 335–340. http://doi.org/10.1109/COMPSAC.2009.156

Julie, B., & (FTC-US). (2015). *U.S. Privacy Law*. Munich, Germany. Retrieved from https://www.ftc.gov/system/files/documents/public_statements/639991/150429munich.pdf

Khajuria, S., Sørensen, L., & Skouby, K. E. (2017). *Cybersecurity and Privacy - Bridging the Gap*. River Publishers. Retrieved from http://www.riverpublishers.com/book_details.php?book_id=434

Marsland, S. (2009). *Machine learning : an algorithmic perspective*. CRC Press.

Munoz, A. (n.d.). Machine Learning and Optimization. Retrieved from https://www.cims.nyu.edu/~munoz/files/ml_optimization.pdf

Nauman, M., Khan, S., & Zhang, X. (2010). Apex: Extending Android Permission Model and Enforcement with User-defined Runtime Constraints. *ASIACCS*, 328–332. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.193.2604&rep=rep1&type=pdf

OECD. (2013). OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data - OECD. Retrieved September 23, 2017, from http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

Rosen, S., Qian, Z., & Mao, Z. M. (2013). AppProfiler: A Flexible Method of Exposing Privacy-Related Behavior in Android Applications to End Users.

CODASPY'13. Retrieved from http://appprofiles.eecs.umich.edu/appprofiler.pdf

The European Parliment and The Council of the EU - GDPR. (2016). GDPR. *Official Journal of the European Union*. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

The Gazette of India. (2009). *Information Technology Act 2008,*. Retrieved from http://meity.gov.in/sites/upload_files/dit/files/downloads/itact2000/it_amendment_act2008.pdf

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., … Sadeh, N. (2016). The Creation and Analysis of a Website Privacy Policy Corpus. *Annual Meeting of the Association for Computational Linguistics*, 1330–1340. Retrieved from http://anthology.aclweb.org/P/P16/P16-1126.pdf

CHAPTER 6. PPET IMPACT AND RECOMMENDATIONS

This chapter consists of mainly two parts. In the first part of this chapter, the feedback on PPET is represented in detail. This includes a description of questions, participants', responses and its detailed analysis. The analysis reveals important findings like grading of the websites and level of usefulness of PPET. Further, the privacy policies and its compliance with privacy guidelines/ regulation are discussed. At the end of the first section, the limitations of PPET are specified.

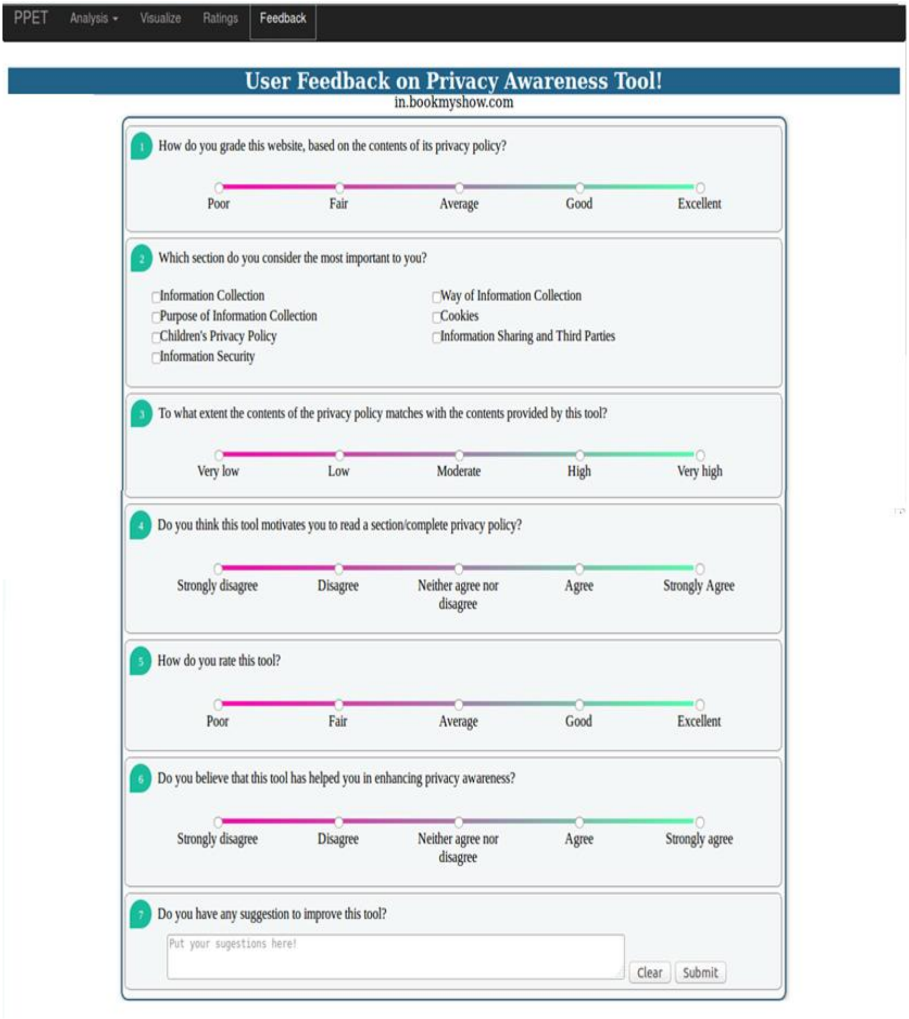
The second part of this chapter is the recommendations. The recommendations include a unified way (standard template) of representing privacy policy of any service provider. This template suggests the important sections that should be part of any privacy policy. Also, this chapter elaborates on the roles of a few entities responsible for privacy awareness. At the end of section part, the recommended measures are mentioned to improve information bargain and visualization.

6.1. WHY FEEDBACK OF PPET

As per the discussion in chapter 5, PPET is a semi-automated tool that provides classification, summarization, and visualization of the privacy policy's contents. This tool also gives a collection of PII by service providers using pie charts. In order to see the level of acceptance of this tool, we have collected feedback from Internet users in India. The evaluation of the PPET was based on the following important objectives:

- To receive an overall rating to the PPET
- To know the important contents/sections of privacy policies from users' perceptions
- To receive grading for websites based on users' understanding of privacy policies
- To know whether the PPET motivate users to read a section or entire privacy policies
- To know whether the PPET help the users to enhance the privacy awareness.

Considering the above objectives, the feedback form was designed and is as shown in fig 6-1. This form appears once the user clicks on the PPET interface. The PPET tool identifies the name of the website, and feedback will be gathered from that same website.



User Feedback on Privacy Awareness Tool!

in.bookmyshow.com

1 How do you grade this website, based on the contents of its privacy policy?

○ Poor ○ Fair ○ Average ○ Good ○ Excellent

2 Which section do you consider the most important to you?

□ Information Collection
□ Purpose of Information Collection
□ Children's Privacy Policy
□ Information Security
□ Way of Information Collection
□ Cookies
□ Information Sharing and Third Parties

3 To what extent the contents of the privacy policy matches with the contents provided by this tool?

○ Very low ○ Low ○ Moderate ○ High ○ Very high

4 Do you think this tool motivates you to read a section/complete privacy policy?

○ Strongly disagree ○ Disagree ○ Neither agree nor disagree ○ Agree ○ Strongly Agree

5 How do you rate this tool?

○ Poor ○ Fair ○ Average ○ Good ○ Excellent

6 Do you believe that this tool has helped you in enhancing privacy awareness?

○ Strongly disagree ○ Disagree ○ Neither agree nor disagree ○ Agree ○ Strongly agree

7 Do you have any suggestion to improve this tool?

Put your sugestions here!

Figure 6-1 Feedback Form of PPET

The feedback comprises mainly six questions which are mandatory for the users. The questions designed are in line with the objectives as mentioned above. We have used a five-point scale question. This will help not only to provide the detailed option but also to get the exact feedback from users. For example, the options for grading a website have five options like poor, fair, average, good, and excellent. The users' have a choice to share their comments if any. Hence, along with six questions, a separate comment/text box is provided.

"in.bookmyshow.com" is the name of the website for which the feedback form is shown in fig 6-1. To hide the identity of users, the feedback was collected without

collecting users' information. The feedback was retrieved from the database in excel sheet for the analysis. The detailed analysis is discussed in the following sections.

6.2. FEEDBACK METHOD/SETUP

It as important to identify the participants/respondents who can give feedback of PPET. For this, the researcher has selected a group of students and faculties. These students and faculties are from the engineering institute where the author is working. One of the reasons to select these students and faculties is that they are easily accessible. Also, most of them are from a age group between 18 and 35 which is the youth of this country.

For this feedback submission, 310 participants were invited. But, only 120 students and faculty turned up for the feedback submission (100 students and 20 faculties).

So, initially, the students were asked to gather in the laboratory in several batches, each batch size was 20. The students were informed about the PPET and its functionality. Further, depending on users' choice, the website is opened by those users. The respondents have looked at the privacy policy of that website. Once the reading was finished, then the PPET tool was opened by the user. The users have tried the functionalities of the PPET and spent some time in reading the contents obtained by the PPET. And finally, the users have submitted the feedback. One important thing is that the student had a choice to give feedback for more than one website.

