MODELING AND ANALYZING USERS' PRIVACY DISCLOSURE BEHAVIOR TO GENERATE PERSONALIZED PRIVACY POLICIES

by
A K M Nuhil Mehdy

A dissertation
submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Computing
Boise State University

August 2021

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

A K M Nuhil Mehdy

Thesis Title: Modeling and Analyzing Users' Privacy Disclosure Behavior to Generate Personalized Privacy Policies

Date of Final Oral Examination: 15 June 2021

The following individuals read and discussed the dissertation submitted by student A K M Nuhil Mehdy, and they evaluated the student's presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Hoda Mehrpouyan Ph.D.	Chair, Supervisory Committee
Jyh-haw Yeh Ph.D.	Member, Supervisory Committee
Casey Kennington Ph.D.	Member, Supervisory Committee
Michael D. Ekstrand Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Hoda Mehrpouyan Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

DEDICATION

to my father, who always inspires me to live a meaningful life with honesty and morality

ACKNOWLEDGMENT

First and foremost, I wish to express my deepest gratitude and sincere appreciation to my advisor, Dr. Hoda Mehrpouyan, who continuously supported and encouraged me during this research endeavor. She also inspired me to do the right thing with professionalism, even when the road becomes harder. Most importantly, she guided me to work on various yet coherent research topics and collaborate with brilliant researchers during my Ph.D. work. I am thankful to her for this exceptional opportunity. All of my accomplishments would not have been possible without her encouragement, inspiration, and direction throughout my entire Ph.D. journey.

I would also like to show my sincere appreciation to Dr. Casey Kennington and Dr. Michael D. Ekstrand for being on my defense committee. Alongside my advisor, they have also been the greatest mentors I have ever had. I have had the incredible opportunity to work with them closely and learned a lot of valuable skills throughout the collaborations. This knowledge will undoubtedly benefit me in every part of my career. I am also very grateful to my other committee member Dr. Jyh-haw Yeh. His valuable advice and effort to provide me with constructive and insightful comments have significantly improved the quality of my research work. I can not thank them enough for their help, support, and advice along the way.

Besides, I am also thankful to all the co-authors of my publications: Dr. Bart Knijnenburg from Clemson University, Stephen Reese from the Idaho National Labo-

ratory, Subin Sapkota, and again Dr. Casey Kennington and Dr. Michael D. Ekstrand from Boise State University. I would like to send special thanks to Dr. Knijnenburg for his valuable contributions to my research work covered in chapter 3.

Most importantly, I would like to appreciate the funding support for all the research works covered in this dissertation. Thanks to the National Science Foundation for its support through the Computer and Information Science and Engineering (CISE) program and Research Initiation Initiative(CRII) grant number 1657774 of the Secure and Trustworthy Cyberspace (SaTC) program: A System for Privacy Management in Ubiquitous Environments.

Additionally, I need to thank all of my teachers, faculty members, staff, and lab mates at Boise State University for making this period of my life memorable. I would especially like to mention Dr. Amit Jain, Dr. Gaby Dagher, Dr. Jerry Alan Fails, Dr. Tim Andersen, Dr. Jodi Mead, and Dr. David Gabbard for their tremendous support. I want to extend my gratitude to my cubicle mates Shariful Alam, Dhanush Ratakonda, and Rezvan Joshaghani, and others for their friendship and the delightful discussions.

Last but not least, I would like to thank my parents and siblings for supporting me spiritually throughout my whole life. I am also grateful to my lovely wife, Mahfuza Khatun, for supporting and inspiring me to achieve the highest degree while being very busy with her own Ph.D. research and our son Marwaan Mehdy. May almighty Allah bless them.

CITATIONS

Material from this dissertation has been published or submitted for review in the following form:

1. Nuhil Mehdy, M. Ekstrand, B. Knijnenburg, H. Mehrpouyan, "Privacy as a Planned Behavior: Effects of Situational Factors on Privacy Perceptions and Plans," 2021 *29th Conference on User Modeling, Adaptation and Personalization (UMAP 21)*, ACM

2. Nuhil Mehdy, H. Mehrpouyan, "Modeling of Personalized Privacy Disclosure Behavior: A Formal Method Approach," 2021 *4th International Workshop on Behavioral Authentication for System Security (BASS 21) @ARES 21*

3. Nuhil Mehdy, H. Mehrpouyan, "A User-Centric and Sentiment Aware Privacy-Disclosure Detection Framework based on Multi-input Neural Network," 2020 *PrivateNLP @13th ACM International WSDM Conference*

4. Nuhil Mehdy, C. Kennington, H. Mehrpouyan, "Privacy Disclosures Detection in Natural-Language Text Through Linguistically-motivated Artificial Neural Network," 2019 *2nd EAI International Conference on Security and Privacy in New Computing Environments (SPNCE 19)*

5. Nuhil Mehdy, H. Mehrpouyan, "A Multi-input Multi-output Transformer-based Hybrid Neural Network for Multi-class Privacy Disclosure Detection," 2021 *2nd*

International Conference on Machine Learning Techniques and NLP (MLNLP 2021) **(Pending Notification)**

ABSTRACT

Privacy and its importance to society have been studied for centuries. While our understanding and continued theory building to hypothesize how users make privacy disclosure decisions has increased over time, the struggle to find a one-size solution that satisfies the requirements of each individual remains unsolved. Depending on culture, gender, age, and other situational factors, the concept of privacy and users' expectations of how their privacy should be protected varies from person to person. The goal of this dissertation is to design and develop tools and algorithms to support personal privacy management for end-users. The foundation of this research is based on ensuring the appropriate flow of information based on a user's unique set of personalized rules, policies, and principles. This goal is achieved by building a context-aware and user-centric privacy framework that applies insights from the users' privacy decision-making process, natural language processing (NLP), and formal specification and verification techniques. We conducted a survey (N=401) based on the theory of planned behavior (TPB) to measure the way users' perceptions of privacy factors and intent to disclose information are affected by three situational factors embodied by hypothetical scenarios: information type, recipients' role, and trust source. To help build usable privacy tools, we developed multiple NLP models based on novel architectures and ground truth datasets, that can precisely recognize privacy disclosures through text by utilizing state-of-the-art semantic and syntactic

analysis, the hidden pattern of sentence structure, tone of the author, and metadata from the content. We also designed a methodology to formally model, validate, and verify personalized privacy disclosure behavior based on the analysis of the users' situational decision-making process. A robust model checking tool (UPPAAL) is used to represent users' self-reported privacy disclosure behavior by an extended form of finite state automata (FSA). Further, reachability analysis is performed for the verification of privacy properties through computation tree logic (CTL) formulas. Most importantly, we study the correctness, explainability, usability, and acceptance of the proposed methodologies. This dissertation, through extensive amounts of experimental results, contributes several insights to the area of user-tailored privacy modeling and personalized privacy systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

TPB	Theory of Planned Behavior
A	Attitude
SN	Subjective Norm
PBC	Perceived Behavioral Control
UI	User Interface
IT	Information Type
RR	Recipient's Role
TS	Trust Source
TS	Trust Source
CI	Contextual Integrity
HIT	Human Intelligence Task
CFI	Comparative Fit Index
TLI	Tucker-Lewis Index
TLI	Tucker-Lewis Index
SD	Standard Deviation
CNN	Convolutional Neural Network
DRER	Disclosure-Related Entity Recognizer
NER	Named Entity Recognizer
IC	Information Content

TFIDF Term Frequency Inverse Document Frequency

SVM Support Vector Machines

AMT Amazon Mechanical Turk

PHI Personal Health Information

PII Personally Identifiable Information

DNN Deep Neural Network

NLP Natural Language Processing

DP Dependency Parser

ReLU Rectified Linear Unit

ReLU Rectified Linear Unit

LSTM Long Short-term Memory

TP True Positive

FP False Positive

TN True Negative

FN False Negative

DLP Data Loss Prevention

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

RNN Recurrent Neural Network

NSP Next Sentence Prediction

MLM Masked Language Modeling

ROC Receiver Operating Characteristic

AUC Area Under Curve

AUC Area Under Curve

FSA Finite State Automata

CTL Computation Tree Logic

FSM Finite State Machine

TCTL Timed Computation Tree Logic

DFA Deterministic Finite Automata

API Application Programming Interface

PIA Privacy Interface Automata

CHAPTER 1:

INTRODUCTION

The concern regarding users' data privacy has reached an all-time high due to the massive increase in communication platforms, social networking sites, and greater users' participation in online public discourse [115, 80]. Also, an increasing number of people exchange their private information arbitrarily via various mediums without knowing the risks and implications. In many cases, improper disclosure of sensitive information could be the root cause of privacy violation, and the negative consequences of the disclosure could be immense [32]. However, the responsibility is mostly on the user themselves to take control of what kind of information should be shared with whom, when, and how [166]. Unfortunately, for an individual, it is quite cumbersome and difficult to manage and control their information sharing activities because of the complexities associated with the platforms [170]. Also, users' decision to share their private information, and the perceptions of risk that inform this decision, depend on the individual and vary from situation to situation [68, 141, 81]. Therefore, ensuring the appropriate flow of information based on the user's unique set of personalized rules, policies, and principles is the core topic of this dissertation.

In this chapter, we first introduce what is data privacy as per the scope of this dissertation. Then we define privacy disclosure and talk about potential privacy threats that are associated with it. Later, we discuss the situational privacy behavior of

human beings and why the analysis and modeling of personalized behavior are essential for ensuring policy-governed appropriate flow of information. Then we discuss the roles of the privacy verification engine and automated text analysis tool to help archive this goal. Finally, we briefly represent a potential privacy management system that integrates these components while depicting the overall significance of the research.

1.1 Data Privacy

Privacy is an ancient concept concerning human values that could be "intruded upon", "invaded", "violated", "breached", "lost", and "diminished" [151]. Each of these metaphors reflects a conception of privacy that can be found in one or more standard models or theories of privacy. The seclusion and non-intrusion theory of privacy has defined the user's privacy as "the right to be left alone" or "being free from intrusion" [163, 55]. In the information domain, privacy refers to the right of a person to monitor and control the processing, exposition, and preservation of personal information [98].

Even though privacy varies from individual to individual and each user may have different views of privacy, there is an imperfect societal consensus that certain information (e.g., personally identifiable information (PII), financial situation, health condition, relationship issues) is more private than the others (e.g., public statements, opinions, comments, reviews) [26]. White et al., for instance made an interesting distinction that certain types of information, when disclosed, can cause loss of privacy, but other types of information, when disclosed, might just cause embarrassment [167]. Researchers have also intended to classify someone's private information into two main categories: objective (i.e., factual information such as age, sex, marital status, etc.) and subjective (i.e., internal states of an individual such as interests, opinions, feel-

ings, etc.) [156]. Unintended disclosure of such private information to inappropriate recipients could lead to serious privacy threats. We talk about this issue in detail in the following section.

1.2 Privacy Disclosure and Threats

As per the scope of this dissertation, we define *privacy disclosure* as an occurrence when a piece of text which is usually a statement/expression from an author, discloses someone's private information/situation. In other words, we focus mostly on the objective disclosure where users explicitly reveal someone's privacy in terms of their financial situation, health condition, relationship issues, etc. For example, a disclosure happens when a user, through an email, text message, or a tweet, shares his/her physical/mental health condition, diagnosis results, medication/drug they are taking, or any diseases they got recently or in the past, etc. Another example of disclosure could be when a user shares his/her economic situation, such as the financial crisis he/she is going through, the profit he/she made recently, his/her investment details, etc. We consider three types of such information as the domain of this research: health, financial, and relationship.

That being said, recent advances in communication technologies such as messaging applications, email services, online forums, and electronic social media have resulted in privacy concerns about analogous information amongst users [151, 127]. This is because an increasing number of people exchange their private information arbitrarily with different types of recipients, without being aware of the risks and implications [149]. Alongside, with the increasingly sophisticated and available technology, the effectiveness of collecting, analyzing, exploiting vast amounts of user information by service providers has also significantly increased [130]. For example, Rosenberg et

al. claim that "the privacy policy" on various websites are generally ignored by users unless they make the effort to truly understand what they are granting permission to, and to whom they are giving their personal information. As a result, their control over data privacy also continues to deteriorate [130].

In many cases, disclosure of someone's sensitive information to inappropriate recipients could cause serious privacy violations, and the negative consequences could be immense [32]. A recent data scandal involving Facebook and Cambridge Analytica shows how personally identifiable information of up to 87 million Facebook users influenced voter's opinions [143, 61]. Likewise, millions of data breach incidents are happening all over the world, and unfortunately, most of them get exposed in public [146]. Therefore, user-centric targeted attacks by exploiting the victim's personal information become a new genre of privacy threat in the present day [162]. It's worth mentioning that the United States is the number one destination for such user-centric targeted attacks based on some recent statistics [150]. Therefore, it is quite evident that individual's private information will always be reachable and attainable to cyber-criminals regardless of the security level and privacy policy at the recipients' end. Hence, Rosenberg et al., uttered "the best and most effective way to control the use of information, without interfering with the conduct of others, is to prevent it from ever coming into others' hands." [130]. That being the case, users' data privacy has become one of the major concerns of today's world, and the requirements for user-centric privacy measures to help each individual protect their private information have been researched extensively [24, 103, 139, 69]. However, to account for privacy perceptions and preferences in user models and develop personalized privacy systems, we need to understand how users make privacy decisions in various

situations[18, 133]. We briefly talk about this topic in the following section.

1.3 Situational Privacy Behavior

Users' decisions to reveal their private information, as well as their risk perceptions, differ depending on the situation [68, 141, 81]. Situations consist of various factors such as the information type, recipient of the information, and the trust source behind the motivation for sharing. Also, depending on culture, gender, age, and other situational factors, the concept of privacy and users' expectations of how their privacy should be protected varies from person to person. Therefore, to understand, model, and possibly predict human privacy behavior in various situated environments, it is important to understand their privacy decision-making process.

In order to do that, there have been several factors and parameters documented to influence users in their privacy decisions. The theory of planned behavior (TPB) [7], an extension of the theory of reasoned action [158], is a behavioral theory that helps modeling users' perceptions and plans based on several influential factors. According to this theory, people's behavior is directly determined by their behavioral *intentions*. These intentions are in turn influenced by their *attitude* (positive or negative evaluation of the decision), perception of the *subjective norms* (generally expected behavior in their social group), and *perceived behavioral control* (ease or difficulty to perform the behavior). Also, the *perceived behavioral control* can, together with *intention*, be used 'to explain the actual *behavior*. However, most privacy research based on this theory has either studied single situations or has considered a very limited set of situational factors [60, 133]. As a result, understanding the characteristics and impact of various situational factors on users' privacy decisions is still an active area of research. In this dissertation, we study users' situational privacy decisions to bet-

ter understand how users make privacy decisions in various situations and how the situational factors have significant effects on users' perceptions of privacy and intention to disclose private information. Results from this behavioral analysis contribute several insights to the area of user-tailored privacy modeling and personalized privacy systems. In fact, we represent a privacy verification engine in this dissertation to assist the users for better privacy practices which is based on the findings of our situational privacy behavioral modeling. We briefly describe the verification engine in the following section.

1.4 Policy-governed Flow of Information

Users' data privacy must be assured not by restricting or preventing the sharing activities but by ensuring the appropriate flow of information based on user-specific rules, policies, and principles. A popular privacy management theory, known as the theory of contextual integrity (CI) [16] also formulates users' data privacy as the appropriate flow of information. Based on this theory, the information sharing activities should conform with the privacy policies of the user. This and other similar theories and researchers have reached the consensus on requirements for user-tailored privacy policies, preferably defined as mathematical expressions [166, 70, 140].

However, in practice, it is quite difficult for an individual to define, manage, and control their information sharing preferences because different devices, applications, and software require different privacy settings from users, and most importantly, they are not designed to be personalized and easy to configure [170]. Existing methodologies and protocols intend to tackle this problem by employing techniques such as access control policies [121, 138], machine-readable privacy policy languages [38, 10], and formal methods [11, 23], etc. Still, most of the works attempt to frame the

problem from a request's perspective, which lacks the crucial involvement of the information owner, resulting in limited or no control of policy adjustment. Moreover, very few of them take into consideration the aspect of personalization and explainability, and 'their practical usability and acceptance remain an important challenge' [90].

In this dissertation, we have presented a methodology to formally model, validate, and verify personalized privacy policies based on finite state automata (FSA) and perform reachability analysis to verify privacy properties through computation tree logic (CTL) formulas. The proposed methodology uses a model-checking based formalism technique to check a set of privacy requirements against a new information-sharing attempt, usually described as a CTL query. This step is achieved by exhaustively searching a system's state space to determine if a given query is satisfied or not, in real-time. If the CTL query does not get satisfied, then the verification engine warns users of a potential violation of their privacy rules. However, even before performing the formal verification step, this engine requires essential inputs from another component. In other words, when a user attempts to share private information in the form of text data, a tool first decides whether or not that piece of text contains private information. If it includes someone's private information/situation, this action is then parameterized (e.g., information type, recipient) and passed to the verification engine. Therefore, in this dissertation, we represent custom models to detect privacy disclosure through natural language processing. We discuss this component in the following section.

1.5 Analysis of Communication Data

Humans use natural language to communicate with each other in the form of text or voice, which has a vibrant structure with different levels of ambiguities [28]. Despite that, the first step towards building privacy systems is to develop a component that can accurately detect users' privacy disclosure through automated text analysis. Researchers from the area of natural language processing have been intensively working to develop different techniques for identifying privacy disclosure [134, 27, 2]. A wide range of methodologies such as dictionary utilization, information theory, statistical models, and machine learning have been shown significant results in identifying private information in textual data [31, 59, 27].

However, most of the existing methods solely rely on the existence of keywords in the text and disregard the underlying meaning of the utterance in a specific context. Hence, in some situations, these methods fail to detect disclosure or produced results are missclassified. In the context of user-centric privacy systems, the task is more crucial than just spotting sensitive keywords. Instead, this is about identifying the actual occurrence of someone's privacy disclosure that may happen through a piece of text. In other words, considering the data subject, actors' (sender, receiver) involvement, sentiment, and grammatical structure to classify texts as a privacy disclosure. Moreover, as we mentioned in the earlier section, it is essential to extract parameters (e.g., information type, recipient) from the text data to generate and validate the privacy properties. Therefore, to help build complete usable privacy tools, we present multiple NLP models in this dissertation that can precisely recognize privacy disclosures through text by utilizing semantic and syntactic analysis, hidden pattern of sentence structure, tone of the author, and metadata content.

1.6 Objectives and Contributions

The main goal of this research is to lay the foundation for developing personalized privacy tools for the end-users of various communication platforms. The key contributions are as follows:

- **Objective 1: Analyzing and Modeling of User's Situational Privacy Decision-making Process**
  - How do people make the decision towards disclosing their privacy under different circumstances, and what are some influential factors behind the choice?
  - How does this process could be explained by utilizing and extending the theory of planned behavior (TPB)?
  - How do users' subjective perceptions of TBP constructs differ in different informational situations?
  - How do situational perceptions affect users' intent to disclose information?
- **Objective 2: Detecting Privacy Disclosure Occurrence through Natural Language Processing (NLP)**
  - Exploring the importance of semantic and syntactic representation of natural language text data.
  - Utilizing transfer-learning, linguistics, and metadata to train deep-learning models for precisely detecting privacy disclosure occurrences.
  - Evaluating the performance of the models on human-annotated ground-truth dataset.

- **Objective 3: Modeling of Personalized Privacy Disclosure Behavior through Formal Method Approach**
  - Model-based analysis of personalized privacy disclosure behavior.
  - Formulating personalized privacy policies
  - Detecting and reasoning about unwanted disclosure behavior.
  - Validating the proposed model-based approach and demonstrating its practicality.

1.7 Integration of the Research Objectives

As mentioned already, this dissertation consists of different modular research goals, from understanding human behavior and their decision-making process, developing text analysis models, and designing a formal method based verification engine. Altogether, this dissertation lays the foundation for developing personalized privacy tools while sharing extensive amounts of experimental results and contributing several insights to the area of user-tailored privacy modeling and personalized privacy systems. Figure 1.1 depicts the overall implications of the research while representing the relationship among the research objectives.

A verification engine is the front-end tool of a potential privacy management system that observes users' privacy behavior and later assists them in maintaining good privacy practices based on their rules and policies. This engine warns, budges, and helps the user whenever a privacy-violating sharing action happens. This component refers to research objective 3 in this dissertation. This component, however, relies on the parameters and disclosure detection results from the text analyzer. In other words, the verification engine needs to know whether the user is *about to* disclose

their privacy by analyzing the text they want to share. This goal is achieved through the work referred in research objective 2. Nevertheless, to allow the privacy management system to be more personalized and behavior-oriented, it should be designed and developed based on the privacy behavior of the user. Therefore, analyzing and modeling the privacy behavior of humans is also very important, referring to research objective 1 of this dissertation.

Figure 1.1: Integration of the Research Objectives.

1.8 Dissertation Outline

This dissertation starts by representing a detailed analysis of the effects of situational factors on users' privacy perceptions and plans in chapter 3. This is done by studying users' situational privacy decisions through a scenario-based survey with 401 participants, each responding to several of 48 different hypothetical scenarios. We perform a path analysis to model participants' privacy perceptions and plans, taking into consideration their attitudinal evaluations on *subjective norm*, *perceived behavioral control*, and *attitude* by manipulating three situational factors: information type, recipient role, and trust source. The results from the analysis reveal how users make privacy decisions in various situations, and how the situational factors have significant effects on users' perceptions of privacy factors and intention to disclose potentially private

information.

Next, in chapters 4, 5, and 6, we represent the novel architectures of three different NLP models and their evaluation reports on ground truth datasets that can precisely recognize privacy disclosures through text by utilizing state-of-the-art semantic and syntactic analysis, the hidden pattern of sentence structure, tone of the author, and metadata from the content. Chapter 4 contains the detail of our first model that applies off-the-shelf natural language processing tools to derive linguistic features such as part-of-speech, syntactic dependencies, and entity relations. From these features, a multi-channel convolutional neural network is trained as a classifier to identify short texts that have privacy disclosures. Later in 5, an updated version of the previous model is represented that takes into account the authorship and sentiment (tone) of the content alongside the linguistic features and techniques. Eventually, chapter 6 represents the detail of yet another multi-input, multi-output hybrid neural network that utilizes transfer-learning, linguistics, and metadata to learn the hidden patterns, and better classify disclosure/non-disclosure content in terms of the context of situation. This chapter also includes the evaluation of this sophisticated model on a human-annotated ground truth data-set. Each of the proposed models is considered as the supporting plugin to help build usable privacy tools.

Chapter 7 represents a methodology to formally model, validate, and verify personalized privacy disclosure behavior based on the analysis of the user's situational decision-making process. In this chapter, by utilizing a model checking tool named UPPAAL, users' self-reported privacy disclosure behavior is represented by an extended form of finite state automata (FSA), and reachability analysis is performed for the verification of privacy properties through computation tree logic (CTL) for-

mulas. Chapter 8 contains the notable limitations of this work, following chapter 9 that summarizes the material presented in this dissertation, makes closing remarks, and discusses future research directions.

CHAPTER 2:

BACKGROUND AND RELATED WORK

Analyzing humans' privacy behavior and the decision-making processes, building text analysis models, and designing a formal method based verification engine are all part of this dissertation's objectives. Therefore, the literature review in this section is further segmented into several sub-sections based on the research objectives. Section 2.1 briefly describes the study about analyzing users' privacy-decision making process and details associated related works. Then in section 2.2, we review the related research works and state-of-the-art techniques for detecting privacy disclosure through text analysis that refers to research objective 1. Later in this chapter, section 2.3 contains a brief literature review regarding different formalism techniques to model, validate, and verify personalized privacy policies.

2.1 Analyzing and Modeling Users' Situational Privacy Decision-making Process

In this section, we first briefly introduce the theory of planned behavior which is the basis of our analysis method for analyzing situational privacy behavior. In the following subsections, we review the related research that uses this theory to model users' privacy decision-making process. We also briefly review the research that models users' contextual privacy decisions.

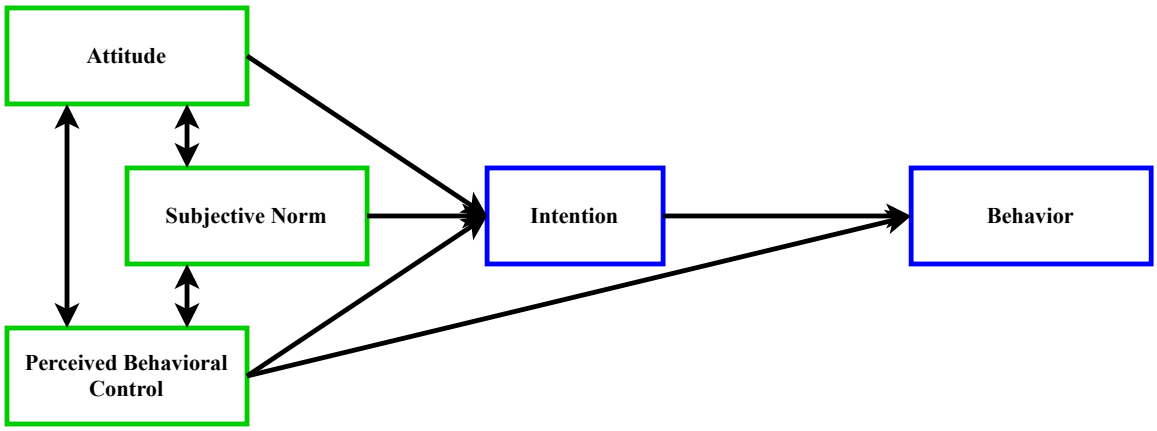


Figure 2.1: Theory of planned behavior and its core components[7].

2.1.1 Theory of Planned Behavior (TPB)

According to the TPB, people's behavior is directly determined by their behavioral *intentions*, which are in turn influenced by their *attitude*, perception of the *subjective norms*, and *perceived behavioral control*. Also, the *perceived behavioural control* can, together with *intention*, be used to explain the actual *behavior*. In the literature [18, 60] these constructs are defined as follows:

- Attitude (A) is defined by the positive or negative evaluation of the decision (e.g., how well the participant understands the value of an action).
- Subjective norm (SN) is defined as a culturally appropriate and desired behavior that is generally expected of a person with in his/her social group (e.g., how a participant's closest relatives act on similar situation).
- Perceived behavioral control (PBC) is defined by the perceived ease or difficulty that the individual addresses to perform the behavior.

The theory states that these constructs or components together shape an individ-

ual's behavioral intentions. Thus, it provides a model to capture humans' behavioral intention (Figure 2.1). Theory of planned behavior is used in many research areas and has demonstrated its effectiveness in predicting human behavior in various fields such as privacy [60, 159], use of the internet [172], health [34], environmental psychology [100], etc.

2.1.2 Modeling Users' Privacy Decisions Using TPB

In spite of the dynamic nature of privacy behavior [68, 141] and the fact that privacy paradox shows that users' intentions and attitudes may not always result in privacy-protective behaviors [4], studies have used TPB to investigate and model the most important factors that influence users' privacy decision-making process [7]. Heirman et al. [60] analyzed the impact of the TPB factors (i.e., attitude, subjective norm, perceived behavioral control) on the disclosure of private information through a structured survey. A similar TPB-based approach was utilized by Saeri et al. to investigate Facebook users' privacy protection tendency based on descriptive norms, risk, and trust [133]. Yao et al. extended the TPB to model users' intention to adopt an online privacy protection strategy [172]. Their analysis showed that "the intention to adopt online privacy self-protection is a function of one's attitude towards protective strategies, the subjective norm of adoption, and the perception of behavioral control". Lwin et al. combined Laufer and Wolfe's multidimensional approach to privacy [92], and an extended version of Ajzen's theory of planned behavior [7] to study the privacy behavior of online users [99]. They partially used a TPB inspired conceptual framework to investigate the reasons behind users' intention to disguise their identities (i.e., private information). While TPB is normally used for grounding designs and analyses related to any type of human behavior towards an action [159],

researchers have successfully used TPB for in-depth analysis of privacy attitudes and privacy behavior [62, 20, 25] with ample justifications [45].

All of the above-mentioned works have one common limitation: they assumed the users' privacy perceptions (TPB construct measures) to be stable and did not take into account the potential impact of contextual factors. The way contextual dimensions influence TPB remains underexplored.

2.1.3 Modeling Users' Contextual Privacy Decisions

Many researchers have studied modeling users' decision-making process in the context of various types and recipients of the information. [83], while exploring the design parameters of social network site's privacy-settings UI (user interfaces), discovered about how the type of information and their specific recipients have significant effect on user's sharing tendency. In their study participants were asked to set their privacy settings on a custom made privacy settings UI of an imagined Facebook-like social network site by indicating which of their profile information they would share with whom. At the end of the study, they measured the users' interpersonal privacy concerns using a post-experimental questionnaire. In another user study, [84] validated the primitive idea of users' privacy calculus (i.e., costs vs benefits which measures the benefits of privacy allowances and the resulting costs [46]) and how it led them to disclose different types of information to different types of websites in a purpose-specific manner. They found that the perceived risk and perceived relevance of the disclosure depends on the interaction between the type of the information and the type of the website/recipient, and that this perceived risk and relevance decreases and increases disclosure, respectively. While both studies show how the perceived relevance and risk of the information—as well as the disclosure activity or intention—depend on the

type and recipient of the information, neither of these studies takes into consideration the impact of ephemeral situations (i.e., scenarios or contexts) on the participants' behavior.

In a contextual setup, [95] investigated the relative effects of the information recipient and the situation towards information disclosure. They conducted a study with 130 participants by providing them with two hypothetical situations (working lunch, social evening) and four inquirers/recipients (spouse, employer, stranger, merchant). They asked each participant to imagine using a mobile phone which is capable of collecting and sharing profile and location information to the requesting parties. Through a web-based questionnaire, they analyzed the user's preferences and found that "identity of the information inquirer is a stronger determinant of privacy preferences than is the situation in which the information is collected." However, they found that the situation is also an important determinant but only when the information inquirer is an employer. Even though they incorporated scenarios and recipients' roles in the study, the characteristics of the scenarios were unchanged and represent only two static situations. In this regard, their contextual behavior analysis is limited to these two situations only. Nevertheless, the above-mentioned work and other similar works have demonstrated the influence of various contextual factors on users' privacy behavior [123, 120].

2.1.4 Representing Contexts with Scenarios

One way of contextualizing a survey is to introduce various scenarios to the participants and ask them to respond to questionnaires linked to each of those scenarios [95]. However, one challenge in this regard is to create proper scenarios with an appropriate level of detail. Researchers from the area of the scenario-based surveys

have introduced many different approaches to create hypothetical scenarios using text, graphics, games, app interfaces, etc. [48, 89, 168, 174]. Among all of these, text-based scenarios are preferred in the case of surveying the participants. A set of methods have been well-established for the development of such scenarios, especially in the privacy survey domain, such as the factorial method, storytelling method, and claim analysis. The factorial method involves creating scenarios "based on a set of predefined factors that describe all or a subset of possible combinations seen in a situation or decision problem" [22]. These factors could be socioeconomic, behavioral, or clinical issues, defined as categorical variables with two or more levels. However, the number of factors and their levels are subject to be decided carefully. Otherwise, the number of combinations of factor categories increases very rapidly, which in turn increases the total number of unique scenarios. On the other hand, the storytelling method suggests creating a few illustrative scenarios, usually based on the experience of the members of the research team. In our work, we adopt the former method to create the scenarios while keeping the number of factors and their categories low.

2.2 Detecting Privacy Disclosure through Natural Language Processing (NLP)

In this section, we review the research works from the area of natural language processing and privacy that are focused on privacy disclosure detection [29, 67, 104, 13]. Traditional research carried out in this area mostly utilize lexicon based technique to leverage the linguistic resources such as a privacy dictionary to automate the content analysis of privacy-related information. Existing automated content-analysis tool

such as LIWC[1] is used with a specific set of privacy dictionaries. Vasalou et al. suggest such a method that utilizes a dictionary of individual keywords or phrases which are previously assigned to one or more privacy domains [160]. To create the dictionary, they sample from a wide range of privacy domains such as self-reported privacy violations, health records, social network sites, children's use of the Internet, etc. However, their technique solely relies on the predetermined sensitive keywords/terms which classify both a medical article (public) and someone's medical condition (private) as a private document. A similar approach was taken by Chakaravarthy et al. for a document sanitization task, where they represent a scheme that detects sensitive information using a database of entities [27]. The database contains different entities i.e., persons, organizations, products, diseases, etc. Each entity is also associated with a set of sensitive terms e.g., name, address, age, birth date, etc. Thus a set of terms is considered as the context of the entity. For example, the context of a person becomes his/her age, birth date, name, etc.

Researchers from the area of information theory leverage large corpus of words along with computational linguistics to identify sensitive information in text documents [134]. Information theory provides the necessary formula[2] for calculating the sensitivity score, otherwise known as the IC (Information Content) score of every term, based on the amount of information it contributes to a corpus. For example, in a database of employees, a term such as *handicapped* carries more information than the common terms such as job, manager, desk, office, etc. All such terms that exceed a threshold score β are considered as sensitive[3]. One of the advantages of this technique

[1] Linguistic Inquiry and Word Count

[2] $IC_{database}(t) = -log_2 \frac{document\_counts(t)}{total\_documents}$

[3] $sensitive_t = \{t_i \in document | IC_{database(t)} \geq \beta\}$

is that a finite collection of named entities is not required for the disclosure detection to be successful. However, this approach suffers from the same limitation as to the previous line of work. In other words, it does not consider any semantic information other than merely relying on the appearance of sensitive keywords. Other popular techniques such as Named Entity Recognition (NER), also known as entity chunking, entity identification, or entity extraction have also been used by many researchers to identify and classify private information in text documents [2]. This line of research is based on the sub-task of information extraction technique that aims to identify named entities (medical codes, time expressions, quantities, monetary values, etc.) and classify them into predefined categories in an unstructured text. Modern NER systems use linguistic grammar-based techniques, statistical models, machine learning, etc. Regardless of the underlying method, the NER based disclosure detection techniques also lack the capability of properly inferring the meaning from a text that could disclose someone's private information if a specific named entity is not detected (see examples 3 in Table 6.1).