The same procedure was repeated before taking the feedback from the faculty. In total 262 responses were received from these 120 participants**.**

6.3. FEEDBACK ANALYSIS/RESULTS

In this section, the analysis of the feedback is presented in detail. The result of the responses includes the grading of the websites based upon the contents of privacy policies. Also, the responses were received to check the usefulness of the PET such as motivation in reading privacy policies, or help in enhancing privacy awareness:

6.3.1. USER'S FEEDBACK: GRADING OF THE WEBSITES

Fig 6-2 represents the analysis of the responses on the grading of the websites based on the contents of the privacy policies. Looking at the classification, summarization and visualization features of PPET, the Internet users have browsed the contents of the privacy policies. Based on the users' understanding, the grading was submitted during the feedback process.

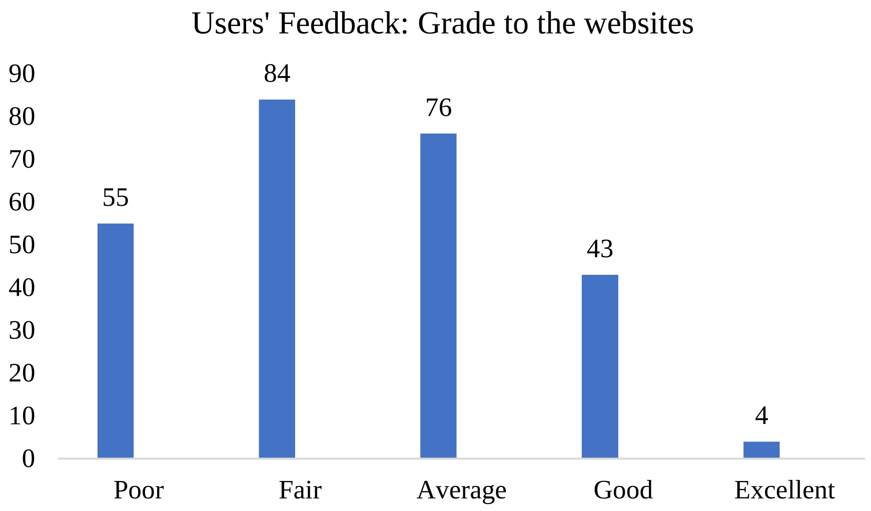


Figure 6-2 Users' Feedback: Grading of the websites based on privacy policies

Fig 6-2 illustrates the various categories of the grading to the websites (poor, fair, average, good, or excellent). On one side, out of 262 responses on the various websites, 55 responses denote that the privacy policies are poor. On the other side, there are only 47 responses that show the privacy policies are good or excellent. The percentage of such privacy policies identified by the respondents are only 17.93 which is very low.

Looking at the responses, of the privacy policies of the websites like "HDFC"[67] which is a fastest growing bank in India or Indian government's rail transport service like "irctc"[68] which is frequently used by the Indian users for train reservations has received a poor or fair grade from users.

A website from sports domain like "ESPNcricinfo"[69], or "Google" (http://www.google.com) have been ranked as best websites by the Internet users.

However, 160 responses (61%) represent the websites that are either fair or average. This major response clearly indicates that there is scope for the improvement in the privacy policies. The shopping website like "Flipkart"[70] or a matrimonial website like

[67] http://www.hdfcbank.com

[68] www.irctc.co.in

[69] disneyprivacycenter.com

[70] http://www.flipkart.com/

“Shadi” (http://www.shaadi.com/) have been rated as average websites based on the users’ responses.

Further, the responses were analyzed and grouped as per the domains. Fig 6-3 represents the grading of the websites of various sectors.

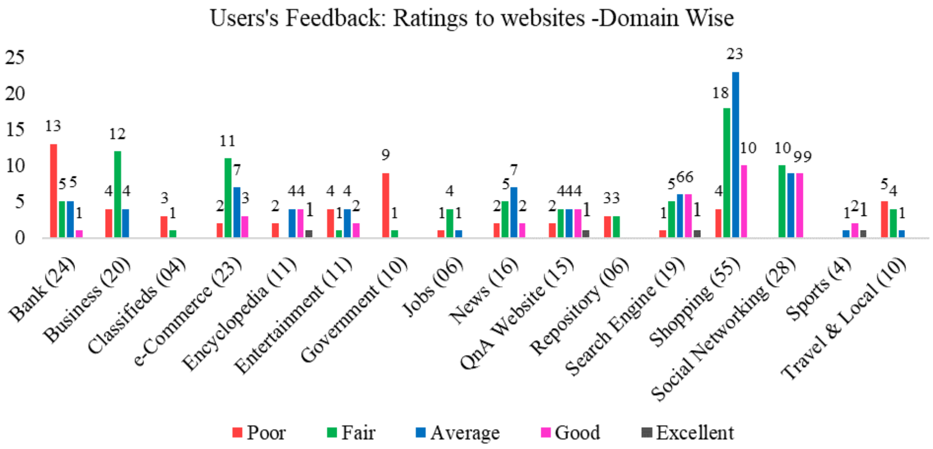


Figure 6-3 Users’ Feedback: Rating to the websites- Domain wise

The responses were received on websites that belonged to various domains from banking to government to travel and local. In the fig 6-3, the x-axis represents the domain names and the number of responses received for the websites under those domains. For example, “shopping” domain has received 55 responses on the websites like Flipkart, Snapdeal, etc. The y-axis represents the number of responses.

The first observation is that the banking websites “HDFC[71], Axis[72], and ICICI[73]” and government websites[74],[75] have received poor grading based on the content of privacy policies. As per the responses, the users believe that the privacy policies are not enough to understand the collection of information, its purpose, sharing, security, etc.

[71] www.hdfcbank.com

[72] www.axisbank.com

[73] *https://www.icicibank.com*

[74] www.irctc.co.in

[75] www.passportindia.gov.in

Considering the use of these websites by Indian Internet users, it is a privacy issue that needs to be addressed.

The second observation is that the shopping websites "Snapdeal[76], Amazon, Flipkart[77], Rediff[78], eBay" have been used frequently. Most of the responses (55 out of 262) have been received for these websites. As per the responses, the grading of the privacy policies of mentioned websites is either fair or average. So, it is important to note that the privacy policies should be good or excellent considering the high number of users accessing them. Hence, the privacy policies should be transparent, easy to understand, represented in a standard way and should follow the privacy principles.

The third observation is that the websites where finance or more personal information is involved should have good or excellent grades. For example, e-commerce websites like "PayTM, PayPal, or BillDeskP" (paytm.com, www.paypal.com, and www.billdesk.com) have been categorized under a fair grading.

Last, but not least, there are only 4 responses where users think that the company's privacy policies are excellent. As per the responses, the contents are readable, easy to understand, etc. These websites are "Wikipedia, StackOverflow, Google, ESPN" (www.wikipedia.org, stackoverflow.com, www.google.co.in, and www.espncricinfo.com).

In short, the important observation is that there is the huge scope and need to upgrade the privacy policies in accordance with the principles of privacy policies or to adhere the regulations of the country.

6.3.2. USERS' FEEDBACK: IMPORTANT SECTION/CONTENTS OF THE PRIVACY POLICY

The next question in the feedback was regarding the important concern to the Internet users in the context of the privacy policy. Considering the manual analysis of privacy policy (chapter 4) and PPET (chapter 5), we identified a few important sections mentioned as follows:

- Information Collection
- Way of Information Collection
- Purpose of Information Collection
- Cookies

[76] www.snapdeal.com

[77] *www.flipkart.com*

[78] http://www.rediff.com/

- Children’s Privacy Policy
- Information Sharing and Third Parties
- Information Security

It is possible that the users have a specific section(s) that he/she might be interested to read thoroughly from the privacy policy. Hence, the answers to this question will give us the idea about the important aspect of privacy awareness that concerns users. During the process, the Internet user can submit one or more than one section as important for them (multiple choice questions). Hence, a total of 886 replies were received up to this point.

All these responses have been analyzed and graphically represented in figure 6-4. Out of the 7 sections, the most important section is Third Party sharing (202), followed by Way of collection (179). The information security and collection stood at the 3rd and 4th position in the list of important sections for the users.

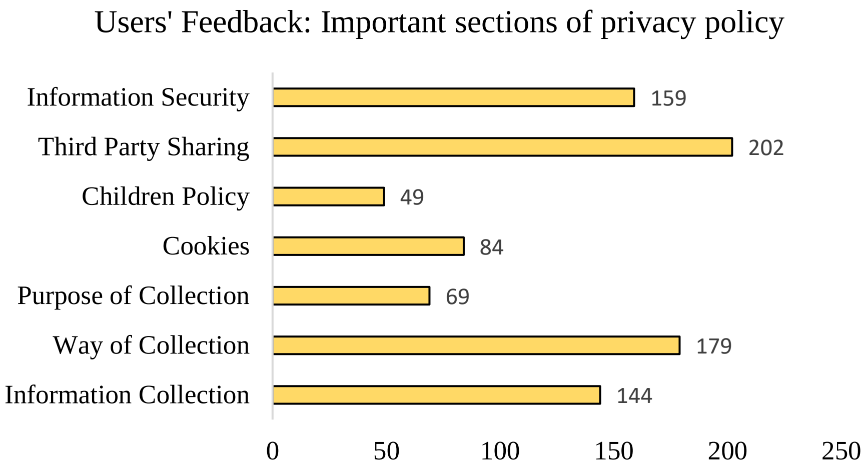


Figure 6-4 User’s Feedback: Important sections of privacy policy

Considering the users’ limited knowledge about privacy as observed in the survey carried in 2014 (chapter 3), the users are now quite interested to know about information sharing. They might be looking for what information is being shared, who are the collecting parties, what are the purposes and their benefits, etc.