Machine learning based techniques such as association rule mining [31], support vector machines (SVM), random forests [156], boosted Naive Bayes, AdaBoost, latent Dirichlet allocation (LDA), etc. have also been used to tackle similar tasks. Hart et al. used a novel training strategy on top of SVM to classify text documents as either sensitive or non-sensitive [59]. Caliskan et al. proposed a method for detecting whether or not a given text contains private information by combining topic modeling, named entity recognition, privacy ontology, sentiment analysis, and text normalization technique [26]. A combination of linguistic operations and machine learning is proposed by Razavi et al. to detect health information disclosure [128]. They first compile a

list of keywords related to a person's health information, and then apply keyword combinatorial web search. Alongside, they implement a machine learning layer to detect and learn any possible latent semantic patterns in the annotated dataset. Mao et al. studied privacy leaks on Twitter by automatically detecting vacation plans, tweeting under the influence of alcohol, and revealing medical conditions [104]. As the classifier model, they implemented two machine learning algorithms, Naive Bayes and SVM, based on the TF-IDF (Term Frequency Inverse Document Frequency) feature space. Their main research goal was to analyze and characterize the tweets in terms of who leaks the information and how. Therefore, in the paper, the focus was less on the architecture and performance of the disclosure detection model.

Bak et al. applied a modified LDA based topic modeling technique for semi-supervised classification of Twitter conversations that disclose private information [14]. This technique is also based on the distributions of terms/keywords across documents and corpus, which again does not consider word meaning inference. The limitations of most of the above-mentioned techniques are based on the fact that they solely rely on the existence of keywords and disregard word meaning inference from the text. We observe through our experimentation that these limitations, in some cases, result in missclassification. This is because, existence/lack of sensitive terms/keywords in a piece of text does not always result in disclosure/non-disclosure of private information.

In order to overcome these limitations, recent research works from the area of NLP and privacy have considered utilizing semantic meaning along with lexical and syntactic analysis, while designing and developing deep learning based models [110, 108, 157, 39]. Accordingly, there has been significant progress in the area of language

modeling through training complex models on enormous amounts of unlabeled data [161, 43]. All the tailored solutions are being outperformed by these generic models. Most importantly, the utilization of transfer learning and pre-trained models has shed light on this area of research. Dadu et al. proposed a predictive ensemble model by exploiting the fine-tuned contextualized word embedding, RoBERTa (Robustly Optimized BERT Approach) and ALBERT (A Lite version of BERT). The authors generated a small labeled dataset, containing Reddit comments from casual and confessional conversations. Through the ensemble implementation, they achieved a 3% increment in the F1-score from the baseline model. Therefore, after considering the importance of transfer-learning and also taking into account the significance of linguistic features, we propose several multi-input hybrid neural network models that utilize both transfer-learning and linguistics along with the metadata from the input text.

2.3 Modeling of Personalized Privacy Disclosure Behavior through Formal Method

Researchers from the field of privacy, decision-making, and personalization have shed light on the area of behavior modeling. They have been exploring how the psychological factors of humans relate to their concerns about their information privacy [166, 3, 96]. Accordingly, many behavioral theories have been established and adopted in the privacy management domain [7, 16, 58]. Theory of planned behavior (TPB) tells that people's behavior is directly determined by their behavioral *intentions*. These intentions are in turn influenced by their *attitude* (positive or negative evaluation of the decision), perception of the *subjective norms* (generally expected behavior

in their social group), and *perceived behavioral control* (ease or difficulty to perform the behavior). The theory also states that these constructs together determine an individual's behavioral intentions and provide a model to capture humans' decision-making behavior. Therefore, researchers from various areas (e.g., privacy, use of the internet, health, environmental psychology, etc.) have used TPB and demonstrated its effectiveness in predicting human behavior in terms of privacy decision-making [60, 159, 172, 34, 100].

Another privacy management theory that is relevant to our work is known as the theory of contextual integrity (CI) [16]. In the CI theory, privacy is formulated as an appropriate flow of information that conforms with the contextual informational norms (i.e., rules governing the flow of information in CI format). An example of a norm in the context of *health* could be: a husband usually shares his diagnosis result with his family doctor, or his wife but not with his friends or on social media. In this example, the husband is recognized as the data subject, and the sender, the doctor or wife as the recipient of the information, health as the information type, and the recipient will hold the information confidentially as the transmission principle. Based on the theory of (CI) [16], privacy is violated if the information is shared or transferred with friends or financial advisers, as they are not usually and explicitly included as part of the 'allowed' recipients of the information.

Consequently, many researchers have studied modeling users' privacy decision-making process in the context of various types and recipients of the information. Knijnenburg et al. discovered how the type of the information and their recipients have significant effect on user's information disclosing tendency [83]. In their study participants were asked to set their privacy settings on a custom-made privacy settings

UI of an imagined Facebook-like social network site by indicating which of their profile information they would share with whom. In another study [46], authors have examined the idea of users' privacy calculus (i.e., costs vs benefits) and how it led the users to disclose their different types of private information to different types of recipients (websites), in a purpose-specific fashion. Lederer et al. [95] investigated the relative effects of different recipients and the situations towards users' information disclosure intention. By surveying 130 participants, given two hypothetical situations, they found that situation is an important determinant and highly correlated with the information recipient.

Despite the existence of many behavioral theories and analyses, only a handful of works address the issue of personalized modeling of human behavior. Most importantly, a few of them acknowledge the issue of practical usability and application of the derived models. Joshaghani et al. extends the concept of CI theory and provide mathematical models that enable the creations and management of privacy norms by individual users [70]. They propose and develop a custom formal verification technique that ensures privacy norms are enforced for every information sharing attempt by the user. Similar to our transition system based formalism, Lu et al. proposed a technique that translates the privacy specification or requirements of web services to LTL formulas [98]. Then they create the privacy policy model by utilizing a privacy interface automata (PIA) that transforms the messaging structure extracted from the web service business process execution language into an automaton. Krishnan et al. propose a semi-formal approach that enforces privacy requirements by leveraging the role-based access control technique along with LTL formulas [88]. Grace et al. propose a technique for modeling user-centric privacy management using labeled transition

systems. The goal of this model is to compare the user's privacy preferences with the privacy policies of the cloud service provider [54]. Thus the users 'can be informed of the privacy implications of the services and warned of potential privacy breaches. However, they mentioned two limitations— i) the requirement of human intervention for creating the initial model, ii) limited extensibility and scalability.

In our work, we address many of the above-mentioned limitations and open questions by representing personalized situational behavior, proposing a technique for automatic translation of activities to FSM, demonstrating the practical usability, and describing the scalability of this formal approach.

CHAPTER 3:

PRIVACY AS A PLANNED BEHAVIOR: EFFECTS OF SITUATIONAL FACTORS ON PRIVACY PERCEPTIONS AND PLANS[1]

To account for privacy perceptions and preferences in user models and develop personalized privacy systems, we need to understand how users make privacy decisions in various contexts. Existing studies of privacy perceptions and behavior focus on overall tendencies toward privacy, but few have examined the context-specific factors in privacy decision-making. We conducted a survey on Mechanical Turk (N=401) based on the theory of planned behavior (TPB) to measure the way users' perceptions of privacy factors and intent to disclose information are affected by three situational factors embodied by hypothetical scenarios: information type, recipients' role, and trust source. Results showed a positive relationship between subjective norms and perceived behavioral control, and between each of these and situational privacy attitude; all three constructs are significantly positively associated with intent to disclose. These findings also suggest that, situational factors predict participants' privacy decisions through their influence on the TPB constructs.

[1]Nuhil Mehdy, M. Ekstrand, B. Knijnenburg, H. Mehrpouyan, "Privacy as a Planned Behavior: Effects of Situational Factors on Privacy Perceptions and Plans," 2021 *29th Conference on User Modeling, Adaptation and Personalization (UMAP 21)*, ACM

3.1 Introduction

Users' decision to share their personal information, and the perceptions of risk that inform this decision, vary from situation to situation. Situations consist of various factors such as the information type, recipient of the information, and the trust source behind the motivation for sharing. Past research has not paid much attention to how these factors can be used to model and predict users' contextual privacy concerns and decisions. This is an important shortcoming, as decision research suggests that users' privacy preferences are malleable rather than stable and that privacy behavior may vary based on situational and contextual factors [68, 141, 81]. Moreover, an individual's privacy expectations depend on the contexts in which the user is sharing information [111, 124, 108, 70].

In order to understand, model, and possibly predict human privacy behavior in various situated environments, there have been several factors and parameters documented to influence users in their privacy decisions. The theory of planned behavior (TPB) [7], an extension of the theory of reasoned action [158], is a behavioral theory that helps modeling users' perceptions and plans. However, most privacy research based on this theory has either studied single situations, or has considered a very limited set of situational factors [60, 133]. As a result, understanding the characteristics and impact of various situational factors on users' privacy decisions is still an active area of research.

In this work, we study users' situational privacy decisions, through a scenario-based survey with 401 participants, each responding to several of 48 different scenarios. Each data point consists of responses to a set of questionnaires that measure participants' attitudinal evaluations of each scenario as well as their perceptions and

intention to disclose private information under the specified situation. Alongside the scenario-specific questions, participants responded to a set of general attitude questions to elicit their general attitude towards information disclosure. We perform a path analysis to model participants' privacy perceptions and plans, taking into consideration their attitudinal evaluations on *subjective norm*, *perceived behavioral control*, and *attitude* by manipulating three situational factors: information type, recipient role, and trust source. The results from the analysis reveal how users make privacy decisions in various situations, and how the situational factors have significant effects on users' perceptions of privacy factors and intention to disclose potentially private information. This paper is the first to our knowledge to combine the Theory of Planned Behavior with a contextual approach to privacy modeling. This study also contributes several insights to the area of user-tailored privacy modeling and personalized privacy systems [81], through the following research questions:

1. How do users' subjective perceptions of TBP constructs differ in different informational situations?

2. How do situational perceptions affect users' intent to disclose information?

3. How do users' situational perceptions and intents relate to their general privacy attitudes?

3.2 Survey Methodology

The overall flow of our experiment can be divided into three main steps: i) recruitment and consent ii) capturing scenario-specific perceptions and planned decisions iii) general attitude survey (Figure 3.1). After consenting to the study, a participant

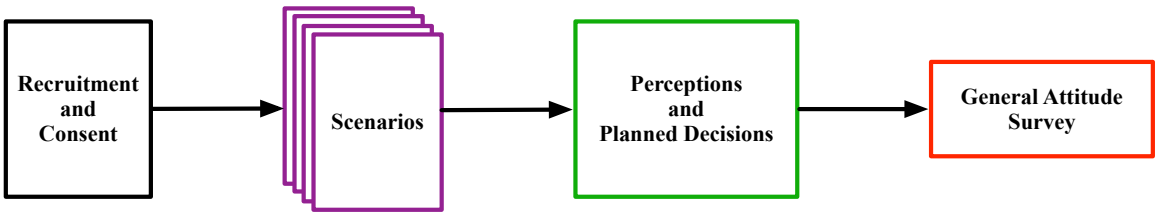


Figure 3.1: Overview of the experimental flow.

is assigned a set of 8 random hypothetical scenarios and asked to respond to those scenarios one after another. Each scenario gives the participant a situation in which he/she must decide whether or not to share a piece of information. This incorporates the situational factors on which participants might have a degree of reliance for their perception and decision towards disclosure intention (see Section 3.2.1). A participant has to read a given scenario and respond to all of the corresponding questions before proceeding to the next assigned scenario.

In the next step, the participant takes a short survey to capture their general privacy attitudes independent of any particular scenario. This step is designed to capture such perceptions that are assumed to be stable over time and do not usually change based on any situation. There is another final step for collecting the demographic information of the participants, in which participants are asked to optionally input their gender, age group, country of residence, and the duration of residence in that country. It is worth mentioning that the presented scenarios are hypothetical; none of the participants' personally identifiable information is collected in any step of the survey, explicitly or implicitly.

3.2.1 Factor Manipulation

We manipulate 3 situational factors in order to measure their effect on participant response:

Information Type (IT) The general category information that may be disclosed. Each scenario is about one of three categories: health, finance, or relationship.

Recipient's Role (RR) The kind of recipient the information may be disclosed to, with their relationship to the participant. We use four roles: family, friend, colleague, and online (e.g., discussion forum).

Trust Source (TS) Where the idea of disclosing this information to this recipient came from. We test four trust sources: family, friend, expert (e.g., physician or financial adviser), and online (e.g., searching the web).

This choice of factors is partly inspired from the theory of contextual integrity (CI) [118, 16]. The CI theory provides the ground of informational norm where, norm is formulated as a tuple of access permission (ρ, τ), environmental conditions (ψ), and transmission principle (η). Hence, a norm, n is represented as: $n = ((\rho, \tau), \psi, \eta)$ where, n = Informational norm, ρ = Recipient's Role, τ = Information type, η = Transmission principle. These factors yield a total of 48 (3*4*4) unique situations. Every situation and the associated questionnaire is intended to measure the situational privacy perceptions of the participant through 3 constructs: i) attitude ii) subjective norm iii) perceived behavioral control

3.2.2 Scenario Generation

For each combination of situational factors, we wrote a scenario in which a trust source encourages the participant to share information with a recipient. To minimize extraneous variability, we made each scenario as similar as possible while presenting the combination of factors in a natural and coherent manner. As an example, the scenario for *health* as information type, *friend* as trust source, and *family member* as

recipient's role is:

> Your doctor called and told you that your lab results came back positive for a disease. One of your friends suggested discussing the situation with a family member and asking their support, saying it could be helpful.

Changing the trust source from friend to family and recipient's role from family to online yields another scenario:

> Your doctor called and told you that your lab results came back positive for a disease. A family member suggested asking other patients and doctors on an online discussion forum, saying they have found it helpful for dealing with their similar condition.

In this study, the domains of the scenarios are health, finance, and relationship. This means, we have generated 3 sets of scenarios for these three types of information. Each of these sets contains 16 different scenarios (i.e., 4 RR x 4 TS values) resulting in a total of 48 scenarios. For each scenario, the participants answered a set of questions to measure their perception of TPB constructs in that scenario and indicated whether or not they would share the information.

3.2.3 Scenario Randomization

As discussed earlier, every participant is assigned a set of 8 random scenarios with associated questionnaires. To ensure a minimum level of variability within each user's situations (and therefore responses), we used rejection sampling to require that each user's 8 scenarios covered all 11 distinct factor levels at least once. Redrawing a fresh, independent set of 8 scenarios if a user's initial assignment excludes a level ensures maximal statistical independence subject to our inclusion requirement. We

further randomly order scenarios for each participant to avoid order effects. Also, we implicitly account for the variability of judgements of the questions and scales across the participants by setting random per-user intercepts while doing the analysis.

3.2.4 Testing the Experiment

We piloted the experiment and surveys with 6 colleagues from our research lab. Their feedback helped fix issues in the survey application, user-experience/user-interface, and clarity of the scenarios and questions. We then soft-launched the survey on Amazon Mechanical Turk with an initial round of 10 participants to collect information on the average time needed to complete the survey and estimate total survey cost.

3.2.5 Participants

We recruited the participants for the final survey via Amazon Mechanical Turk, an online crowd-sourcing marketplace. We filtered for Workers from the USA with a good reputation (i.e., at least 95% HIT approval rate and 50 hits approved) who are at least 18 years old. We paid $2.00 per survey based on pilot trials indicating Turkers could complete it in about 15 minutes.

3.2.6 Data Collection and Cleaning

We employed a number of filters to ensure the quality of the data. First of all, we capture the time a participant spent on each scenario step and removed the data points (i.e., responses associated with a specific scenario) from our analysis if the spent time was too low (less than 15 seconds per scenario) to be realistic. Secondly, we embedded attention check questions randomly in between survey questions on two surveys per participant, and removed 9 data points for failing the attention check. Since participation is anonymous and therefore a participant could potentially submit

several responses, we restricted this incident by setting a browser cookie for 3 days after a successful submission.

We converted the 5-point scale responses to TPB questions (ranging from *Strongly Disagree* to *Strongly Agree*) into a numeric format (1 to 5). We represent the *Share* and *Not Share* options for the final decision question in logical numeric form, 1 and 0. We dummy-coded categorical variables for the situational factors. We then computed a standardized scale-score for each TPB construct by taking the mean of the responses on its questions (see Section 3.3.2), after inverting negative questions, so that 5 represents the opinion most in favor of sharing for each question.

3.3 TPB-Based Questionnaire and Path Model

As previewed in Section 2.1.1, we designed our survey to measure participants' behavioral intention and their situational perception of three constructs from TBP: attitude (A), subjective norm (SN), and perceived behavioral control (PBC). We followed the scenario-specific questionnaires with a short survey to assess participants' general attitude towards privacy. We integrated the TPB constructs, manipulated factors, and general privacy attitude into an initial path model shown in Figure 3.2. The colors on the figure follow the convention of Knijnenburg et al.'s evaluation framework [82], where purple = manipulations, green = subjective evaluations, red = personal characteristics, and blue = behavior. We evaluate this path model through a causal modeling technique called *path analysis* to determine if our causal model fits the survey data well. Note that path analysis is "not intended to discover causes but to shed light on the tenability of the causal models that a researcher formulates" [125]. We apply this technique to examine the relationships between the observed variables in terms of the strength and direction of the path beta coefficients.