In short, the Internet users believe that purpose of collection, cookies, and children’s policy is the least important sections as compared to the remaining sections. However, it is observed that more than one major section can be read/looked into from the privacy policy and accepted by the users.

This also implies that the privacy policy is an important document from the users' view as the users are interested in various sections and the strategy of the service providers in those sections. It is surprising to see that the some of the websites do not mention few sections precisely in the privacy policy.

For example, the Indian government website that issues a passport "Passport India" (http://passportindia.gov.in/) or travel planning website like "Yatra" (http://www.yatra.com/) doesn't specify any kind of security measures in their privacy policy.

Hence, the privacy policy should cover the necessary and some mandatory sections in their privacy policy. It should be more precisely stated and should be governed by the country's laws.

6.3.3. USERS' FEEDBACK: PRIVACY POLICY CONTENT MATCHING USING PPET

In this section, we have asked the participants to what extent the contents of the privacy policy match with the contents produced by the PPET. The analysis revealed that the contents of the privacy policy and the content given by the PPET are matches well for most of the websites. Fig 6-5 illustrates the matching level as per the responses received from the users.

Almost 97% of the responses (254 out of 262) indicated that the privacy policy contents are very well fetched, classified, and visualized using PPET. This shows that the PPET has the ability to work as per the sections requested by the users. The section could be information collection, sharing, a way of collection, etc.

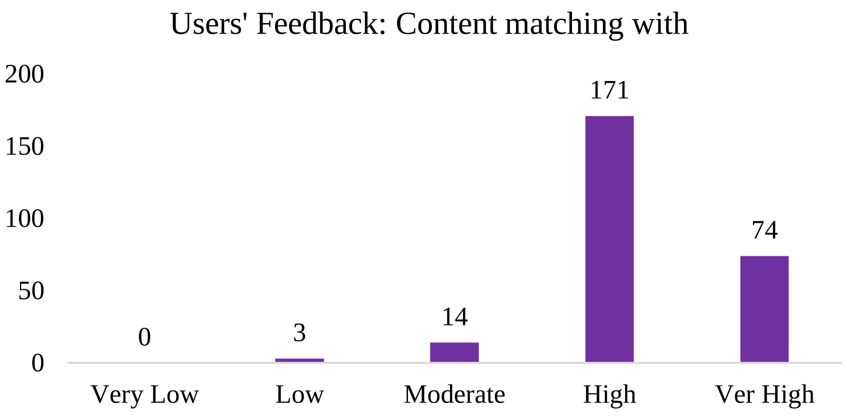


Figure 6-5 Users' Feedback: Content matching

Wikipedia[79], Twitter[80], Flipkart[81], Rediff[82] are a few examples of the websites for which the contents of the privacy policy and the contents from the PPET are well matched. But also, there are a few responses where the user felt that the level of content matching is either low or moderate. The websites like Yatra[83], Olx[84], Yahoo India[85]

6.3.4. USERS' FEEDBACK: MOTIVATION TO READ PRIVACY POLICY USING PPET

The challenges of the privacy policy are considered while developing the PPET. The main functionality of the PPET is to classify and visualize the contents of privacy policy so that the user will be engaged in reading the privacy policy. To take the feedback on this functionality, it was important to ask the users about their views on the motivation for reading the sections/ privacy policy using PPET.

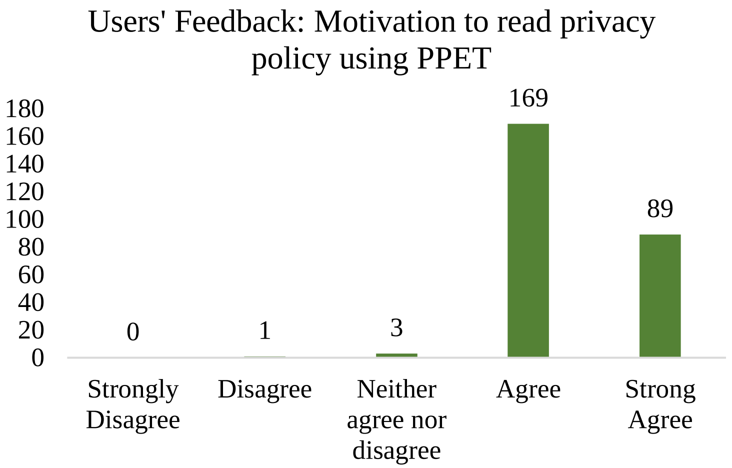


Figure 6-6 Users' Feedback: Motivation to read privacy policy using PPET

[79] *https://www.wikipedia.org*

[80] *https://twitter.com/*

[81] *http://www.flipkart.com*

[82] *http://www.rediff.com/*

[83] *http://www.yatra.com/*

[84] *www.olx.in*

[85] *https://www.yahoo.com/*

Fig 6-6 elaborates the assessment of users on whether the PPET motivates them to read the privacy policy or not. The analysis of 262 responses represents that almost all the users' responses (258 out of 262) agree or strongly agree with the fact that the PPET motivates them to read the privacy policy or a section of the privacy policy.

This shows that the PPET is providing a platform/way to browse the content of privacy policy where the users can spend the time to engage themselves in reading the contents of privacy policies. Hence, this is implying that the tool has the ability to produce the desired output as the classification and visualizations are concerned.

6.3.5. USERS' FEEDBACK: USERS' RATING TO THE PPET

During the feedback, the users were asked to give an overall rating to the PPET. The analysis shows that the tool is liked and appreciated by most of the users. Fig 6-7 illustrates the overall rating to the PPET.

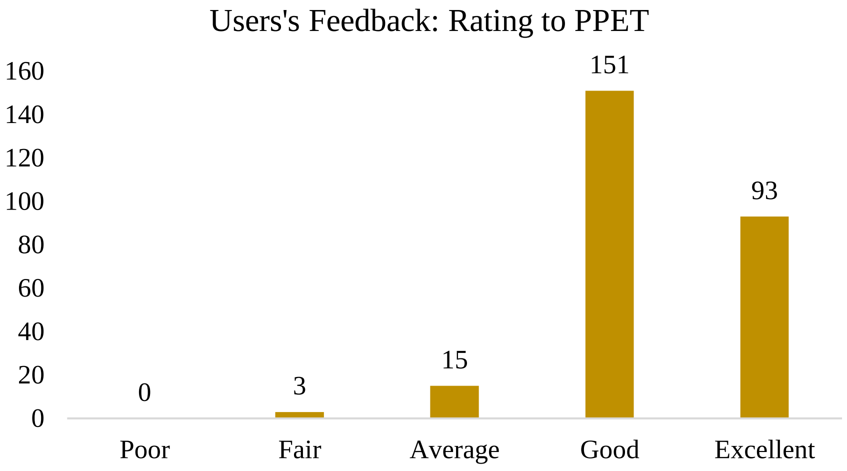


Figure 6-7 Users' Feedback: Rating of the PPET

From fig 6-7, it is clearly observed that the responses 244 out of 262 (93%) gave the ratings to the PPET as either good or excellent. This shows that the tool has performed well as the expectations of the participants of this feedback.

There are hardly few responses (18 out of 262) shows that the PPET is fair or average in the classification and visualization of the privacy policy contents.

6.3.6. USERS' FEEDBACK: PPET HELPS USERS BY ENHANCING PRIVACY AWARENESS

The PPET is one tool that helps the users to know more about the strategy of service providers about personal data management in detail mentioned in the privacy policy. So, another aspect of the PPET is to know at what level the PPET help the users to enhance their privacy awareness. Fig 6-8 illustrates the users' thinking on the enhancement in privacy awareness using the PPET.