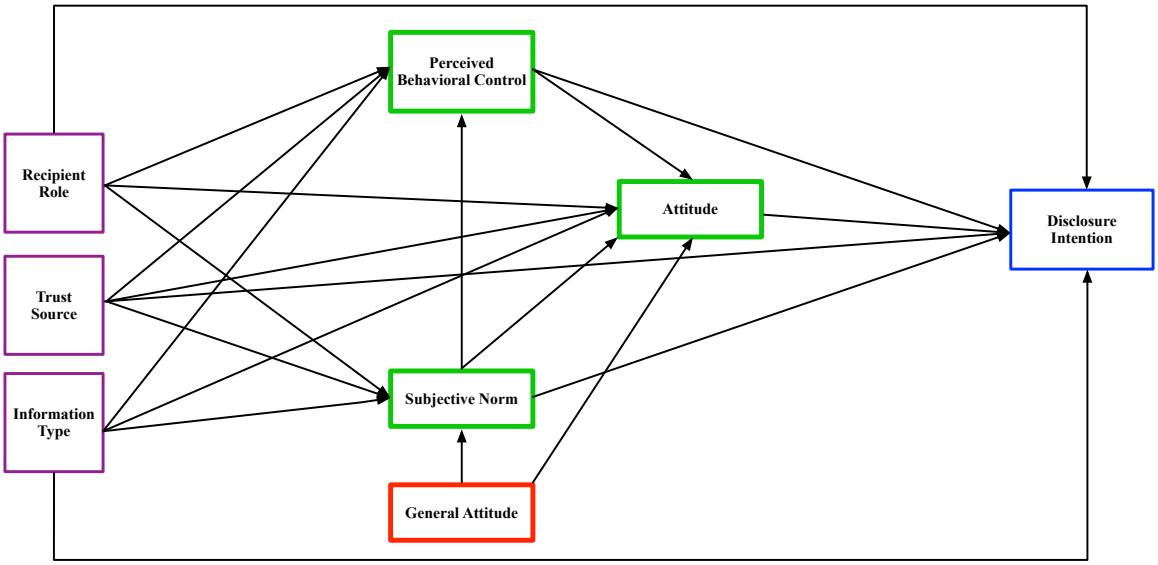


Figure 3.2: The initial path model.

3.3.1 Model Specification

The dummy variables representing the three scenario parameters—information type, recipient's role, and trust source—comprised the *exogenous* variables (variables that have arrows outbound from them and not caused by any other variables of the model [147]) in the preliminary path model, together with general attitude. Trust source was eliminated from the final model because of its non-significant association with the TPB constructs. In our initial model, the exogenous variables were causally related to attitude, perceived behavioral control, and subjective norm, although some of these relations (e.g., from information type to perceived behavioral control) were removed due to a lack of significance. Relationships among attitude, perceived behavioral control, and subjective norm were also modeled. Finally, all variables were causally related to disclosure intention, although only the attitude, perceived behavioral control, and subjective norm were found to be significant. The final model has a

total of 27 free parameters, and 28 fixed parameters whose values are estimated from the data.

We fit the model with Mplus, a statistical analysis tool for conducting the analysis as well as constructing the diagram of our path model [116].

3.3.2 Questionnaire

The survey contains two sets of questions - *i) scenario specific questions (12 in total) ii) general attitude questions (4 in total).* The first set of 12 questions are repeated for each of the 8 assigned scenarios to each participant. The second set of questions are presented at the last step of the survey. The scenario-specific questionnaire is inspired by [60], which operationalized the constructs in the theory of planned behavior [7]. The second set of questions is inspired by prominent privacy research [24, 3].

The following questions were asked once per scenario:

1. *Attitude (Cronbach's alpha: 0.68)*
    (a) I would benefit from sharing this situation. (Scale: Completely disagree (1) to Completely agree (5))
    (b) I am concerned about where this information would be stored or recorded if I shared it with *$recipient*. (1-5, reversed)
    (c) I do not expect any significant risks if I share this situation. (1-5)
    (d) I have concerns about who will learn about this situation. (1-5, reversed)
2. *Subjective Norm (Cronbach's alpha: 0.79)*
    (a) I think my friends or family would share in this situation. (1-5)

(b) A friend or family member would likely suggest that I disclose this situation. (1-5)

(c) My friends would approve of me disclosing this situation. (1-5)

(d) Some people in my life would disapprove if they knew I shared this situation. (1-5, removed from the scale)

3. *Perceived Behavioral Control (Cronbach's alpha: 0.66)*

(a) I have control over how my information will be used after I share it in this situation. (1-5)

(b) I trust the recipient of my information to honor my wishes if I ask them to keep my situation a secret. (1-5, removed)

(c) Sharing this situation would put me at risk. (1-5)

4. *Disclosure Intention*

(a) What would you do in this scenario? (Scale: Not share (0) or Share (1))

The following questions were asked once per participant:

1. *General Attitude (Cronbach's alpha: 0.68)*

(a) In general, I am concerned about threats to my personal privacy. (1-5, reversed)

(b) I am generally concerned about my privacy while using the internet. (1-5, reversed)

(c) I believe other people are too concerned about online privacy issues. (1-5, removed)

(d) I think I am more sensitive than others about the way my contacts handle information I consider private. (1-5, reversed)

We performed Cronbach's alpha test to measure the items' scale reliability. Thus, item (d) was removed from the subjective norm scale because of its negative effect on the alpha score. We removed item (c) from perceived behavioral control, and item (c) from general attitude because of the same reason. Items (b) and (d) in the attitude scale were reversed while calculating their score because of their negative phrasing. All items in the general attitude scale were reversed to align this factor with the context-specific attitude.

3.4 Results

This section describes the path analysis results in detail. First, we talk about the descriptive analysis and the quality of the model fit. Then we describe the direct and indirect effects of the factors and constructs in subsequent sections. Figure 3.3 depicts our final path model.

3.4.1 Descriptive Statistics

Table 3.1 reports the demographic information of the participants. We share this information not because these are relevant factors in this context but for those who may attempt to reproduce these results with a similar setup. Figure 3.4 reports the differences in attitude, subjective norm, perceived behavioral control, and disclosure intention between the different value of the scenario parameters "information type" and "recipient role", including standard error bars. For example, we can see how the participants perceive a higher level of behavioral control when the recipient is a family member or friend than that of a colleague or online platform.

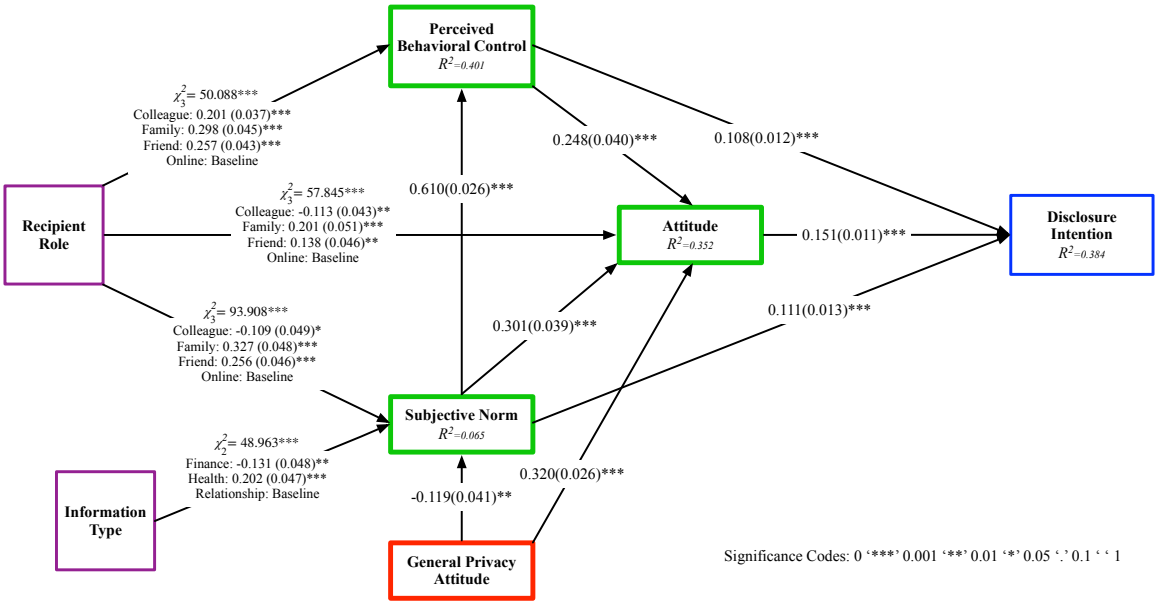


Figure 3.3: Path model results. Paths that are non-significant ($p > .05$) are removed from the model.

Table 3.1: Demographic information of the participants.

Constructs	Distribution
Gender	Man: 252 Woman: 144 Not Answered: 3 Non Binary: 1 Woman,Man: 1
Age	18-30: 108 31-40: 148 41-50: 75 51-60: 49 60+: 19 Not Answered: 2

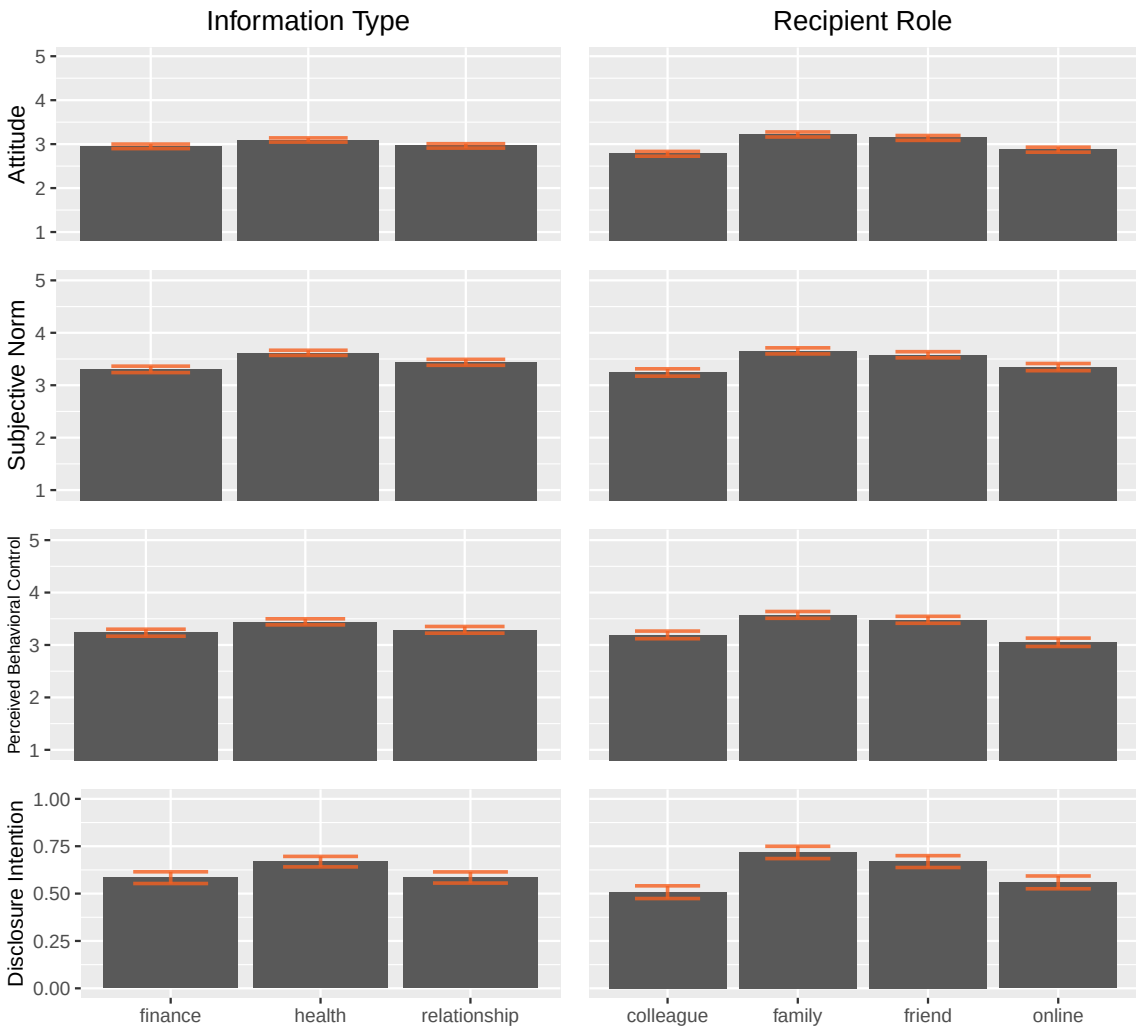


Figure 3.4: Constructs vs Mean Scale-score based on Information Type and Recipient's Role.

3.4.2 Model Fit

Figure 3.3 depicts the final results of the path model analysis in detail. The model fits the data very well with $\chi^2_{11} = 12.017$, $p = 0.3623$, $CFI = 1.0$, $TLI = 0.99$, $SRMR = 0.008$, $RMSEA = 0.005$, 90% $CI = 0.000$ to 0.020. A non-significant χ^2 value ($p > .05$) is indicative of a path model that fits the data well [137]. Also, the comparative fit index (CFI) and Tucker-Lewis index (TLI) values, which range from 0 to 1 show near-perfect scores. Moreover, the relationships in the model explain 38.4% ($R^2 = 0.384$) of the variance in disclosure intention, 35.2% of the variance in attitude, 6.5% of the variance in subjective norm, and 40.1% of the variance in perceived behavioral control.

3.4.3 Effect of the Scenario Parameters on TPB Constructs

This section describes the significant effects of the scenario parameters (recipient role and information type) on the TPB constructs (the privacy perceptions of the user). These effects are measured by the paths from the purple (square) boxes to the green (rectangular) ones in Figure 3.3.

1. The recipient's role in the scenario has a significant influence on perception of behavioral control. Compared to "people online", participants are estimated to perceive significantly more control when the recipient is a colleague (0.201 SD higher), a family member (0.298 SD higher) or a friend (0.257 SD higher).

2. Likewise, the recipient's role in the scenario has a significant influence on attitude. Compared to "people online", people are estimated to have significantly more positive attitude toward disclosure when the recipient is a family member (0.201 SD higher) or a friend (0.138 SD higher), but more negative attitude

when the recipient is a colleague (0.113 SD lower).

3. The recipient's role in the scenario has a significant influence on subjective norm. Compared to "people online", participants are estimated to believe that individuals close to them would be more likely to agree with the scenario when the recipient is a family member (0.327 SD higher) or a friend (0.256 SD higher), but less when the recipient is a colleague (0.109 SD lower).

4. The information type in the scenario has a significant influence on subjective norm. Compared to "relationship information", participants are estimated to believe that individuals close to them would be more likely to agree with the scenario when the information type is health (0.202 SD higher) but less reliance when the information type is finance (0.131 SD lower).

3.4.4 Effects between General Attitude and Situational Perceptions

We now turn to the relationships between constructs, both situational TPB constructs and the influence of general attitude on these constructs. The following effects refer to the paths among the green (rectangular) boxes and between the red (rectangular) box and the green ones in Figure 3.3.

1. The participants' perceived subjective norm regarding the scenario is positively associated with their perception of behavioral control. A 1 SD difference in subjective norm results in an estimated 0.610 SD difference in perceived behavioral control.

2. Participants' subjective norm is also positively associated with their attitude towards disclosure. A 1 SD difference in subjective norm results in an estimated 0.301 SD difference in attitude.

3. The perception of behavioral control of the participants regarding the scenario is positively associated with their attitude towards disclosure. A 1 SD difference in perceived behavioral control results in an estimated 0.248 SD difference in attitude.

4. The participants' general attitude is positively associated with their situational attitude towards disclosure. A 1 SD difference in general attitude results in an estimated 0.320 SD difference in attitude.

5. General attitude is also negatively associated with perceived situational subjective norm. A 1 SD difference in general attitude results in an estimated -0.119 SD difference in perceived subjective norm.

3.4.5 Effects of Situational Perceptions on Disclosure Intention

This section briefly describes the significant effects between the situational TPB constructs (the privacy perceptions of the user) and users' disclosure intention. The following effects refer to the paths between the green (rectangular) boxes and the blue (rectangular) one in Figure 3.3.

1. Participants who perceived a higher level of behavioral control were more likely to engage in the disclosure described in the scenario. Particularly, the odds of

disclosure of participants who have a 1 SD higher level of perceived behavioral control are estimated to be 11.4% higher.

2. Participants who have a higher level of perceived subjective norm were more likely to engage in the disclosure described in the scenario. Particularly, the odds of disclosure of participants who have a 1 SD higher level of perceived subjective norm are estimated to be 11.7% higher.

3. Participants who have a more positive attitude were more likely to engage in the disclosure described in the scenario. Particularly, the odds of disclosure of participants who have a 1 SD higher level of attitude are estimated to be 16.2% higher.

Although not directly comparable, it's worth mentioning a comparison with the results from [60] in this section while showing the relationships between the TPB constructs and disclosure intention. According to their analyses which take into account only the stable factors, an individual's intent to disclose is influenced primarily by a subjective norm and subsequently by attitude, not significantly by perceived behavior control. In contrast, our study shows the order of significant influence of the TPB constructs to disclosure intention as, attitude > subjective norm > perceived behavioral control. It should be noted that in our study, the TPB constructs are already affected by the situational factors.

3.4.6 Total Effects of the Scenario Parameter on Disclosure Intention

All effects of scenario parameters on disclosure intention were fully mediated by perception of TPB constructs—that is, after controlling for scenario effects through TPB

constructs, there were no statistically significant residual effects of scenario parameters on disclosure intention. This section describes the total significant (indirect) effects of the scenario parameters on the users' disclosure intention. The following effects do not refer to any direct paths between the purple (square) boxes and the blue (rectangular) one in Figure 3.3. Rather, they refer to the paths from the leftmost boxes to the rightmost box via the mediator rectangular boxes in between. These total effects describe *how* users' intention changes from one scenario to another; the mediating TBP factors provide an explanation for *why*. The latter may help with future generalizability.

1. With regard to the recipient's role in the scenario, compared to the recipient "people online", the odds of disclosure were estimated to be 16.6% higher when the recipient was a family member and 12.9% higher when the recipient was a friend. Both of these differences were significant ($p = 0.000$ and $p = 0.000$, respectively). There was no significant difference between the recipient "people online" and a colleague.