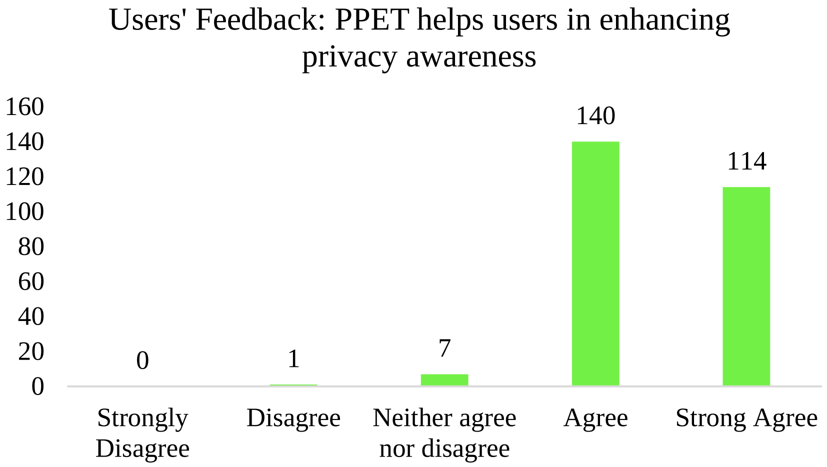


Figure 6-8 Users' Feedback: PPET help users in enhancing privacy awareness

The analysis mentioned on fig 6-8 represents that the maximum responses agree or strongly agree upon the fact that the PPET tool helps them to enhance their privacy awareness. This is to be noted that, the PPET do not enhance privacy awareness, but is acting as a step that leads to privacy awareness.

Hardly, few responses have shown a neutral reaction on privacy awareness using PPET.

6.3.7. COMMENTS/SUGGESTIONS FROM USERS

This is the last question from the feedback form which is asked to users. The text was provided to the users to give comments or suggestions. The comments received from several users (students and faculty). The comments are classified in two categories as 1. Users overall feedback on the PPET and 2. Suggestions to improve the PPET.

The overall feedback includes the words/sentences like "good", "great tool", "nice one", "wow…", "not seen before", and similar.

However, the few feedbacks include concrete suggestions like “Tool should give details on the risk level”, “shd show the personal information based on websites”, etc.

These suggestions and comments helps to define the scope for the future work.

6.4. PRIVACY PRINCIPLES VS REALITY

Looking at the feedback on PPET, most of the responses show the grading of the services/service providers as average or below average considering the contents of their privacy policies. Hence, there is a necessity to extend this study to know the reasons for such gradings. One of the reasons could be the – difference in the users’ expectations from service providers on privacy and the reality observed in the privacy policy documents.

Hence, considering the feedback and manual analysis of privacy policies mentioned in chapter 4, the vital parameters like the information collection, its use, sharing with other entities, security measures, etc. are a few concerns and need to be expressed in detail. Also, there is need to see the service providers who follow the principles of privacy given by the OECD or similar organizations/forums. The following section discusses the few privacy principles from OECD and some of the service providers who do or don’t comply with those principles.

From 1970, the OECD has been continuously encouraging and contributing in promoting to value and respect for the users’ privacy. Also, the guidelines help to limit the free flow of users’ information across the boundaries.

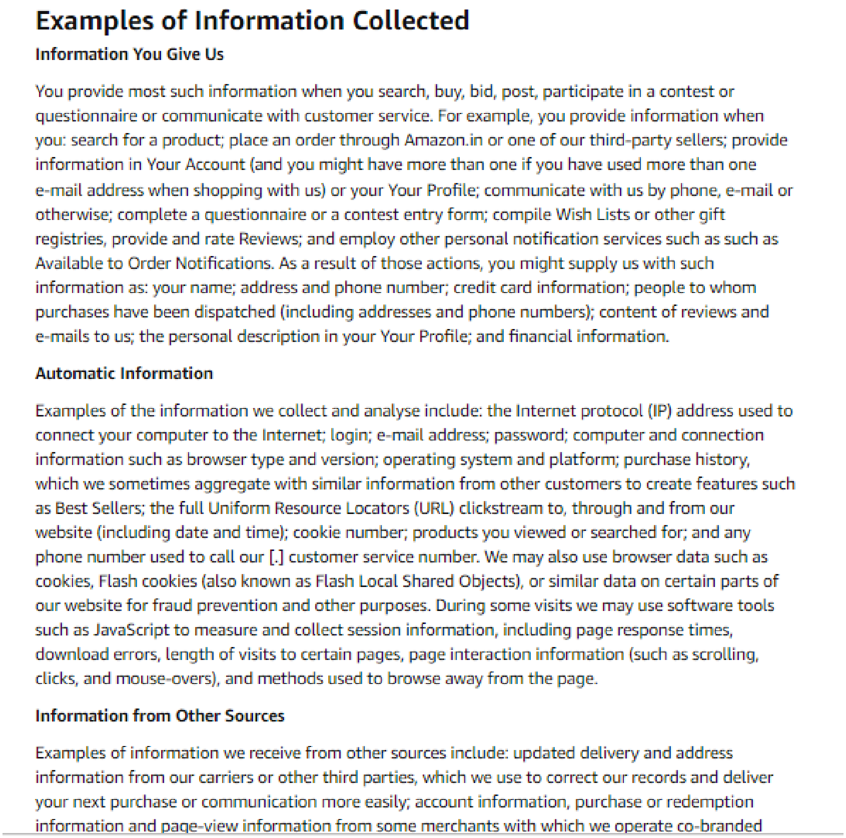
The set of revised privacy principles given by OECD (OECD, 2013) focus on privacy issues, its protection, its implementations, and the privacy at the global level. Specifically, the revision to the 1980s privacy guidelines includes several concepts like privacy management and enforcement, awareness and education programs, privacy strategies at national and at the global level, notification of security breaches.

The guidelines are applicable for private and public data. Some of the principles are discussed here.

- Collection Limitation

As per the collection limitation principles, the data collector/processor should specify a limit on the collection of information from users. Considering this principle, the privacy policies studied in this research do not have limitations on data use, and collection. Hence, the privacy policies of service providers should have an explicit specification on a limited collection of data and its limited use.

For example, it is very difficult to understand the complete list of information collected by the Amazon company. Fig 6-9 represents the information collection section of the privacy policy of Amazon. This section describes information collected from users. The collected information includes users' search details, account information/ profile, communication details, questionnaires/contact forms, wish list, reviews. The information also includes personal details like name, address, phone number, contents of emails, personal details in the profile, along with the credit card information.



Examples of Information Collected

Information You Give Us

You provide most such information when you search, buy, bid, post, participate in a contest or questionnaire or communicate with customer service. For example, you provide information when you: search for a product; place an order through Amazon.in or one of our third-party sellers; provide information in Your Account (and you might have more than one if you have used more than one e-mail address when shopping with us) or your Your Profile; communicate with us by phone, e-mail or otherwise; complete a questionnaire or a contest entry form; compile Wish Lists or other gift registries, provide and rate Reviews; and employ other personal notification services such as such as Available to Order Notifications. As a result of those actions, you might supply us with such information as: your name; address and phone number; credit card information; people to whom purchases have been dispatched (including addresses and phone numbers); content of reviews and e-mails to us; the personal description in your Your Profile; and financial information.

Automatic Information

Examples of the information we collect and analyse include: the Internet protocol (IP) address used to connect your computer to the Internet; login; e-mail address; password; computer and connection information such as browser type and version; operating system and platform; purchase history, which we sometimes aggregate with similar information from other customers to create features such as Best Sellers; the full Uniform Resource Locators (URL) clickstream to, through and from our website (including date and time); cookie number; products you viewed or searched for; and any phone number used to call our [.] customer service number. We may also use browser data such as cookies, Flash cookies (also known as Flash Local Shared Objects), or similar data on certain parts of our website for fraud prevention and other purposes. During some visits we may use software tools such as JavaScript to measure and collect session information, including page response times, download errors, length of visits to certain pages, page interaction information (such as scrolling, clicks, and mouse-overs), and methods used to browse away from the page.

Information from Other Sources

Examples of information we receive from other sources include: updated delivery and address information from our carriers or other third parties, which we use to correct our records and deliver your next purchase or communication more easily; account information, purchase or redemption information and page-view information from some merchants with which we operate co-branded

Figure 6-9 Information collection section of the Amazon.in[86] the privacy policy

The information collection continues using another method called automatic collection. The collected information includes IP, email, password, browser details, OS, platform, purchase history, URL, page response times, page visit length time,

[86] https://www.amazon.in/gp/help/customer/display.html/ref=footer_privacy?ie=UTF8&nodeId=200534380

clicks, scrolling, mouse overs, etc. The other sources of information include third parties, merchants, credit bureaus, etc.

In short, Amazon usually collects every information from users, which is not limited. Hence, the limit of information collection is not defined precisely. On the other hand, the BookMyShow website collects users' information without specifying the complete details in the privacy documents.

Hence, we conclude that the privacy policies of websites do not have specific limits on the data collection and its use. Therefore, there is a need to make a mandate on limited information collected and its wise use.

Considering this principle, none of the privacy policies have specified the limit of information collection in specific ways.

- Purpose Specification Principle

This principle suggests that the privacy policy should clearly specify the purpose of data collection at the time of its collection. Also, it should be precisely stated if the purpose is changed. Considering this principle and the privacy policies examined in this research, there are several service providers whose privacy policies are unclear for the purpose of collecting each personal attribute.

The information collection sentence from *BookMyShow.com*[87] is:

> *"There are times when we may collect personal information from you such as name, physical address or telephone number. It is our intent to inform you before we do that and to tell you what we intend to do with the information. Generally, you will have the option not to provide the information, and in the future, you will be able to "opt out" of certain uses of the information. If you choose not to provide the information we request, you can still visit the Bookmyshow website, but you may be unable to access certain options, offers and services."*

Looking at the above statements, the purpose of collection of name, physical address, and telephone number is unclear. It will lead to several questions like what are the intentions, how are they going to inform, how will they deal with users' information, etc. Hence, it is doubtful whether such privacy policies follow the purpose specification principle.