2. With regard to the type of information, compared to relationship information, the odds of disclosure were estimated to be 3.1% lower when the scenario involved financial information and recipient was a family member and 5.1% higher when the scenario involved health information. Both of these differences were significant ($p = 0.007$ and $p = 0.000$, respectively).

3.5 Discussion

The results from our path analysis show how users make privacy decisions in various situations: the situational factors have significant effects on users' perceptions of pri-

vacy factors, which in turn have an effect on their intention to disclose their private information. Unlike most existing studies of privacy perceptions and behavior modeling, we developed a set of unique scenarios by manipulating parameters to imitate various situations and used a TPB-based model to introduce mediating factors that explain the effects of these situational factors on participants' disclosure intentions. This situation-specific extension of the TPB fulfils our initial goal of understanding users' contextual privacy decision-making process.

This study reveals that the recipient's role in the scenario has a significant influence on peoples' perception of behavioral control, their attitude, and subjective norm (RQ1). People are estimated to perceive a higher level of behavioral control when the recipient is a family member, a friend, or a colleague than when the recipient is people online (e.g., social media, forum, etc.). Likewise, people are estimated to have a more positive attitude toward disclosure when the recipient is a family member or a friend than people online, but a less positive attitude when the recipient is a colleague. Users' subjective norm also shows similar order of perceptions. As a result of these effects, people are more likely to disclose their information to friends and family than to colleagues or people online.

The information type in the scenario also has significant influence on participants' subjective norm. The model shows that people believe that individuals close to them would be most likely to agree with the scenario when it involves health information, followed by relationship information, and finally financial information. These differences propagate to small differences in disclosure intentions as well.

The results from the analysis also show that participants' perceived subjective norm regarding the scenario is positively associated with their perception of behav-

ioral control and attitude towards intention to disclose (RQ2). In other words, one can make a hypothesis that when users perceive an expectation to share, they also expect that sharing to be respected? Likewise, their perception of behavioral control is a good predictor of their attitude. Moreover, from the results, we can see the positive effects of these three constructs on users' disclosure intention. Users' attitude has the strongest effects on their disclosure intention relative to the other two constructs. Participants with a higher level of positive attitude were more likely to engage in the disclosure described in the scenario. Section 3.4.5 contains specific detail of these effects. Additionally, our results reveal the significant influence of general attitude on some TPB constructs (RQ3). Participants' general attitude is positively associated with their situational attitude towards disclosure. In contrast, general attitude is negatively associated with perceived situational subjective norm.

Most importantly, our study demonstrates that the effects of the contextual parameters (the recipient's role and information type) on the users' disclosure intention was fully mediated by participants' attitude, subjective norm, and perceived behavioral control. As such, these TPB constructs serve as significant and sufficient mediators explaining why users disclose more information in some scenarios than in others. These findings contribute important insights to the area of user-tailored privacy modeling and personalized privacy systems by providing a quantitative analysis of the privacy decision-making factors.

3.5.1 Limitations

Even though path analysis is often referred to as a causal inference technique [15], readers should be advised that this model reveals the predictive properties between the factors and constructs. These properties are measured in terms of path coeffi-

cients. Therefore, our path analysis shows how the hypothesized model fits the survey data which in turn aims to explain users' privacy decision-making process. We also acknowledge that we are only manipulating a few levels per factor in our study, and there could be much more granularity in the information type, recipient's role, and trust source factor; future work should explore this. Additionally, since the results reveal significant relationships between situational factors and disclosure intention, we feel the necessity to integrate additional factors in future studies.

We also note that our scenarios had a hypothetical nature, and hence did not measure actual disclosure but rather users' *intention* to disclose their private information. This is a limitation that our work shares with many other privacy studies [172], especially in light of the "privacy paradox" which shows a discrepancy between disclosure intentions and behaviors, as behaviors tend to be influenced by extraneous factors like default settings and choice framing [9]. Arguably, though, the absence of such extraneous influences makes users' disclosure intentions a more honest representation of their privacy preferences.

3.6 Conclusion

In this paper, we have presented the results of a scenario-based survey to understand users' *situational* privacy perceptions and disclosure intentions. These results constitute a contextualized understanding of users' privacy behaviors, connected to the Theory of Planned Behavior, and provide new insights that can help build future user-tailored privacy models. The impact of various situational factors on users' privacy decision is still an active area of research; one particular need is more study of the gap between users' intention versus reported and actual behavior. In future work, we plan to bridge the gap between intention and behavior by incorporating reported

or actual behavior in the model. We also plan to evaluate the predictive power of the current path analysis by surveying a new sample of users. Moreover, we plan to increase the sample size significantly and employ machine learning based algorithms along with the statistical approaches, as a means to compare various analysis methods for explaining contextual privacy behavior. For now, we can advise the user-modeling community to take the recipient and information type into account when modeling users' situation-specific privacy concerns, and to perhaps build these models not as a uni-dimensional construct, but to include aspects of behavioral control, social norms, and attitude, as suggested by the Theory of Planned Behavior.

CHAPTER 4:

PRIVACY DISCLOSURES DETECTION IN NATURAL-LANGUAGE TEXT THROUGH LINGUISTICALLY-MOTIVATED ARTIFICIAL NEURAL NETWORKS[1]

An increasing number of people are sharing information through text messages, emails, and social media without proper privacy checks. In many situations, this could lead to serious privacy threats. This paper presents a methodology for providing extra safety precautions without being intrusive to users. We have developed and evaluated a model to help users take control of their shared information by automatically identifying text (i.e., a sentence or a transcribed utterance) that might contain personal or private disclosures. We apply off-the-shelf natural language processing tools to derive linguistic features such as part-of-speech, syntactic dependencies, and entity relations. From these features, we model and train a multichannel convolutional neural network as a classifier to identify short texts that have personal, private disclosures. We show how our model can notify users if a piece of text discloses per-

[1]Nuhil Mehdy, C. Kennington, H. Mehrpouyan, "Privacy Disclosures Detection in Natural-Language Text Through Linguistically-motivated Artificial Neural Network," 2019 *2nd EAI International Conference on Security and Privacy in New Computing Environments (SPNCE 19)*

sonal or private information, and evaluate our approach in a binary classification task with 93% accuracy on our own labeled dataset, and 86% on a dataset of ground truth. Unlike document classification tasks in the area of natural language processing, our framework is developed keeping the sentence-level context into consideration.

4.1 Introduction

In this era of global communication, individuals often share stories, news, and information with each other. It is not easy for these users to keep track of what information they have shared, whether or not that information was a private disclosure, and to whom they shared that information. While the importance of user-centric privacy management systems is widely studied [101, 102, 113], only some of these works are concerned with real-time text analysis and identifying text that contains private information. An important step in constructing an effective privacy management system is to concentrate on identifying and discriminating private information from public information.

For example, a very common medium of social communication between people is messaging using text; e.g., email, SMS/text messages, chat, social media, etc. While interacting, people sometimes disclose personal and sensitive information, unintentionally. For example, a sentence, *Let's meet at the Joe's Coffee Shop tonight at seven* is disclosing someone's meeting place along with the time. Whether or not these disclosures are intentional, it could potentially be an unwanted security threat and cause for alarm–or for harm. This example illustrates a common problem in a multitasking environment where users are simultaneously using in both public and private communication mediums. Our approach serves as an automated privacy check in these kinds of situations, warning individuals regarding risky communications in

both private and public contexts. This framework could also be effective while processing large amounts of off-line text documents. An example case study could be filtering out all the privacy disclosures from a batch of documents that belongs to a person before its disposal or archival.

Privacy concerns exist wherever personally identifiable information (e.g., name, address, age) or other sensitive information (e.g., health, finance, mental status) is involved [87]. Therefore, improper disclosure control can be the root cause for many privacy issues and the negative consequences of disclosing information could be immense [32]. A recent data scandal involving Facebook and Cambridge Analytica shows how personally identifiable information of up to 87 million Facebook users influenced voter opinions [143, 61].

The requirements for privacy measures to protect sensitive information about organizations or individuals have been researched extensively [24, 103, 139, 69]. One approach to protect the disclosure of private information is to detect them in textual data. However, automating the process of classifying private information prior to their disclosure is challenging [2]. One of the difficulties results from the volume of textual data that would need to be processed, and further the automation process is complicated even more by the number of real-time requirements that need to be analyzed [6, 142]. Moreover, it remains a challenge to analyze and dissect the details of private information from the text data due to the ambiguities that arise from natural language [56].

In this paper, we identify a potential approach that brings this challenge within reach: recognizing disclosures in a piece of text, which could be a short phrase (i.e., a sentence) within a longer content (i.e., a paragraph or document). Specifically, we

focus on identifying whether or not sentences have disclosures in them. Our approach enriches text data with linguistic features such as part-of-speech tags, syntactic dependency parse information, and entity relation information using off-the-shelf language processing tools. We then use these features to train a Convolutional Neural Network (CNN) to learn a mapping from the features to a binary label: disclosure/non-disclosure. This is a structured approach to train a machine learning model for detecting privacy disclosures and then automating that knowledge to classify certain types of privacy breaches.

The contributions of this paper can be summarized as follows-

- **Sentence level privacy disclosure identification:** While there exist similar techniques for classifying an entire document as private (i.e., confidential) or public, most of these approaches rely only on the existence of the privacy-related keywords in a document regardless of their semantics. In this paper, we consider detecting privacy disclosure at a sentence level, which is based on not only the existence of privacy-related keywords (i.e., disclosure related entities) but also on the valid grammatical structure of each sentence. This reduces false positive results by verifying the construction of a statement.

- **Disclosure Related Entity recognizer:** A Disclosure-Related Entity Recognizer (DRER) is developed by extending a trainable Named Entity Recognizer (NER) model. The developed DRER is later utilized to prepare a unique labeled dataset as well as to provide tagged entities for learning word embedding (i.e., similarities among disclosure related entities).

- **Case study and performance comparison:** We represent a comparison of the efficiency of different neural network architectures to detect privacy disclosure. Further, the proposed framework was evaluated to other similar datasets for a baseline comparison.

4.2 Methodology

In this paper, we leverage a multichannel convolutional deep neural network (DNN) to utilize lexical and sentence-level features. Our model takes all of the word tokens, part-of-speech tags, and dependency parse tree information of a sentence as input. First, lexical analyses are done at sentence-level. Then, the tokens are transformed into word vectors by learning word embeddings. Later, these features are concatenated to form the final feature vector. Finally, sentence-level structure, and privacy-related keywords are learned using the convolutional approach.

In this paper, privacy-related keywords are defined as disclosure related entities (DREs). These fall into the super set of all possible named entities (NE) but contextually different (i.e., not all Named Entities are Disclosure Related Entities by our definition). We develop a DRE recognizer by extending an off-the-shelf NE recognizer tool to assist the proposed model.

Definition 1 *(**Disclosure Related Entities**) - Let sentence S be a set of words, $S = \{w_1, w_2, \cdots, w_n\}$. A word w_i is considered to be a DRE if it indicates private information such as the name of disease, amount of debt, location of meeting, time of outing, etc.*

However, dis-joined existence of such entities in any random part of a sentence does not always prove the occurrence of a valid disclosure of private information *(e.g.,*

My son nothing morning no sense makes spoofing not $100 dollars). A sentence has to carry a reasonable meaning after being constructed by disclosure related entities (DRE) *(e.g., We are planing to leave for Paris on 31st December in early morning).* Moreover, non-machine learning methods seemed to perform well based on rules and reference datasets, but they are not scalable and adaptable when the time comes to analyze large amount of data. In order to overcome these challenges, this paper employs a framework which is based on typical convolutional neural network with extended capabilities. It first looks for disclosure related entities in a sentence, retrieves syntactic information, identifies grammatical validation, learns semantic information, and then determines the occurrence of disclosure or non-disclosure of information.

4.2.1 Data

The proposed framework consists of a neural network model that requires labeled data to learn patterns of disclosure and non-disclosure sentences from text data. Unfortunately, no particular data set with ground truth (i.e., set of sentences labeled as disclosure/non-disclosure) is available so far to work with. Therefore, after collecting textual data, we use a state-of-the-art Natural Language Processing (NLP) Toolkit named Spacy [63] to conduct a preliminary labeling (i.e., labeling raw dataset for training) of the dataset as well as to pre-process before feeding into the DNN model. The left section of Figure 4.3 demonstrates the usage of the NLP Toolkit for both data labeling and pre-processing; the following subsections describe the process in detail.

4.2.2 Data Collection

In order to collect data from different domains, we consider online platforms where people post reviews, ask questions, post tweets, and discuss from a first-person

perspective. Online forums like medical, psychiatric, and relationship communities mostly contain private information through users' conversations. However, we also wanted to see whether private information is disclosed by a user unintentionally in public forums (e.g., Stackoverflow, Amazon). This is why we introduce domain diversity here to give the model more generalized data. We sampled the same number of user posts from each domain, such as medical forums, social sites, food reviews, place and service reviews etc. All of the domains are selected randomly. This is summarized in Table 4.1. All the posts are written in the English language, and each of them is comprised of 4 to 15 sentences. The average sentence length throughout the whole data set is 9 words. As this research requires data that are related to privacy, we carefully avoided any sensitive resource that could have caused privacy violation. Anonymity has also been assured while collecting these data sets from reliable public sources.

Table 4.1: Summary of data sources.

Source	**Amount of Posts**
Medhelp Forum Posts [171]	3000
Amazon Product Reviews [41]	3000
Amazon Food Reviews [106]	3000
Hotel Reviews [40]	3000
Place of Interest Reviews [51]	3000
Psychriatic Forum Posts [114]	3000
Twitter Posts [122]	3000
Stack Overflow Questions [50]	3000
Total	**24000**

In each of the above-mentioned domains, people shared their views, feedback, or comments in a set of sentences (i.e., a product review, a twitter status, a question regarding health). Thus they expressed their overall opinions about a product, lo-

Table 4.2: Example disclosure and non-disclosure sentences

	Text	Is Disclosure
1	I have been living in W Boise Avenue for last few months	Yes
2	I got unexpected divorced after 2 years of relationship	Yes
3	1 pound is equivalent to 1.41 dollars	No
4	My company lost $1 million dollar revenue in last quarter	Yes
5	Spending $100 dollars for a lunch in restaurant is too bad	No
6	Our meeting will be at 3pm in the US Bank building	Yes
7	Yesterday to garbage keywords am nothing Houston more keywords	No
8	I got the Flu	Yes
9	My son nothing morning no sense makes spoofing not $100 dollars	No
10	We are planing to leave for Paris on 31st December in early morning	Yes
11	Houston is a very populated city to live in	No

cation, situation etc. Our focus is to analyze each piece of content, and evaluate whether or not an individual is disclosing private information through any of the sentences while expressing his pronouncements. Some examples of private disclosures and public information can be found in Table 4.2.

4.2.3 Data Labeling

As mentioned above, no ready-made labeled dataset is found for our experiment where various types of sentences are marked as discloser or non-discloser. Both the privacy policy of available data sources and complexity in classification of such textual data, might be the cause. Yet, this is the most important factor from the model's perspective which learns in a supervised fashion. So, our collected dataset is labeled using an algorithm that is built upon the idea of rule-based approach used by [148, 160], and obeying the following definitions.

Definition 2 *(**Disclosure Related Entity Type**) - Each $DRET_f$ is a set of DREs that belong to a type f, where $f \in F = \{$Person, Location, Money, Health, Date, Time, Interpersonal Relationship, Business Information$\}$. Having D as an infinite set of all possible DREs then*

$$\forall\, DRE_d \in D \nexists i, j \in F \text{ where } i \neq j, DRE_d \in DRET_i \cap DRET_j$$

By applying an entity and relation extraction tool [63], we implemented the following formal definition of disclosure to classify the dataset:

Definition 3 *(**Disclosure**) - Let sentence S be a set of words, $S = \{w_1, w_2, \cdots, w_n\}$. S is disclosing if it satisfies the following condition:*

$$\exists\, w_i, w_j \in S \text{ where } i \neq j, w_i \in DRET_{Person} \wedge w_j \in \bigcup_{f \in F} DRET_f$$

In order to label a sentence as disclosure (Definition 3), we examine the sentence. If it contains one or more entities (i.e., mention of a person, place, location, etc., explained below) and if one of those entities is of type *person*, then it's labeled as disclosure. This is a simple, yet effective rule which allows us to label our data set with disclosure/non-disclosure classes. A more structured guideline for manual labeling is given below:

1. Start with an example sentence
    (a) Look if that contains one or more DRE (by Definition 1) which falls into the set of DRET (by Definition 2).

(b) If Count of DRE > 1 AND at least one of the DREs is type of PERSON go to Step 2 otherwise label it as a Nondisclosure sentence.

2. Is it a grammatically valid sentence?

   (a) If YES go to Step 3 otherwise label it as a Nondisclosure sentence.

3. Label the sentence as Disclosure and return to step 1.

This produced 5000 disclosure sentences and 5000 non-disclosure sentences from the collected dataset (Table 4.1), that yields proper labeled information with ground truth. Human evaluation on the labeled examples (i.e., 20% of the data) was also done for the verification of the applied techniques. We use this data to train our model which we hypothesize will generalize to new data, that we show in our evaluation. Although, those 24,000 posts contained more than 100,000 sentences, we picked only those with disclosure related entities in it. Hence, the final quantity becomes lower after eliminating most of the sentences with non-disclosure content.

At this stage of our work, we consider the following entity types while discriminating sentences with privacy disclosure: `Person` *(e.g., I, He, Robert)*, `Location` *(e.g., Starbucks, Airport, Main Street)*, `Money` *(e.g., $100, 1 million)*, `Date` *(e.g., Tomorrow, 31st December)*, `Time` *(e.g., 7pm, Evening)*, `Interpersonal Relationships` *(e.g., Married, Divorced)*, `Health Information` *(e.g., Flu, Pregnant)*, and `Business Information` *(e.g., Revenue, Loss, Profit)*. It's worth mentioning that the types mentioned above are just a few from all possible categories that might be related to privacy and security. The number of considerable categories could be extended or reduced as per problem domain.

4.2.4 Data Pre-processing

As can be seen from the examples in Table 4.2, many DREs (e.g., I, divorce, 3pm, $100 dollar, Houston) can be used in both private disclosures and in public posts. This makes the problem particularly challenging because we cannot simply rely on the lexical items in the text; we have to consider the intent of the author of the text, and somehow determine if the intent was for the text to be public (i.e., DRE used in a public statement) or private (i.e., DRE used in personal context). To this end, we do special tokenization and enrich our data with additional information using linguistic details such as part-of-speech tags and syntactic dependency relations. We make use of the NLP toolkit Spacy [63] for all of our data pre-processing. This tool is also used for feature enrichment by creating synthetic features (e.g., dependency tree, POS tags) out of existing features (i.e., word tokens, sentences).

Tokenization In many text-based natural language processing tasks, the text is pre-processed by removing punctuation and stop words, leaving only the lexical items. However, we found that the way people punctuate their texts helps give the clues as to whether or not it is a valid private or public information. That is, we considered tokens from an example sentence like *Ok... I will meet you; tomorrow morning,, in-front of the Coffee Shop!... :)* are *[“Ok”, “I”, “will”, “meet”, “you”, “tomorrow”, “morning”, “,”, “in”, “front”, “of”, “the”, “Coffee”, “Shop”].* Therefore, we use the NLP Toolkit to tokenize the sentences in a customized way that ignores redundant tokens such as *“,,”, “;”, “!”, “:)”* but keeps the important ones. This step of considering all the valid sequential tokens helps our model learn important arrangement of tokens for validating relationships of entities. This is somewhat in contrast to other text analysis

literature where clearing off all the punctuation tends to improve task performance. However, keeping the punctuations showed better performance than removing them, throughout our experiment.

Syntactic Structure Present linguistic theory, classifies certain formal properties of language as "purely stylistic." That is, two sentences can have different forms but express the same meaning [132, 49]. For example, a sentence with the structure *subject verb direct-object preposition object* is semantically equivalent to *subject verb object direct-object*, though they are syntactically different. Also, as per our experiments, dependency parse information, and parts of speech tags are two synthetic features that improved the performance of the neural network model. This helps the model to observe the common sequence of tokens as well as co-occurrence of dependency tags. We use a Dependency Parser (DP) Toolkit [63] to extract the syntactic relation information (which is different from, but in some ways similar to, entity relation information). This allowed us to enrich our data with dependency parse information.

Parts-of-Speech Even though we use syntactic structure, we also include parts of speech as a slightly less structured representation of the input text that is also non-lexical. (We found, however, that including Parts-of-Speech did not dramatically increase the performance of our model.)

Figure 4.1 shows an example of the linguistic feature enrichment for the example sentence *Me and Steve will meet you tonight* for parts-of-speech (which appear below the words) and the dependency parse tree. Figure 4.2 shows the entities with their tagged entity types.

In summary, our data set is comprised of the original tokenized text enriched with

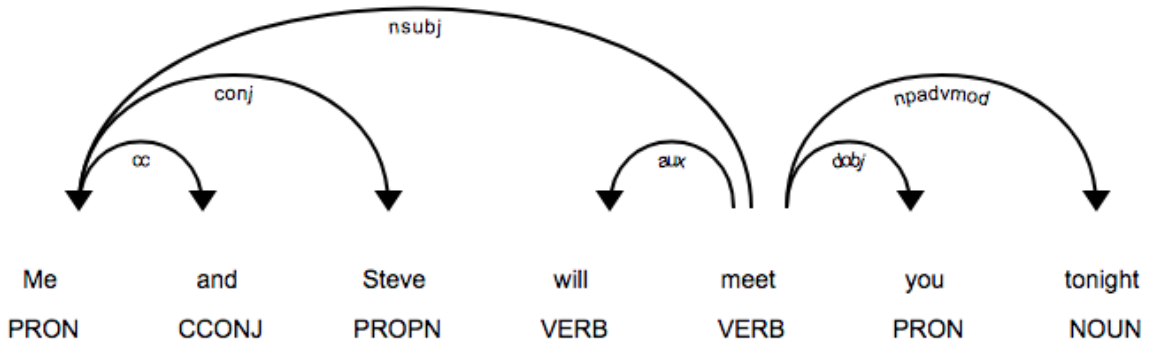


Figure 4.1: Parts-of-speech and dependency parse tree of an example sentence.

Figure 4.2: Recognized entities in an example sentence.

parts-of-speech, tagged entities, and syntactic information.

4.2.5 Model and Approach

Our model composes together multiple channels of a convolutional neural network to perform the disclosure/non-disclosure classification task, where each channel refers to different representations (i.e., word tokens, dependency parse tree, parts-of-speech tags) of the same candidate piece of text. All the channels use similar hyper-parameters (e.g., input/output dimension, activation function, dropout) applied to them to keep computational consistency. Shared input layers are combined together at the first stage of the neural network which is described in this section.

4.2.6 Neural Network Architecture

The primary task is a supervised optimization problem while minimizing the error of classifying disclosure/non-disclosure sentences. An overview of our proposed frame-

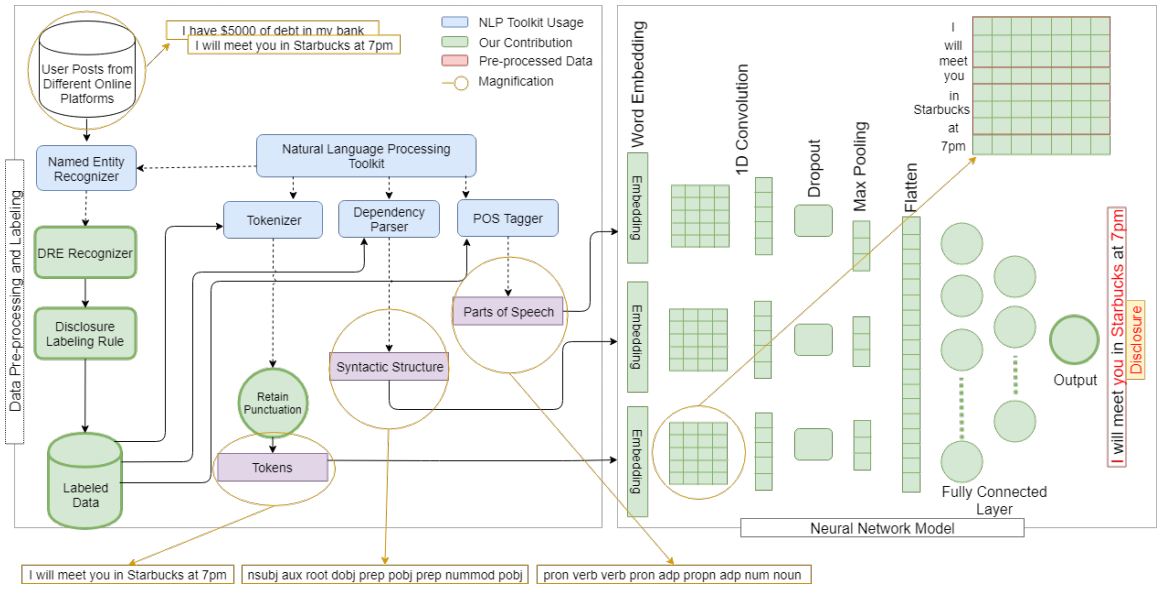


Figure 4.3: The bigger picture of the whole framework combining linguistics and neural network stages.

work, along with the core model, is represented in Figure 4.3. We explain most of the important constituents of the system below.

Word Embedding Layer Word embedding represents words as a dense vector representation in high-dimensional space [35] [154, 73]. Unlike the typical bag-of-words model, where words are represented as very sparse high-dimensional (e.g., 1-hot) vectors, in word embeddings, words are represented by dense vectors where a vector represents the projection of the word into a continuous vector space. The most important benefit of utilizing word embedding is that the position of a word or token within the vector space is learned from text and is based on the words that surround the word where it is used. This is useful because words that have similar semantic meanings are close (in terms of Euclidean distance) to each other; which is more semantically useful than one-hot encodings, in which all words are semantically

equidistant from each other.

In the proposed neural network architecture, we apply word-embedding as the first layer of the model to learn embeddings through training. Specifically, three separate embedding layers are used as the first hidden layer of each of the multichannel input of the network. We prefer this technique of learning embeddings because we did not observe better accuracy while using pre-trained word vectors like GloVe [112], rather it caused computational overhead. Glove, for example, contains 800 billion of tokens which in turn incorporate 800 billion of word vectors. On the other hand, these embedding layers learn semantic relationships from DREs, words, and tags from our data throughout the process. This is particularly crucial, as we apply embeddings not just to words, but also to three types of derived linguistic features: parts-of-speech, entities, and dependency parse, as explained in Section 4.2.1). We observed better performance while implementing this approach.

Convolution Layer CNN is a neural network architecture which is useful in mapping 'togetherness' of information (i.e., image of objects, sentence of tokens) onto class labels. These are feed-forward neural networks that became popular in image processing by work of LeCunn et al. [94]. While traditional CNNs used in image processing are 2D, 1D CNNs can be successfully used for sequence processing [93, 86]. This is because, text data (e.g., a sentence of tokens) have a strong 1D (sequential) locality that can be successfully extracted by convolution. LSTM neural network seems a good fit for this task in the first place, these networks are more computationally intensive than CNN-based networks. In this work, sequences of tokens in-between entities is observed deeply by utilizing one-dimensional convolution with smaller kernel for learning about valid syntactic structure among entities in a way where one or

more entities are modifying other entities.

Another challenge that makes the problem of validating sentence structure difficult is that the sequences (i.e., the input sentences and accompanying linguistic features) can vary in length. Sequences could be short as 2-3 words in length or, as long as 8-10 words. It's obvious that the model needs to learn the co-occurrence of tokens or dependencies between symbols in the input sequence. Unlike two-dimensional convolution in an image processing area which focuses on spatial visual structure, a one-dimensional convolution suits perfectly in this approach for looking into sentences. As with the word embedding layers, there is a convolution layer for each channel–one for each linguistic feature type.

Following each convolution layer, we introduce a dropout layer, a pooling layer and, a flatten layer before going into the concatenation layer where inputs from different channels are merged.

Concatenation In this layer, the three channels are brought together. Our final goal is a single, composed neural network that uses the three linguistic feature types then performs a single binary classification task. Concatenation is the simplest form of bringing these different channels together by simply representing the output layer of the respective CNNs from each channel as a single input into the following layer.

Fully Connected Layer After concatenation, we apply several densely connected network layers. These hidden layers are comprised of one hundred neurons in the input, then ten neurons in the hidden layer and, finally an output neuron for binary classification at the end. We implemented the well-known Rectified Linear Unit

(ReLU) neurons for the first two layers and Sigmoidal neuron in the output layer.[2] Our final resulting model is depicted in Figure 4.3 where three separate channels take in three different linguistic feature sequence types, each channel begins with an embedding layer, followed by a CNN layer; those three layers are concatenated, then a three-layer feed-forward network made up of dense layers (using standard ReLU and sigmoid activations) outputs a distribution over a binary class.

In summary, the model is not only learning about the private information but also learning about the correct grammatical structure of such sentences. We train it with words themselves as well as with two other representations (i.e., parts of speech and dependency tree) of the example sentences. This helps the machine learning model to learn both privacy-related tokens and the pattern of a correct sentence.

4.3 Experiment

This section and the subsequent portions contain details about the experimental environments and tools, along with the implementation of the proposed model in the processed data set and results from an off-line evaluation.

4.3.1 Data Preprocessing

In the data pre-processing step, we applied Spacy [63] to derive the linguistic features of each sentence. This tool comes with several features to analyze natural language text. Parts of speech tagging, deriving syntactic structure, and tokenization are done by this toolkit. The reasons behind selecting Spacy include - its trainable statistical model (we trained its existing NER model), dependency parser, tokenizer, noun chunk separator in a single toolkit. Two peer-reviewed papers in 2015 confirm that spaCy

[2]It is worth mentioning that we get little fluctuation in the accuracy value while changing the number of neurons in these layers. It seems obvious because this layer might have needed more neurons for better non-linearity understanding when it sees relatively more data.

offers the fastest syntactic parser in the world and that its accuracy is within 1% of the best available. It also contains a statistical entity recognition model in it, but does not have an entity recognizer for more specific types in which we are interested, such as Interpersonal Relationships, Health Information and, Business Information. The default model identifies a variety of named and numeric entities, including companies, locations, organizations and products, falling somewhat short of identifying some additional entities according to our problem scope.

For example, out of the box, it can not identify *flu* as a disclosure-related entity, whereas it should be identified as a Health Information type entity as a task of the first step toward the whole disclosure recognition system. We were, however, able to leverage Spacy's model extension provisions [63], resulting in an extended entity recognizer model that was trained to identify Interpersonal Relationships, Health Information and, Business Information such as *divorce*, *marriage*, *flu*, *cancer*, *fever*, *loss*, *profit*, etc. as valid recognizable entities. An annotator tool by Spacy called Prodigy [145] is used to train the NER model further for identifying these new types of entities. Prodigy has a loop model architecture by which it shows relevant keywords based on the annotation of previous steps.

After this, text encoding is done using Keras [78]. At the end of integer encoding, post padding with zeros is also done for all the sequences or sentences to a certain value which is the maximum length of a sentence in the whole training data set. The post padding is needed to make all the input sequences the same length, which is required by the later neural network architecture.

4.3.2 Neural Network Implementation

For implementing the word embeddings we use the *Embedding* layer of Keras [73] that turns positive integers into dense vectors of fixed size [73]. As per its requirement, the integer encoding of all text data is completed in the earlier stages. In the beginning, the embedding layers are initialized with random weights and then learn embeddings for all of the words in the training dataset.

For the *Convolution* layer, we use the Conv1D layer of Keras. To avoid the over-fitting problem of this neural network, we applied 20 percent dropout rate after each convolution layer using Dropout layer of Keras. This is a common practice which means setting the values of 20% input units to 0 at each update during each iteration of the training life cycle. A pooling layer is also added just after the dropout layer by utilizing *Pooling* followed by a *Flatten* layer of Keras.

The Keras functional API provides some methods to define complex model structure such as multi-input and or multi-output models that best suits our case. The `concatenate` method of Keras takes all the output vectors from the convolution layers and merges them into a single vector which then acts as the input to the later fully connected layers [74].

4.3.3 Model Hyper Parameters

This section describes all the needed model hyper-parameters and the intuition behind the selection of those parameters and associated values. First of all, random seeding is used for maintaining reproducibility while experimenting with different architectural values. For the `Input` layers that define the shape for each of the three multi-channel inputs, are determined by the length of the longest sentence (by tokens).

In each of the three embedding layers, all the mandatory parameters are chosen as

follows: input dimension is the vocabulary size and, output dimension that describes the size of output vectors where words are embedded is 100 and increased to 200 while working with more than twenty thousand sentences.

Convolution layers are comprised of `32` filters with kernel size of `4`, and `relu` as the activation function, keeping all other parameters to default values as determined by Keras. Some default parameters are worth mentioning such as, `valid` (no padding) as padding type, `1` as the strides and dilation rate, `zeros` as bias initializers, with no kernel regularization (regularizers allow to apply penalties on layer parameters during optimization).

Pooling layers are responsible for the max pooling operations on the temporal data which are comprised of `2` as pooling window and, strides for downscaling. This layer uses `valid` as the padding type by default. To prepare the data for concatenation, we flatten all the multi-channel inputs separately after the max pooling.

We use ReLU (Rectifier Linear Unit) as the activation function for all the neurons in the dense hidden layers, whereas Sigmoid is used as the activation function in the only neuron of the output layer where we get a probability value towards disclosure or, non-disclosure. The model is trained using 50 epochs, with a batch size of 100.

4.3.4 Model Summary

A high-level summary of the multi-channel convolutional neural network goes as follows - each embedding layer produces 100-dimensional word embeddings, and connected to the earlier input layers. Also, each of the convolution layers contains 32 filters with no padding. After the convolution, dropout layers and pooling layers are employed. Later, three separate flatten layers are used. Eventually, a concatenation layer merges all the input vectors into a single one, and forwards to the fully con-

nected layers. Finally, the output layer that contains a single neuron produces the probability score for the desired binary classification.

4.3.5 Task and Procedure

Our task is a binary classification task of identifying whether a piece of short text contains a personal disclosure or not. We compare our model (as described above) to several other known classification models after the data pre-processing step (i.e., all models had the same inputs). The procedure of applying those models and their outcomes are described below.

Simple Convolutional Neural Network A simple CNN with only word tokenization is first applied for identifying disclosure and non-disclosure events. This simple network also uses a word embedding layer along with 32 filters with kernel size of 3 by maintaining same padding for convolution, max pooling of 2, using binary cross-entropy as loss function and, ReLU as the activation function. This network serves as our baseline.

LSTM Recurrent Neural Network We also compare to a recurrent neural network, LSTM, because LSTMs have been shown to produce good results in sequential language processing tasks. We use a word embedding, LSTM (with 100 neurons), and dropout (20%) layer.