Also, it should be clearly noted that the "opt out" is not possible in the present state as the privacy policy of "BookMyShow" states that this will be possible in future.

[87] https://in.bookmyshow.com/privacy

Similar cases have been observed in the privacy policies of ICICI bank[88] and indianairforce[89], a government website.

- Security safeguards principle

According to this principle, the information of users should be protected against the various risks. The type of the risks could be use, disclosure, alteration in users' information, or unauthorized access, etc. To protect the information from above-stated risks, the service provider should take reasonable security safeguards.

As mentioned in our research study (Dhotre, Olesen, & Samant, 2016), only 57% of privacy policies have mentioned general description about security measures which are very abstract and may not give the security assurance.

The banking websites which are very vital in the area of information collection and its protection are concerned. However, the analysis of privacy policies has revealed the fact the banking websites did not mention or mention little about security and privacy of users' information.

For example, the privacy policy of "HDFC" bank has no statements on the protection of users' information which is really surprising.

However, an example of banking website is "Axis" bank where the general statement on security in the privacy policy[90] is:

> *"The security of personal information is a priority and is protected by maintaining physical, electronic, and procedural safeguards that meet applicable laws. Employees are trained in the proper handling of personal information."*

Another example is the privacy policy of "Yatra"[91], where the statements on security measures are:

> *"Yatra takes appropriate steps to protect the information you share with us. We have implemented technology and security features and strict policy guidelines to safeguard the privacy of your personally identifiable information from unauthorized access and improper use or disclosure"*

[88] http://www.icicibank.com/privacy.page

[89] http://www.icicibank.com/privacy.page

[90] https://www.axisbank.com/

[91] http://www.yatra.com/online/privacy-policy.html

The two examples above denote that security measures are not specific like the privacy policies of various service providers have mentioned measures as firewalls and data encryption, Secure Socket Layer (SSL), technical, administrative and physical security measures, etc.

To conclude here, the privacy policies should adhere to the principles of privacy to avoid the free flow of information among service providers and other entities. Such principles, if followed thoroughly, could act as an obstacle for misuse of personal information in the important sectors like banking and insurance. Also, the principles are necessary to study to help to have a standardized way of representing the contents of the privacy policy which is completely missing in the privacy policies analyzed in this research. The later section discusses a proposal of standard template for privacy policy.

6.5. LIMITATIONS OF PPET

As mentioned earlier, the PPET is a visualization tool that analyses the privacy policy and using visual aids, it shows the contents of privacy policies. This PPET is an awareness tool that helps the users in reading privacy policies of service providers in the easiest way. We believe that the PPET also helps the users to enhance privacy awareness.

However, this tool has some limitations:

1. PPET Based on privacy policy: This research is analyzing, summarizing and visualizing the contents of privacy policies. The research is completely based on the privacy policies and doesn't use any other information except the corpus developed during our research.

2. No Semantic Value: During the research, we focused on the actual contents of the privacy policy. We have not considered the semantic value of the privacy policy statements. In the future, we would like take up work on the meaning and other aspects of privacy policies.

3. The PPET doesn't define Privacy: The research work mentioned in this thesis is not defining privacy in the Indian context, even though we focused on the Internet users and possible service providers.

4. Work at Micro level- Looking at the feedback of PPET, this research motivates the users to read the privacy policy and help in enhancing privacy awareness to a certain extent. Also, this tool is designed and executed as a browser extension. So, the user must install it which is very easy. Once the user visits the website, the contents will be displayed once the user clicks on

the extension icon. The challenge is to spread this tool on a large scale (Macro level) where most of the users can use it.

Despite the above limitations of this research/tool, we identified that the privacy policies are dissimilar in many ways. Hence, there is a need for a uniform way of representing privacy policy. Therefore, the recommendations are given in the following section.

6.6. RECOMMENDATIONS

This section contains two parts. The first recommendation is the standard template for privacy policies. The second part is the recommendation of the new law along with responsibilities of entities for enhancing privacy awareness.

6.6.1. STANDARD TEMPLATE FOR PRIVACY POLICY

For a user, there are several difficulties when it comes to privacy policy. Hence, for making informed decisions or check legislations easily, the privacy policy should be well structured. The privacy policy should be simple, meaningful and understandable to the Internet users. To stimulate users' interest and strengthen the brand of service providers, it is important to have a transparent privacy policy.

Hence, we propose and recommend a standard, unified template for privacy policies in the Indian context. The template should have the following elements:

1. Accessibility and Format:

   Considering the importance of Privacy policy, we recommend that a link to the privacy policy should be easily accessible from the first/home page of the websites. Even the link should be at once defined on the home page. We recommend the link should be at the bottom of the page.

   Currently, the format of the privacy policies is not only unstructured but is in several formats (for example, bullet points, tabular format, or plain text, etc.). Hence, the privacy policies should have a unique standard format.

2. Readability

   To create an interest in the policy document, the language should be easy, and shouldn't use complicated words/terms that the users cannot understand. The service providers should follow the readability score provided by the Flesch–Kincaid (ReadabilityScore.com, 2016). The use of layered format (Sections) will help the users to read it fully or a section of a privacy policy.

We also recommend using icons or a short video to represent the contents of privacy policy just to know its meaning or how it works? This will surely engage the user in reading or listening to the contents of the privacy policies.

3. Objective and Scope

Once the user starts reading, the objective should be presented in the beginning. The objective should focus on the protection of users' information, personal data management and the respect of the users' privacy. The online and offline practices on service providers should be well described in the privacy policy.

4. Information Collection

Looking at the feedback on the PPET, most of the users are interested to know about information collection. Hence, this element/sections should be specified in detail in the privacy policy by the policymakers of the service providers.

It would be very effective if the policy specifies the list of users' PII and non-PII collection. The policy should mention the information collection methods (many ways like registration, cookies or other technologies).

Our analysis of privacy policies shows that the information is collected and kept by the service provider forever. A finite retention time (number of days, weeks, etc.) should be specified by the service providers in the privacy policies.

5. Purpose of Information Collection/Use of Information

Another important section about which the users would like to know more about is the purpose.The purpose should be clearly mentioned and it should as per each attribute being collected. For example, if the first name is collected, the policy should have clear statements on its use.

Also, if the user realizes the information is being used unlawfully, then there should be an option of withholding the consent. An effective dashboard view should be provided that will show the complete list of attributes collected followed by its use.

A control mechanism or rights to the users should be given so that it will help the users to modify or delete some information if needed.

6. Information sharing and disclosure

Looking at the privacy policies in this study, most of the policies share users' data to other entities (like govt, affiliates, third party). However, none of them have clearly specified the list of those entities. Hence, in this recommended standard template, the information (or link) to the entities should be given the purpose of information being shared with them. This will help to understand the business strategy of the companies.

7. Customer consent and choice

   The present method of service providers to avail the services is to accept the policy completely, even if the user is not happy with few terms or sections of the privacy policy. Hence, there is major revision needed on user consent and choices. The privacy policy should be clear, simple, and must have explicit statements on the type of information collected, sharing strategy, etc.

8. Security Measures

   The analyzed policies in this study have variable security mechanism adopted and some of them have defined general statements on users' information security. Hence, the recommendation is that the privacy policy should be mention clear statements on security that use updated measures/techniques. Similarly, users' information should be governed by the country's laws where the information is stored (data storage location may be mentioned in the privacy policy or should be communicated privately if the users are interested).

9. Notification and User Feedback

   As part of the privacy ecosystem, the user is also important and responsible for privacy awareness. Hence, the feedback from users is vital and must be considered by the service providers. Using a contact officer or a forum, where the users can share their opinions for improvement of the privacy policy should be encouraged.

   The privacy policy should be updated and it should be communicated to the users effectively. This could be achieved using email.

   Apart from the above elements as a part of the standard template, the privacy policy should be validated by the government authorities (if available) and then the policy will be named as or sealed as *"privacy-aware policy"*.

Apart from the user side, the will help the service providers in garnering the users' trust. Hence, users/customers will be retained and the business-relationship can go a long way in customer retention.

6.6.2. PRIVACY AWARENESS PARTIES AND RESPONSIBILITIES

This section is about the discussion of the entities responsible for enhancing or spreading awareness about privacy in the Indian context. As a part of the research, we identified entities as shown in Fig 6-10.