CNN with LSTM Network We also compare to a combination of the CNN and LSTM models as they are explained above. This allows the model to combine the benefits of the sequential LSTM and filters from the CNN in a single model. The data of this experiment contains one-dimensional spatial structure in the sequence of

words in conversational text and the CNN (Convolutional Neural Network) tries to pick out invariant features for disclosure and non-disclosure events. These learned spatial features are then treated as sequences by the subsequent LSTM layer. This combined neural network shows a very good improvement in accuracy but going through an obvious computational overhead.

The Multichannel CNN Eventually, our proposed multichannel convolutional neural network is applied for the classification of disclosure and non-disclosure sentences by providing word tokens in one channel, dependency parse tree to another channel, and parts of speech tags to the third channel. This is the final model we integrate into the proposed framework (after the data simplification stage) because of its ideal performance. It's worth mentioning that, a multichannel LSTM recurrent neural network was also applied for the classification of the data set, just like the final multichannel CNN. This network also gets different data representations into different channels but could not beat the final model. Even though, LSTM based network seems best suit for learning patterns from sequential data, our convolutional network makes best use of learning togetherness of tokens on the pre-processed data and outperformed all of our other experimental models.

4.3.6 Metrics

Classification accuracy (Equation 4.1), F-Measure, and Receiver Operating Characteristic (ROC) are used as the evaluation metrics. We consider these different types of evaluation metrics because we take it as a binary classification task where accuracy, precision, recall, and diagnostic ability of disclosure identification are equally important. We use labeled data to train our model in a supervised fashion, and evaluation

is also based on similarly labeled data-set (actually a split from the original data set by 30%). The remaining 70% of data was used as the training and validation set, containing 50% and 20% in each group, respectively.

$$Accuracy(ACC) = \frac{\sum Truepositive + \sum Truenegative}{\sum Totalpopulation} \tag{4.1}$$

For observing the precision and recall of our final model, we consider F-Score as per the following equation (Equation 4.2). We try to look at how precise our model is while identifying disclosure sentences as well as its capability of pulling out disclosure sentences as much as possible from the test data set.

$$F_1 = \frac{2}{\frac{(TP+FN)}{TP} + \frac{TP+FP}{TP}} = \frac{2TP}{2TP + FP + FN} \tag{4.2}$$

A ROC curve is used to evaluate the association of true positive rate against the false positive rate to examine the sensitivity, and fall-out of the model. We also calculate the AUC (area under curve) value of the ROC curve.

4.3.7 Results

For experimenting with different models to achieve a strong classification result, the model variants described above with different architectures are applied in the same data set. Each variant gets the same simplified and entity-marked data.

The simple convolutional neural network that uses only word tokenization shows 69.2% accuracy in identifying disclosure and non-disclosure occurrence. Simple LSTM network shows 70.6%, and the combined neural network of convolution and LSTM layers shows 74.1% of accuracy. The multi-channel LSTM neural network model achieved 81% accuracy.

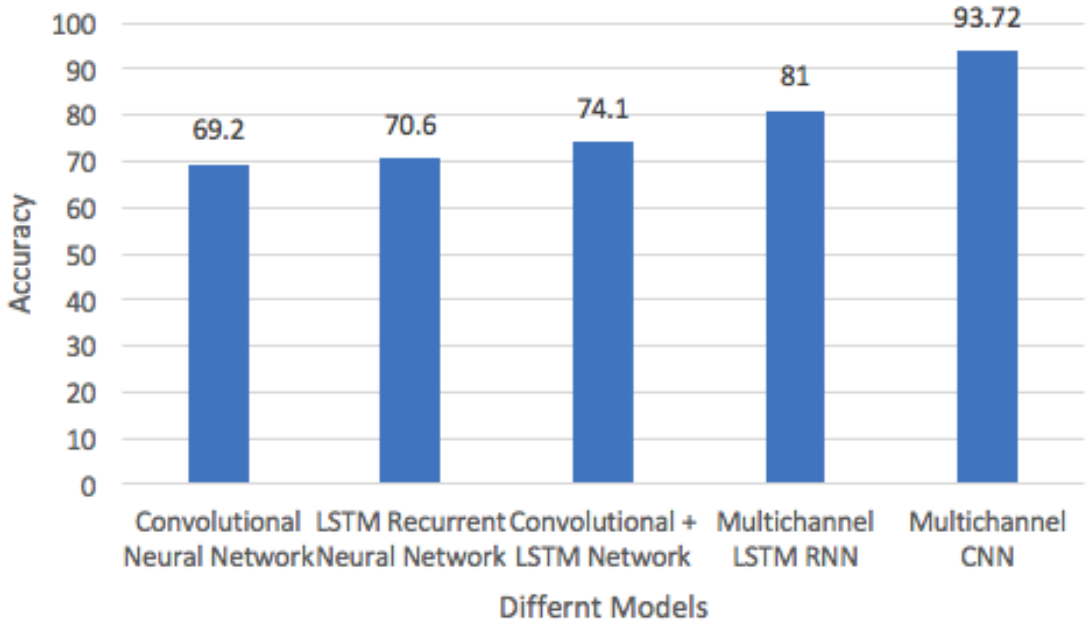


Figure 4.4: Comparison among different models

Our proposed model that uses multi-channel inputs and convolution layers along with word embeddings shows 93.72% accuracy on the data set of labeled disclosure and non-disclosure sentences. Also, it shows significant learning improvement on the amount of training data set. Figure 4.4 shows the comparison of accuracy among all the experimented models along with the final proposed one. Accuracy is measured on the test data that is basically a split of the whole data set and unseen to the model while training.

The model shows 0.94 F-Score on disclosure label and, 0.93 on non-disclosure with an overall weighted F-Score of 0.93. Figure 4.5 shows the ROC curve that is generated as per the predicted labels and, true labels of the test data set. We find a significantly large area under the curve that is 0.98 that clearly indicates the strength of the classification model. The ROC curve tells us where we can reliably set the

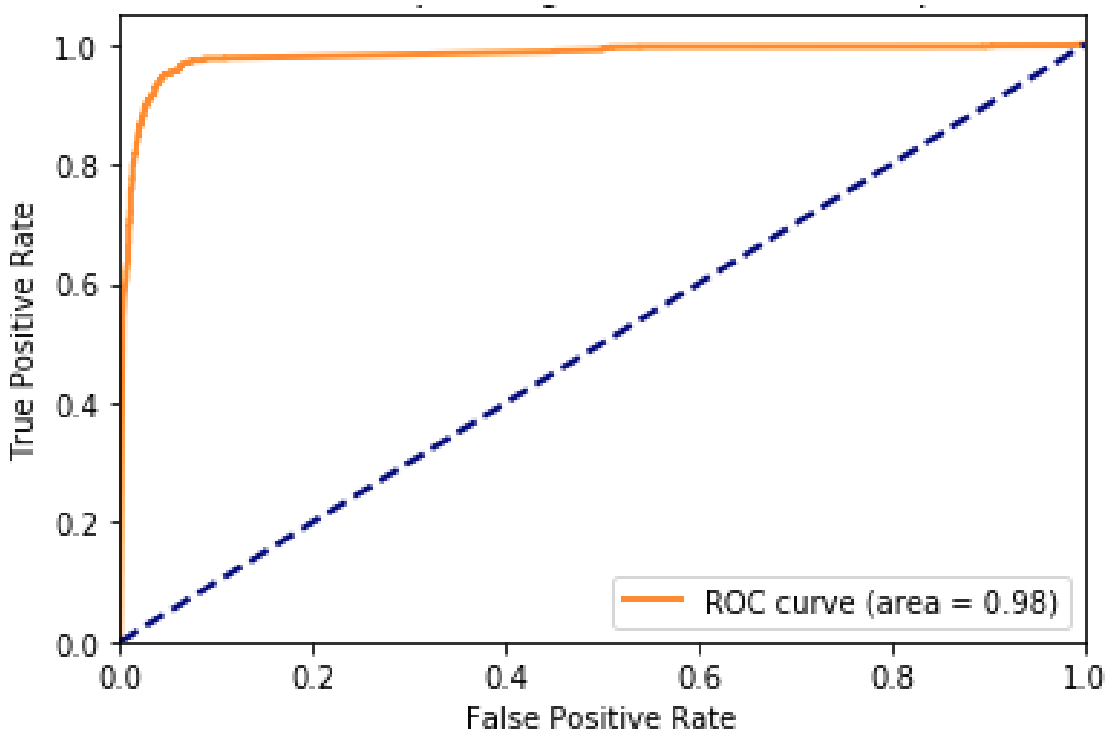


Figure 4.5: Receiver operating characteristic curve

model to disallow false negatives. It's important to know because in this particular task, the system should notify users about information disclosure in a lower threshold (i.e., positive if prediction beyond 0.40) to be strict in information leakage.

These are overall positive results. They show that, despite a lack of large amounts of labeled data, we can train a classifier that goes beyond simple keyword spotting and uses linguistic features to determine if a text contains a disclosure or not with a useful degree of accuracy.

Table 4.3 shows how we get different accuracy scores on the same data set based on the effect of different input channels.

These results, however, are only applied to learning the automatically labeled data. We further evaluate on 200 manually labeled data (i.e., English sentences which may

Table 4.3: Impact of using multichannel data.

Channel	Accuracy %
Single Channel with Word Tokens	70.6
+ Dependency Parse Tree Information as Second Channel	87.4
+ Parts of Speech Tags as Third Channel	89.0
Multi-channel Input	93.7

or may not have the same characteristics required for our labeling rule, as described above) yielded 86.4% accuracy in disclosure identification. This dataset of ground truth was labeled by the human who had no idea about the working principle of this model. Those were evaluated from the natural perspective of the human agents. This experiment simulates one of the many possible case studies of the developed disclosure identification system.

In order for the proposed framework to be integrated into a global solution for the end users' privacy management problem, a web browser extension is developed to detect privacy disclosures as users are typing their text messages. The implementation is based on a server-based request-response architecture. The client (i.e., Browser Extension) captures user-side text and sends it to the server for classification where the trained model is already deployed. If any sentence contains privacy disclosure then the color of that text changes to red, as depicted in Figure 4.6. On the other hand, as represented in Figure 4.7, the color of the text does not change, since no disclosure is detected.

This implementation of the proposed framework is one of the many possible use cases. It is also important to note that we recognize the limitation of the developed tool, since sending personal data to a remote server for a classification purpose might result in users' privacy violation. For the future version of this tool, we will implement

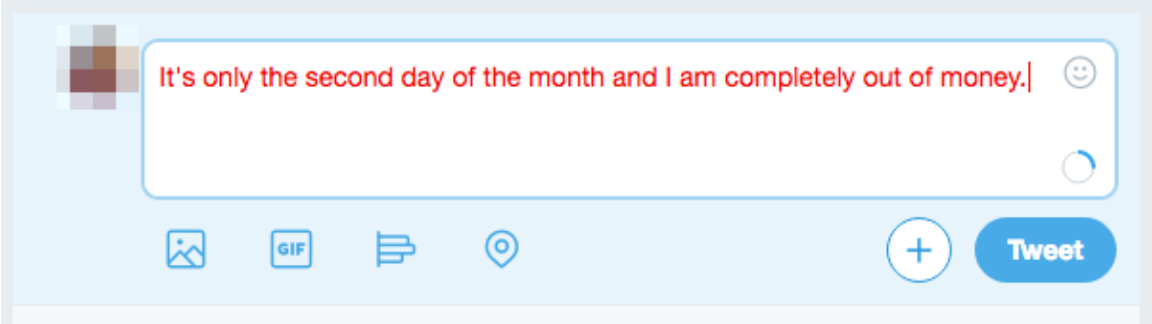


Figure 4.6: Information disclosure marked as red automatically by the browser extension.

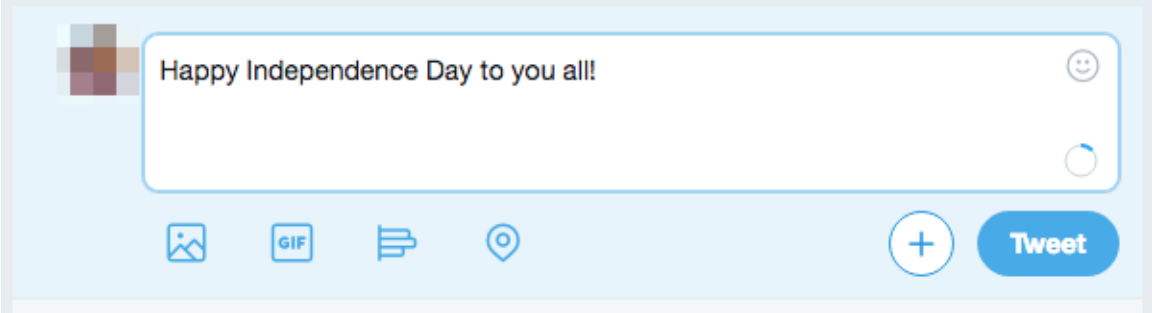


Figure 4.7: Non-private information keeps default color.

an architecture based on a pre-trained model stored in the client side (e.g., Using TensorflowJS). Source code of this implementation (i.e., the web browser extension and the API server) along with other resources regarding this work are made available for interested researchers [3].

4.4 Discussion and Analysis

For a baseline evaluation and assessment of the generalizability of the proposed framework, a dataset that was created by Schrading et al. and Choudhury et al. [136, 42] is utilized to detect privacy disclosures in Reddit users' posts and comments. This dataset was created mainly to analyze and study the dynamics of domestic abuse in electronic social media (i.e., Reddit). This dataset is comprised of posts and comments from Reddit users under several sub-reddits such as *abuseinterrupted, domesticviolence, survivorsofabuse, casualconversation, advice, anxiety, anger, relationships, and relationship_advice.* All the posts and comments are labeled with one of the above classes.

For the purpose of creating a comparable result, we divided the posts into two classes of Disclosure or Non-disclosure. Submissions under the sub-reddits - *abuseinterrupted, domesticviolence, suervivorsofabuse, and relationship* are considered as *Disclosure* class and *casualconversation, and advice* as *Non-disclosure* (Table 4.4). With this new binary classification, the proposed framework was able to detect each post or comment as a disclosure or non-disclosure with an accuracy of 95%.

Further analysis revealed that even if a post is labeled as an *Abuse*, not all the sentences in the post represent the labeled class and that is the limitation of the work by [136, 42]. However, the framework proposed in this paper is able to classify text at

[3]https://anonymous.4open.science/repository/3c84ab7b-02ce-4fd7-b982-f278d6f3c4f4/

a sentence-level and provide a more detailed analysis. Therefore, in order to be able to compare the result of our classifier with the work of [136, 42], we implemented a rule that if at least 70% (i.e., 7 out of 10 sentences of a submission) of the sentences of a post are classified as disclosure, then that entire post is classified as a disclosure. The result of the classifier was assumed correct if that same post was classified as abuse by [136, 42].

From all our experiments and the evaluation explained in this section, we have been able to recognize that with considering each sentence as the unit piece of information, the proposed framework faces some limitations while working on conversational context. At the moment, discourse information that spans beyond sentences cannot be handled in the way the system works. For example, a chat conversation like - *:How is your son? ...*, *:Bad ...*, *Got flu ...* can mislead the whole system for identifying both the disclosure and the actual nominal subject of this context. Whereas the understandable and rephrased version of the sentence is actually *My son got the flu* and is certainly a disclosure. One possible workaround is to implement the exact same procedures with an extended lookup window. For example, an information extraction step can be implemented in a sliding window style where each window will contain more than one phrase or utterance. Thus, it might be able to find the semantics, and the dependency parse tree of the conversation.

Another limitation of this proposed system is related to incorrectly (i.e., grammatically) written sentences. People often do not care about sentence structure while texting (which is more like speech than standard text) with close friends, and family members. On the other hand, this system moderately depends on sentence structure, specifically structure in-between entities.

Table 4.4: Summary of the Reddit dataset.

Sub-reddit	Class	Quantity	Target Class
abuseinterrupted	Abuse	1653	Disclosure
domesticviolence	Abuse	749	Disclosure
suervivorsofabuse	Abuse	512	Disclosure
relationship	Relationship	8201	Disclosure
Total	**-**	11,115	-
casualconversation	Not-abuse	7286	Nondisclosure
advice	Not-abuse	5913	Nondisclosure
Total	**-**	13,199	-

4.5 Conclusion and Future work

A practical model of privacy protection is in dire need by users in the era of social networks that results in activities such as posting online, chatting, text messaging, blogging, and playing online games, etc. Therefore, the development of algorithm and tools that helps users to identifying privacy disclosure in textual data is important. While many research studies in this area mainly focus on classifying textual data as public or private at the document or paragraph level, only a few of those are concerned with the privacy detection at the sentence-level analysis. Hence, these approaches can not be used for managing privacy for users, and they are mostly designed for the privacy protection of organizations and corporations.

To address this limitation, this paper proposes a privacy disclosure identification framework, comprised of a neural network model with linguistics. The proposed framework is capable of: I) detecting disclosure related entities more effectively by utilizing natural language processing techniques rather than relying on random keywords from an unbound set of tokens, II) conducting disclosure detection analysis

only on sentences with correct subject-verb agreement to increase performance time.

For the proof of concept, we conducted several experiments, examining various machine learning based algorithms (Figure 4.4) with the different types of data pre-processing techniques, and parameter tunning approaches, while experimenting with various neural network architectures. Throughout this process it was proven that the entity-based evaluation, and enriching the input data with additional underlying features helped improving the performance of the model. Convolution over the feature vectors resulted in learning about the sentence structure as well as to overcome the computational overhead.