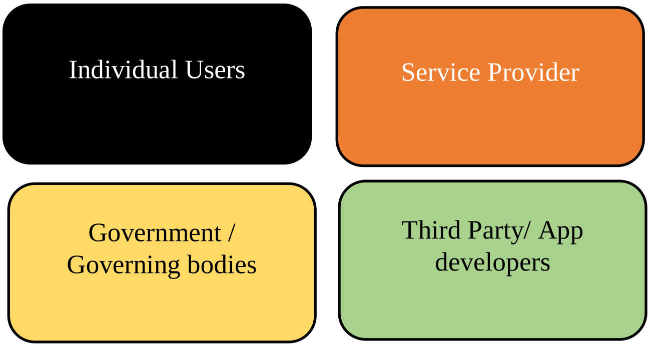


Figure 6-10 Privacy awareness parties

1. Individual Users

   The users should be more careful and should be able to take decisions on information disclosure. They should be aware of privacy and related issues. Once they are online, they should be able to take a decision on which type of information should be shared with whom. Users should believe in minimum information disclosure and should share what is required and necessary.

   The users should read the privacy policy before accepting these conditions. For simplicity, the users can use awareness tool like PPET that help them in reading the privacy policy and maybe enhance their privacy awareness.

2. Service providers

   A simple, transparent and fair privacy policy is the main responsibility of the service providers. The panel (or dashboard like PPET) will help the users to know about the information practices of service providers.

   The user consent, revoke mechanism should be changed effectively, which would provide more user control. The user choice and feedback should be considered for the better enhancement in the privacy policy

   Also, there could be a forum/platform from where the service provider can exchange information as agreed upon by both the users and the service providers to avail the service. This win-win situation helps both entities doing business without any risks

3. Government or Governing Bodies

The standard and legitimate efforts should come from government at least in India. India doesn’t have dedicated laws on privacy protection or cybercrime, but it has set of rules on misuse of personal information in India’s mother legislation ‘Information Technology Act 2000’ (The Gazette of India, 2009). Specifically, in its chapter 11, under the section 65-78, it is punishable with imprisonment up to 3 years or a penalty of Rs. 5 Lakh for a person who gains users' information in a wrongful way or reveals the information to other entities without user consent.

Hence there is a need for a separate law on privacy. To start with this, the government of India can look at the GDPR as a base.

The GDPR is introduced to specify how users’ information should be used and protected from significant risks. This GDPR has been adopted in April 2016 by the European Parliament and will become enforceable in May 2018 in EU.

This includes several parameters/elements as:

- Consent: The consent should be easy and companies should not use indecipherable terms and conditions filled with jargon. Hence, in the privacy documents should make it easy to give or withdraw the consent

- Breach Notification: Once the data is breached, it’s the job of the data processor to notify the users about the possible risks within 72 hours.

- Right to Access: There should be a specific right where the users can have access to know whether their information is being processed. A free electronic copy of users' data should be given the users.

- Right to be forgotten: The users should ask the controllers to erase personal information if that is not needed for the intended purpose.

- Data Portability: There should be a provision where the user can obtain their data across various IT environments.

- Privacy by Design: From the design step, the data protection should be included. It should implement appropriate technical and infrastructural measures.

- Data Protection Officers: qualified, professional officers must be appointed who is responsible to check systematic monitoring and processing of personal information.

There are several influences in the context India if new laws come in the near future. This will include a restriction on commercial use of personal information, inspire trust and confidence, right to users' information privacy, etc.

4. Third parties and App Developers

   The third parties sometimes unknown to the users are responsible to get the needed information by asking the users in a clear way. The collection of information, their methods, purpose and other important details must be specified in a clear manner.

6.7. REFERENCES

Dhotre, P., Olesen, H., & Samant, K. (2016). Interpretation and Analysis of Privacy Policies of Websites in India. In *WWRF Meeting 36*.

OECD. (2013). OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data - OECD. Retrieved September 23, 2017, from http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm

ReadabilityScore.com. (2016). Measure the Readability of Text! - Improve your writing and your website marketing with Readability-Score.com.

The Gazette of India. (2009). *Information Technology Act 2008,*. Retrieved from http://meity.gov.in/sites/upload_files/dit/files/downloads/itact2000/it_amendment_act2008.pdf

CHAPTER 7. CONCLUSIONS AND FUTURE WORK

This chapter is a summary of all the contributions discussed in thesis followed by the scope for future work.

7.1 CONTRIBUTIONS OF THIS THESIS

The wireless and mobile technologies have facilitated efficient and convenient access to services that make users' life easier. In the advanced digital world, the users share personal information to avail of the services offered by the service providers. The users are not aware of the business practices of the service providers which are obtainable in the form of privacy policies. These policies should be easy to read and understand. But, the thesis has revealed that the privacy policies are unstructured, unclear with respect to management of the user's information. Furthermore, the contents of the privacy policy have been systematically analyzed and visualized. And finally, a solution has been proposed that will help not only the users to read and understand the privacy policy easily but also the service providers by providing a standard template for privacy policy. The highlights of this thesis are as follows:

- A survey of privacy awareness and online practices of Internet users to know the issues and challenges in enhancing privacy awareness.

- Manual analysis and interpretation of the privacy policies of service providers in India show that the privacy policies are unstructured, and unclear about information collection, use, consent and choice, notification, etc.

- Careful design and implementation of semi-automated PPET tool for classification, summarization and visualization of privacy policies.

- Analysis of feedback on PPET and a recommendation on a standard template for privacy policies of service providers in the Indian context.

The main objective is to design an analytical and visualization tool with a focus on engaging the users to read and understand the privacy policies that will empower the users to enhance their privacy. The contributions presented in this thesis focus on the issues and challenges faced by the user in privacy policies. The main challenges with respect to privacy awareness and privacy policies are presented in this thesis. For the known challenges, the existing solutions/methods/tools are evaluated and new way/tool has been proposed that help the user to read the privacy policies and understand it.

Chapter 1 elaborates on information sharing and service utilization - the "actual scene"- where users and service providers meet. The interaction leads to a situation where the users are characterized or described in the form of users' information, behavior, interest, social relations, etc. The information collected by the service providers is used for two purposes. The first purpose is constructive and beneficial to the users in that it helps service providers to authenticate and provide customized services to the users. This is how the information collected should be justified by the government laws and guidelines perspective.

However, the second purpose of information collection is to build user profiles which are shared with other entities making it quite difficult for users to manage their personal information. This "actual scene" is the motivation to identify the challenges in enhancing users' privacy awareness. This chapter presents the research question (how to enhance users' privacy awareness) along with research objectives. This chapter proposes a solution that contains a survey, extensive analysis of privacy policies and a semi-automated tool. The condensed view of the thesis is presented in thesis organization.

Chapter 2 presents the research context where the definition of privacy and its interpretation is discussed. Interpretation of privacy varies from user to user depending upon various demographic characteristics of users like region, country, culture, educational background, job-profile, and much more. However, according to the survey and research work, various threats to users' privacy include an extensive collection of users' information and its trading, tracing users using logs, cookies, and creating user profiling and its trading etc.

Privacy awareness is one of the ways to address these issues where user empowerment is quite necessary. Privacy by Design (PbD) and the regulations like GDPR focuses on protecting users' privacy will form a base for privacy laws in countries like India. We studied various privacy supporting tools like Disconnect-Me[92], MyPermission[93], Web of Trust (WOT)[94], Ghostery[95], Terms of Service; Didn't read (ToS;DR)[96], etc. and concluded that the most of these tools are blocking third parties to access users' activities and data.

However, this research has identified and exposed a need to have a solution that helps to enhance users' privacy awareness before accepting the terms and conditions. This will help users to make informed decisions.

[92] *https://disconnect.me/*

[93] *https://mypermissions.com/*

[94] www.Mywot.com

[95] *https://www.ghostery.com/*

[96] *https://tosdr.org/*

To understand online users' perception of privacy and practices of online users in India, a survey and its detailed analysis are mentioned in chapter 3. Inspired by the privacy principles from PbD, and OECD, the questions were based on information access, sharing, security, along with current practices, users' knowledge about privacy and laws, etc.

The analysis of this survey shows that most of the Internet users have limited knowledge about privacy, are worried about information collected by service providers, have difficulties in reading privacy policies, etc. The monitoring factors for improving personal information privacy are effective consent mechanism, the appointment of personal trust manager, etc. Considering the difficulties in the protection of users' privacy, this thesis focuses on privacy awareness. Understanding privacy awareness among Indian users was achieved with the help of quantitative analysis. However, the interviews could have helped to get understanding and feelings about privacy awareness in detail. Also, the survey would have been completed statistically if all the participants (age, location, profiles) had participated in this survey.

In Chapter 4, the aim is to retrieve and interpret information on the business practices of service providers through an extensive manual analysis of their privacy policies. In this chapter, our work has analyzed over 50 privacy policies of the most popular web services in India from different domains. The domains are in banking, traveling, entertainment, government, etc. The work presented in the thesis shows that there is a mismatch between the users' expectation and assurance given by the service providers. Several issues of privacy policies include the complexity of language, unclear aims, the absence of specifications for the purpose of information collection, trading of information without users' explicit consent, etc.

According to Flesch-Kincaid readability evaluation tool, only 3 websites' privacy policies are easy to read and understand. Most of the websites are collecting some user data with explicit knowledge of users through a users' login, registration to use the service provided by websites. But most of the time the information used is for the purpose other than that it is specified for. The study ascertains that there is little or no clear explanation given for the use of cookies in the privacy policies. Security mechanisms or measures are important to protect users' information. Only a few websites have clear specified security mechanism/measure/method opted for the security of users' data. The service provider's strategy for sharing, trading personal information to other entities (third parties, affiliates, etc.) is not clearly understood the privacy policies.

Apart from the above findings, the outcome of this chapter is an emphasis on privacy policy analysis, which can process unstructured, unreadable complex language of the privacy policy to shorten/summarize into easily understandable text or visual cues. To spread awareness about the content of privacy policies, the research continued to use machine learning methods and develop a tool instead of relying purely on human intervention.

Chapter 4 discuss about the manual analysis of the privacy policies performed by the author. However, the analysis is based on the sentences of the privacy policies. This work could have been implemented using automated by developing machine learning based algorithms.

Chapter 5 elaborates the design of a semi-automated tool that will try to identify the gap between the service providers' assurance and the users' expectations. Our research has proposed a tool "PPET" to classify, summarize and visualize the privacy policy contents using visual cues. Again, following the principles of privacy, we have aimed to classify the contents of privacy policies into various sections for easy reading. These sections are Information Collection, Way to collect information, Purpose of Information collection, Information sharing, Cooking, Children Policy, Information Security, Other.

Privacy policy document is classified by the PPET into one or many categories mentioned above by using Naïve Bayes Classifier. To develop the machine learning based model, the training data was not available. In the development of the automated tool, the training data to Naïve Bayes Classifier is manually constructed, called as Corpus. The corpus was constructed by labeling each statement of 52 privacy policies manually, which has produced 43,544 statements in the corpus.

The PPET works in several steps where the first step is pre-processing of privacy policies. (Tokenization, Removal of Pronouns, stop words, transforming a statement to sequence of important words by identifying important words, stemming, Lemmatization, etc.). Finally, a classifier labels the paragraphs of the privacy policy as a category using the token matched as per the corpus. Later by using the notion of the frequency of words, the rank of statements is calculated which helps to summarize the contents of privacy policies. This will help the users to enhance readability.

The PPET provides an extension to use services provided by Web of Trust (WoT) which are to understand various aspects of the website of the service provider like trustworthiness, children policy, etc. The visual output of the PPET motivates the user to read and understand the privacy policy thereby enhance the privacy awareness of the user.

The PPET is a static tool that has focused on over 600 privacy policies. This tool could have been implemented dynamically that will perform classification, summarization, and visualization of any privacy policy.

Chapter 6 is a discussion about the PPET, its impact and a recommendation for a standard template for privacy policies. The feedback of the PPET collected from over 250 Internet users, revealed grading of the websites based on the privacy policies. The interesting fact about website grading shows that the government and banking websites are poor in terms of personal information management.

The feedback also illustrated that the PPET has been able to fetch and visualize the contents of privacy policies to a significant extent. The PPET has helped the users to read the privacy policies easily. In this thesis, we have analyzed how the PPET has enhanced the privacy awareness of the Internet users in India. Hence, we can conclude the PPET has achieved the goal of our research.

7.2 OPEN PROBLEMS

The primary aim of the PPET was to interpret privacy policies and convey it to the users with the help of the visualization tool. This tool could be extended by enhancing the quality of corpus that will increase the accuracy of the results. Also, the PPET works offline- the analysis and visualization are not dynamic in nature. This is because of resource constraint; the runtime analysis and visualization could be a further improvement in the PPET to enhance privacy awareness.

However, the topic of privacy awareness is broad and covers many issues. Considering information flow between the users and the service providers, other issues like privacy risk assessment, information bargain, etc. can be combined with the proposed tool discussed in this thesis.

A (semi) automatic mechanism could be proposed to measure users' privacy risk level by considering privacy-related settings, third-party observers, privacy policy, and internet cookie data being exchanged. This mechanism could be responsible to quantify users' privacy risk level for all online services irrespective of their services. Inputs for privacy risk level measuring mechanism are the users' environment parameters (e.g. Privacy related setting, the presence of cookies), User interaction with web services, code of web services, the output of privacy threat analysis, and the privacy concerns of users.

Further, the ambitious idea is to improve Internet users' privacy by tackling the service providers' "Take it or Leave it" attitude, which holds the users to ransom, forcing them to accept the privacy policy. The new challenge is to design an approach that will balance the interests of users and service providers. This could be achieved by introducing a framework where the user will grant certain permissions to the service based on the personal information they ask and purpose of information use.

There is a need to extend the work presented in this thesis, by introducing a bargaining mechanism where the users will see the benefits from the service providers once the user shares some personal information. This will help the users to accept the user-provider liaison.

Another possibility could be to assign a value to users' personal information where users will be ready to pay for privacy. The service providers will charge the users based on the level of privacy they wish to have.

APPENDICES

Appendix A. Survey Questionnaire

Email contents to the Internet users

The link to online survey questionnaire was sent by email to the Internet users with some description about the context. The email contents are:

Dear All,

I am Prashant Dhotre, Asst Prof, Comp Engg Dept. I am pursuing PhD from Aalborg University, Denmark.

My research work is "Security in Big Data". We get free services from Gmail, google, facebook, WhatsApp and other service providers [flipkart, MakeMyTrip, bookmyshow..etc].

The cost of using all these services is that service providers extensively collect and make use of our personal information. Using 'Big Data techniques hidden patterns can be revealed and additional value can be extracted from the data, giving rise to serious problems with leakage of our personal data.

In the same context, I am doing a survey of Indian users. Please find your free time and give your valuable feedback.

Your feedback is valuable to me!
I hope I can contribute in the favor of all Indian users security.

Link is:

https://docs.google.com/forms/d/1onJQ2n31nDfIFmKqlBhiEM-m4wytfGJFtBN9eYxUewg/viewform?c=0&w=1&usp=mail_form_link

Prashant S. Dhotre,
Asst Prof, Computer Engg. Dept.

A Questionnaire on Survey on "Personal Information Privacy Awareness"

1:Introduction:

1.1. You're Name Please:

1.2. Location (City):

1.3. Gender Male Female

1.4. Your age in years

Under 16

between 16 and 25

between 26 and 35

between 36 and 45

between 46 and 55

Between 56 and 65

Above 65

1.5. Please select your present state of the profession:

Full-time employed

Part-time employed

Self-employed

Non-paid employment

Working student

Retired

Other:

1.6. In related to the computer field, what you think about yourself:

A beginner

an intermediate user

a legitimately experienced user

a very experienced user

2:Understanding User And His /Her Current Practices:

2.1. From where do you access services/utilities provided by service providers, please choose:

Home

School/college/university

cybercafé

Office/work

Other:

2.4. Do you think your life will be easier by using the Internet for banking/ shopping/ reservation services /others?

Yes

No

2.2. Which of the internet facilities you use from computer/ laptop/ mobile?

(Please specify like Social web site, banking, reservation, online shopping, etc.)

2.3. Name the services/utilities you use most of the times? (Keep blank if you don't want to share)

like gmail, google, makemytrip, flipkart, bookmyshow, facebook)

2.5. Mention the list of electronic documents/ forms, which require that

your personal information is provided to service providers.

(Information like offline and online applications, online eligibility/enrolment records, electronic health records, online government certificates, authorization during online transactions, service payment claims, tax, credit cards, Law enforcement, etc.

3:General Perceptions And Awareness Of Privacy:

3.1.What would you think about the primary usage of collecting personal information by service providers/researchers/organization ?

3.2.Are you aware of any federal institutions that help Indians to deal with privacy and the protection of personal information from wrong collection, use and exposing publicly?

Absolutely Yes

Absolutely No

3.3. What would you think about your knowledge level for your privacy rights under the personal information protection laws?

1 2 3 4 5

Very High

Don't Know

3.4. How do you rate yourself if you think about Internet Security?

(E.g.searching some material, using social web sites, while reading emails, doing online transaction, etc.) The security can be protection of data, trust, privacy etc.

Not at all worried

Somewhat worried

Very worried

I know I should be worried, but I'm not worried

For Following Statements Please specify your Agreement or Disagreement

3.5. In today's era of computers, I feel that I have less protection of my personal information as compared to 10 years ago

Agree

Disagree

Don't Know

3.6. I feel confident that I have enough information to know how new technologies might affect my personal privacy.

Agree

Disagree

Don't Know

3.7. Protecting the personal information of the Indians will be

3.8. Have you ever actively hunted out for information about your privacy

the most important issues facing our country in the next 10 years.

Agree

Disagree

Don't Know

3.9. It is observed that many organizations ask the Indian people for personal information for some reasons. In general, have you frequently shared your personal information with organizations that they ask for it?

Always

Regularly

Sometimes

Rarely Never

3.11. What is the most serious risk to your privacy, if at all any? Please Select

Bank/finance fraud

Credit card fraud

Computer privacy/internet security

Identity theft/Identity fraud

Personal information is being accessed

Personal information is being shared

Cell phone privacy

Monitoring/activity tracking

Medical/Cronic Health Information

Other:

3.13.In your view, how seriously does the businesses/service provider perform their duty to rights, for example by visiting a website, searching on the internet, contacting an agency / organization, or reviewing a standard publication for help?

Yes

No

3.10. Have you ever asked an organization that requested personal information from you why they want it, what they will do with it or where they store and process it?

Yes

No

3.12. In your view, how seriously does the central/state government perform their duty to protect personal information about the Indian citizen?

Rarely

Some what

Not really

3.14.Do you think that the organizations/service providers would inform you, if your collected personal

protect the personal information of Indian consumer?

Rarely

Some what

Not really

information is lost during business, unintentionally stolen or made public/open?

Absolutely

Possibly

Possibly Not

Absolutely Not

don’t know

4:Are You Serious About Your Online Information? [Please tell me how nervous you are about each of the following when you think about the information available about you online.**]**

4.1: Using this information, some companies/organizations send you junk mail/spam,

Very nervous

somewhat nervous

not nervous

4.2: Without a warrant, the law administration/ national security agencies, collects it for general surveillance/investigation purposes.

Very nervous

somewhat nervous

not nervous

4.3. Marketing companies use this information to analyze your likes and dislikes.

Very nervous

somewhat nervous

not nervous

4.4. Organizations/companies use this information to determine your suitability for a job or promotions.

Very nervous

somewhat nervous

not nervous

4.5.Governments use this information for other purposes, such as to confirm claims for social benefits or payment of taxes.

Very nervous

somewhat nervous

not nervous

5:Online Activities Information

5.1. What do you most prefer PRIVACY or UTILITIES/SERVICES

Privacy

Utilities/Services

5.2. Do you use a mobile device, such as a cell phone, smart phone, tablet, or like this?

Yes

No

5.3. Have you came across a situation where your password (email/internet banking/ credit

5.4. Has anyone posted online about you so that it has negatively affected your life in any way? (Ex: something you

card) is stolen (either online or offline)

Yes

No

5.5. How worried are you about posting your comments, photos, videos, online?

Very worried

Somewhat worried

Not worried

posted some words/pictures/album/video of yourself or someone else posted about you on social web)

Yes

No

5.6. If you enter a URL/Request, then the response will come to your machine. As per your knowledge which of the following information is possible to record from your request?

Your machine address

Your residential address

Email ID

date and time of your request

A requested page

Type of browser

present operating system and version

monitor /screen size

Your location from where the request was sent

Other:

6:Users Behaviour During Advertisement And Tracking

6.1. Are you aware that your Internet activities, like your browsing, what you search and purchases, service interactions, and etc. can be monitored by…

(The websites you visit, both commercial and government websites, Other companies, such as marketers and researchers)

Yes No

6.3. Do you believe and think that service provider/Internet companies should ask your consent/permission to track what you do on the Internet?

Absolutely Perhaps

6.2. Do you personally take any steps/actions to limit tracking of your Internet/online activities?

(The websites you visit, both commercial and government websites, Other companies, such as marketers and researchers)

Yes

No

6.4. How important and critical do you feel it is that websites should actively inform you about what kinds of personal information they are collecting and how they use it

1 2 3 4 5

Perhaps Not Not Absolutely Not

6.5. Complete the following statement. I would like to share my critical information to a website/service provider.(Please check all that apply.)

If and only if, website/service provider tells me what information they will collect

If and only if, website/service provider tells me how the information will be used.

I will not share my critical information under any other terms and conditions.

6.7. Presently the bulk mailing (a.k.a. Mass/ spamming) often contain eye catching advertisements, political influencing comments, get-fast-rich schemes, etc. In your opinion, which of the following policies is most similar to your approaches about what should be done?

I am fine with present practice.

The Government must pass a law to make it illegal.

To allow message filtering, a blacklist of spammers should be constructed

A registry must be created which contains a list of members not wishing to receive mass mailings.

Not important and critical at all	Very important and critical

6.6. Most of the time after receiving a bulk mailing (bulk mailing means same e-mail is sent to a number of people) what you do?

Read the message

Delete it once I understood it was a mass mailing.

Send a message back asking not to be included in future mailings.

I will not answer

I have not experienced any bulk mailing yet.

I will do some other action.

No option

6.8. Why do you deny from filling out online registration forms at different sites?

Takes too much time

Requires me to give my name

Requires me to give an email address

Requires me to give my residential address

Information is not provided on "how the data are going to be stored and Used"

Accessing the site is not worth to disclose requested information

I do not trust the entity collecting the data

I always register

Other:

I would like to have some other policy.

I Don't Know

Please say your views on each of the following statements in terms of agreement or disagreement

6.9.There should be new laws to protect user privacy on the Internet.

Agreement

Disagreement

6.10. I support the establishment of a personal trust manager (where a trusted entity/party keeps my preferences/experience to build trust between me and the service provider)

Agreement

Disagreement

6.11. Web sites need information about their users to market for their own advertisement

Agreement

Disagreement

6.12. Content/service providers have the right to share/resell information about their users to other agencies/companies.

Agreement

Disagreement

6.13. A user should have complete control over which sites that collect the user's important and critical information.

Agreement

Disagreement

6.14. Online e-books to which I subscribe have the right to sell my information (name, cell number and address) to companies they feel will be interested in my information

Agreement

Disagreement

6.15. I would be very happy if I receive mass/ bulk mailings that were specifically targeted to my interests/preferences

Agreement

Disagreement

6.16. I would be interested to take on different roles/profiles at different websites while on the Internet.

Agreement

Disagreement

6.17. I will be interested in being anonymous when visiting sites on the Internet

Agreement

Disagreement

6.18. I should be able to communicate over the Internet without other people being able to understand and read the content

Agreement

Disagreement

6.19. I would prefer to be become anonymous when I do Internet payment

Agreement

Disagreement

6.21.The terms and conditions of service providers should be shortened and simple to understand while doing online registration

Agreement

Disagreement

6.20. I will allow third party advertising companies/ agencies to track my "online experience" during different web sites for their own advertisements

Agreement

Disagreement

7:Privacy Policies And Laws

7.1.Do you read the terms and conditions (privacy policies) while doing your registration for websites/service providers?

Yes

No

7.3. Do you think that the privacy policy is very lengthy, unable to understand and descriptive?

Yes

No

7.2. How regularly do you read the privacy policies of websites/service providers you visit?

Continuously

Often

Sometimes

Hardly

Never

7.4. How you find service provider's privacy policies about what the service provider will do with your personal information?

Understood completely

understand somewhat

some what not understandable

not at all understandable

7.5. Have you come across any situation where you decided not to select a website or a service because you are unhappy with the terms that set in privacy policy?

Yes

No

7.7. Is there a policy concerning processing and travelling with

7.6. If personal information is collected over the Internet, is there a specific policy to manage this practice?

Yes

No

7.8. Where do you ask the question if security/ privacy is breached?

Information Protection Act

personal and important information?

Yes

No

Security Specialist Service Provider

Police Dept

Government Authority

System Administrator

7.9. Do you know any law for data/information protection in Indian Constituency if any?

Yes

No

7.10. Which kind of privacy do you feel is more important in your life?

Personal

Professional

Social

7.11. How do you rate the present privacy system as per your experience while dealing with service providers?

(Very good, good, neither good nor bad, bad, very bad)

1 2 3 4 5

Very Good Very Bad

8:User Organizations/Service Providers

8.1. Do you feel that the service provider should notify you when they deal (access, sale, share) your personal information?

Yes No

8.2. Do you feel there should be a mechanism to rank the service provider/services according to your experience collection. For example, Do you think the rating of ""how secure"" a specific site is helpful for you?

Yes No

8.3. Does your agency discloses individual ADHAR Card Number (or Passport Number) or discloses user records that contain ADHAR Card Number (or Passport Number)?

Yes No

8.4. Does your agency print an individual's entire ADHAR Card Number (or Passport Number) on any card required for the individual during the course of business?

Yes No

8.5. Does your company/organization require individuals to transmit the individual's entire ADHAR Card Number (or Passport Number) over the Internet during the course of business?

Yes No

8.6. Does your organization/agency lease, trade, sell, rent, or otherwise deliberately release user's ADHAR Card number (or Passport Number) to a third party?

Yes No

9: User On Mobile and Medical and Health Information

9.1. Keeping cell phone in use, tell us what you do with it

Use a password to lock it

Download apps even I know it access my personal information

So as to limit access personal information, I adjust some privacy settings

Store personal information on it.

9.3. Have you ever turned off the location disclosure feature on your mobile device because you were worried about others may access your location information?

Yes

No

9.5. How concerned are you about providing your medical information when trying to obtain insurance or applying and joining for a job profile?

	1 2 3 4 5	
Not worried at all		Extremely worried

9.2. Have you ever decided not to install, or uninstall, an app because of the amount of personal information you need to provide?

Yes

No

9.4. Do you think there are more chances of losing/compromising of personal information stored on a mobile device as compared to the information stored on your home computer?

(Significantly more issue, moderately more issue, About the same, Moderately less issue, significantly less issue)

	1 2 3 4 5	
Significantly More Issue		Significantly Less Issue

9.6. Would you fear to undergo genetic (medical) testing?

Definitely

Perhaps

Perhaps not

Never

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-077-7

AALBORG UNIVERSITY PRESS