The future work will concentrate on extending the number of Disclosure Related Entity Types (DRET) to improve the disclosure detection process. Further, the proposed framework will be made more intelligent to be able to infer from the text analysis the interpersonal relationship (i.e., relationship among friends, family members, colleagues, and public), the context in which the disclosure occurs, and the timing of disclosure to provide an effective privacy management tools an algorithms for users. In order to achieve this objective, inter-annotator agreement measures and annotation guidelines will be used to ensure consistent annotations while developing a generalized dataset that will include human annotation through crowdsourcing.

CHAPTER 5:

A USER-CENTRIC AND SENTIMENT AWARE PRIVACY-DISCLOSURE DETECTION FRAMEWORK BASED ON MULTI-INPUT NEURAL NETWORK[1]

Data and information privacy is a major concern of today's world. More specifically, users' digital privacy has become one of the most important issues to deal with, as advancements are being made in information sharing technology. An increasing number of users are sharing information through text messages, emails, and social media without proper awareness of privacy threats and their consequences. One approach to prevent the disclosure of private information is to identify them in a conversation and warn the dispatcher before the conveyance happens between the sender and the receiver. Another way of preventing information (sensitive) loss might be to analyze and sanitize a batch of offline documents when the data is already accumulated somewhere. However, automating the process of identifying user-centric privacy disclosure in textual data is challenging. This is because the natural language has an

[1]Nuhil Mehdy, H. Mehrpouyan, "A User-Centric and Sentiment Aware Privacy-Disclosure Detection Framework based on Multi-input Neural Network," 2020 *PrivateNLP @13th ACM International WSDM Conference*

extremely rich form and structure with different levels of ambiguities. Therefore, we inquire after a potential framework that could bring this challenge within reach by precisely recognizing users' privacy disclosures in a piece of text by taking into account - the authorship and sentiment (tone) of the content alongside the linguistic features and techniques. The proposed framework is considered as the supporting plugin to help text classification systems more accurately identify text that might disclose the author's personal or private information.

5.1 Introduction

Privacy is an ancient concept concerning human values that could be "intruded upon", "invaded", "violated", "breached", "lost", and "diminished"[151]. Each of these analogies reflects a conception of privacy that can be found in one or more standard models or theories of privacy. Users' privacy has been defined as "the right to be left alone" or being free from intrusion by the seclusion and non-intrusion theory[163, 55]. Even though privacy varies from individual to individual and each user may have different views of privacy, there is an imperfect societal consensus that certain information (e.g., personal information, situation, condition, circumstance, etc.) is more private than the others (e.g., public statements, opinion, comments, etc.)[26].

Recent advances in communication technologies such as messaging applications and social media [151] have resulted in privacy concerns [127] about analogous information amongst the users. In this era of digital communication, an increasing number of users are sharing information through text messages, emails, and social media without proper awareness of privacy threats and their consequences. Moreover, in the context of the information society, historical documents of entities (e.g., peo-

ple, organization) are needed to be made public and shared among authorities every day [135]. In such cases, improper disclosure [2] of a user's information could increase his/her security/privacy vulnerabilities, and the negative consequences of disclosing such information could be immense [32].

A recent data scandal involving Facebook and Cambridge Analytica reveals how personally identifiable information of up to 87 million Facebook users influenced voter's opinions [143, 61]. Likewise, millions of data breach incidents are reported all over the world, and unfortunately, most of them expose users' personal data [146]. Therefore, user-centric targeted attacks by exploiting the victim's Personally Identifiable Information (PII) has become a new kind of privacy threat in the present-day [162]. It's worth mentioning that the United States is the number one destination for such user-centric targeted attacks based on recent statistics [150]. That being the case, users' data privacy has become one of the major concerns of today's world, and the requirements for privacy measures to protect sensitive information about individuals have been researched extensively [70, 71, 24, 103, 139, 69].

As part of this efforts, researchers in the area of Natural Language Processing (NLP) have focused on developing techniques and methodologies to detect, classify, and sanitize private information in textual data. However, most of these works tend to solve these tasks by just detecting set of keywords, leveraging dictionaries of terms, or applying regular expression patterns. These types of detection do not consider the context and the relationship of the keywords in the text, therefore they result in a high amount of false positive (e.g., a doctor's article about a disease is considered public and not private). However, it is considered sensitive and private when associated with

[2] In this work, disclosure is defined as revealing personally identifiable information (e.g., name, address, age) or sensitive data (e.g., health, finance, and mental status) to others.

other entities (e.g., a patient himself) in certain ways that yield different meaning and actually reveals someone's privacy. Therefore, its equally necessary to look into the keywords, data subject (i.e., users), authorship, tone, and overall meaning of the content before classifying it as privacy disclosure (Refer to Figure 5.1). While a few of the recent works are concerned with disclosure detection techniques by considering user-centric factors, most of them still omit other important decision-making factors such as sentiment and authorship of the content. Therefore, this paper aims to review the existing methodologies and techniques from the area of NLP and proposes a novel disclosure identification framework by keeping the following factors in mind:

- **Considering users-centric circumstances, tone, and authorship of content:** content having - sensitive information but no data subject, sensitive keywords but public ambience, analytical tone should not be classified as disclosure.
- **Checking sentence coherence and grammatical structure:** appearance of random keywords, ambiguous and meaningless information, or invalid utterances should not be classified as disclosure.

The rest of the paper is organized as follows: Section 5.2 describes the dataset used in this paper. The methodology is described in detail in section 5.3. In addition, the detail of the deep neural network architecture, data cleaning, pre-processing, featurization, and the experiment is presented in section 5.4. Lastly, section 5.5 represents the experimental results following the conclusion.

5.1.1 Our Contribution

The limitations of the current studies are based on the fact that they solely rely on the existence of keywords and neglect the sentence coherence, ignore grammatical

Figure 5.1: Example of disclosure post, non-disclosure post, and highly similar to disclosure but actually a non-disclosure post (from top to bottom respectively).

validation, and disregard meaning inference in a piece of content. It has been addressed that these limitations, in some cases, result in missclassification and could be resolved by integrating parts-of-speech tags, dependency parse tree information, and word embedding [109]. However, a novel approach is required to take into account the emotional tone or sentiment of the users that are hidden in the textual contents. For example, in Figure 5.1, the text from the red box is revealing someone's private (health) information (the patient has cancer), and the text from the green box is about the Idaho state that represents some public ambiances. It's quite easy to distinguish these two pieces of texts based on the keyword spotting techniques [17]. However, in another example, the text from the yellow box (comment from a doctor about cancer) has similar keywords as the patient's post, in the red box, containing valid word sequences and the presence of grammatical subjects (i.e., first person) with references etc. This piece of text is definitely not revealing private health situation (i.e., the doctor himself does not have cancer). Hence, it is quite challenging to distinguish between the types of contents without taking into consideration the sentiment of the statements. To this end, this paper focuses on distinguishing highly similar contents based on the users' involvement, sentiment, authorship, and grammatical structure to classify texts containing someone's privacy disclosure. However, one of the assumptions of this work is: the proposed model does not solve all the privacy and security requirements of users by providing an entire threat model, rather it provides a better NLP tool to be integrated into any comprehensive privacy framework.

5.2 Dataset

We collected 10,000 users' (patients and doctors) posts from a public online health forum, based on the observation (inspired from the example of figure 5.1) that, patients'

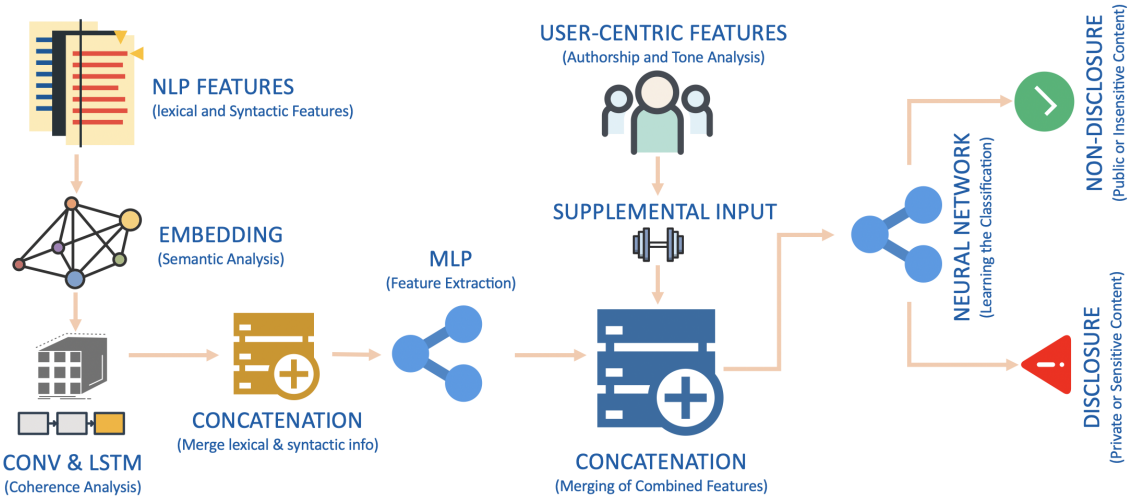


Figure 5.2: Bigger picture of the disclosure detection framework.

posts are somewhat disclosing their health status in that forum. Whereas, doctors' comments on patients' posts are highly similar content (having similar keywords and syntactic representation) but usually do not disclose doctors' health status (doctors' do not have those diseases). Therefore, we labeled patients' posts as disclosure (private) and doctor's comments as non-disclosure (public). For this paper, we crawled 5000 posts and 5000 comments and narrow down our privacy domain to health only. The length of the posts and comments varies from 10 words to more than 100 words comprised of several sentences.

5.3 Methodology

Combination of both linguistic operations and artificial neural network is the core of our methodology. A bigger picture of the framework is depicted in Figure 5.2. In this section, the data pre-processing, representation, and featurization steps are briefly explained, following the detail of the neural network architecture.

5.3.1 Featurization and Data Representation

As can be seen from the examples in Figure 5.1 many domain-specific keywords can be used in both private and public posts. This makes the problem particularly challenging because we cannot simply rely on the lexical items in the text; we have to consider the intent of the author of the text, and somehow determine if the intent was for the text to be public or private. To this end, we do custom tokenization and enrich our data with additional information using linguistic details such as syntactic dependency relations.

Tokenization In many text-based natural language processing tasks, the text is pre-processed by removing punctuation and stop words, leaving only the lexical items. However, we found that the way people punctuate their texts helps give clues as to whether or not it is valid private or public information. Therefore, we use NLP Toolkit to tokenize the sentences in a customized way that ignores redundant tokens such as *",,", ";–", "!!!", ":-)"* but keeps the important ones such as *",", ";", ":", ".", "he", "the", "in"* etc. This step of considering all the valid sequential tokens helps our model learn important arrangement of tokens for validating relationships of entities. This is somewhat in contrast to other text analysis literature where clearing off all the punctuation tends to improve task performance.

Syntactic Structure In the experiments, dependency-parse-tree information is also utilized as additional underlying features that improved the performance of the neural network model. This helps the model to observe the common sequence of tokens as well as co-occurrence of dependency tags. We use a Dependency Parser (DP) Toolkit to extract the syntactic relation information (which is different from, but in

some ways similar to, entity relation information). This allowed us to enrich our data with dependency parse information.

Supplemental Features In addition to the features mentioned above, more user-specific features or metadata are prepared and provided to the extended variant of our models as supplemental input. Some of those auxiliary data are - i) number of pronouns ii) emotional tone iii) number of negations found in the post etc. This additional information are supposed to give the neural network model some distinguishable features about highly similar contents of different class.

5.3.2 Deep Neural Network Model

After doing all the necessary pre-processing steps, the data is then fed into a multi-input deep neural network to learn the hidden patterns and features to distinguish between texts having disclosure and non-disclosure occurrences. It takes lexical (word tokens) features through one input, syntactical features (dependency parse tree information) through another input following a merging of those feature vectors. Later these vectors additionally get merged with supplemental (auxiliary) inputs before going through a further multi-layer perceptron stage. At the end of the deep neural network, a single neuron is used to provide the probability toward each of the above-mentioned classes. More detail about the architecture is depicted in appendix B.2.

5.4 Experiment

In the data pre-processing step, we apply Spacy [144] to perform the linguistic operations on the text. The Keras functional API is utilized to create the multi-input architecture [74]. For implementing word embeddings, we use its *Embedding* [73]

layer where pre-trained word embedding (glove) is used with the trainable flag set to true. In another input of the multi-input model, the same type of embedding layer but without pre-trained vector, is used to learn the embedding space from the dependency parse tree information. For the *Convolution* on the information of the first input channel, we use the Conv1D layer [77] following a pooling layer just after it.

In the other input of the model, a long short term memory (LSTM) layer is used over the dependency parse tree information. The `concatenate` method of Keras then takes the output vectors from the convolution layer and the LSTM layer and merges them into a single vector which then acts as the input to the fully connected layers. At this step, supplemental input, prepared by utilizing IBM Watson Tone Analyzer [66] are added with the concatenated vector following another stage of dense layers. Finally, a single neuron with sigmoid activation function outputs the probability of each class with 0.5 as the cutoff value. As false negatives of the classifier may bring dangerous consequences, it would be wise to lower this probability cutoff value towards the negative class, depending on the usage of the model. The detail of the hyperparameters is listed in appendix B.1.

5.5 Results

Prior to the experiment with the multi-input model, the classification task was examined using baseline models such as Naive Bayes classifier and simple convolutional neural network. Figure 5.3 shows in detail the comparison of accuracy among all the models along with the model which uses user-specific supplemental input. The results show that, despite a lack of large amounts of labeled data, neural network based classifier can be trained that goes beyond simple keyword spotting and uses

linguistic features to determine if a text contains a disclosure or not with a useful degree of accuracy. Moreover, it is observed that, integration of user-specific meta-data to the models increases the classification accuracy, significantly (up to 97%). However, the generalizability of the model has not been well evaluated because of the lack of data set with similar characteristics (i.e., indistinguishable utterances yet carrying different meaning).

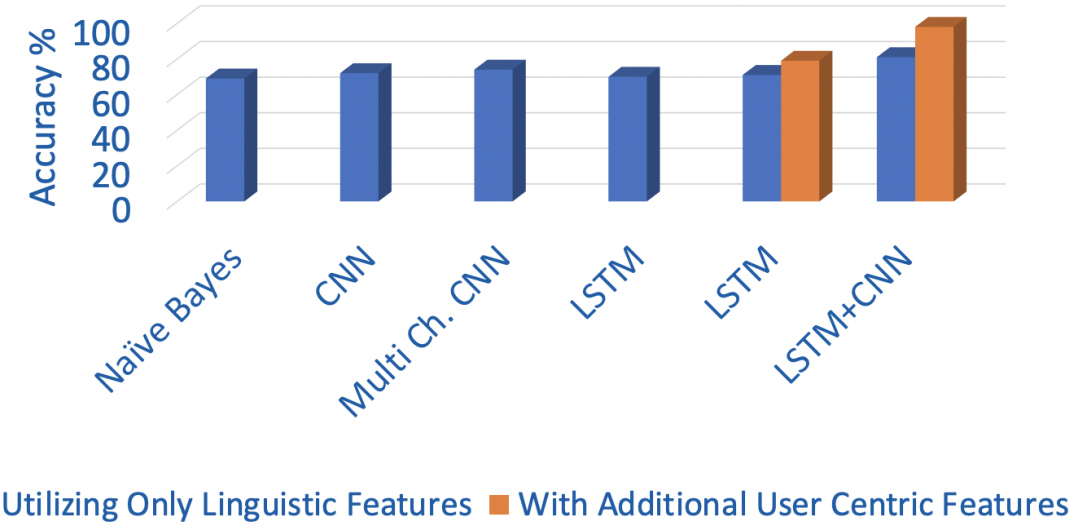


Figure 5.3: Accuracy of the model as a binary classification.

5.6 Conclusion

A practical model of privacy disclosure detection is in dire need by users in this era of social networks that result in activities such as online forum posting, emailing, text messaging, etc. Accordingly, the development of algorithm and tools that helps identifying privacy disclosure in textual data is important. While many of these works in this area mainly focus on classifying textual data as public or private at the document level by just spotting keywords, only a few of those are concerned with the privacy detection, taking the users' context into account.

CHAPTER 6:

A MULTI-INPUT MULTI-OUTPUT TRANSFORMER-BASED HYBRID NEURAL NETWORK FOR MULTI-CLASS PRIVACY DISCLOSURE DETECTION[1]

The concern regarding users' data privacy has risen to its highest level due to the massive increase in communication platforms, social networking sites, and greater users' participation in online public discourse. An increasing number of people exchange private information via emails, text messages, and social media without being aware of the risks and implications. Since a significant amount of data is shared in textual form, researchers from the area of Natural Language Processing (NLP) have focused on developing tools and techniques to detect, classify, and sanitize private information in text data. However, most of the detection methods solely rely on the existence of pre-identified keywords in the text and disregard the inference of underlying meaning of the utterance in a specific context. Hence, in some situations, these tools and algorithms fail to detect disclosure or the produced results are missclassified. In this

[1]Nuhil Mehdy, H. Mehrpouyan, "A Multi-input Multi-output Transformer-based Hybrid Neural Network for Multi-class Privacy Disclosure Detection," 2021 *2nd International Conference on Machine Learning Techniques and NLP (MLNLP 2021).* (Pending Notification)

paper, we propose a multi-input, multi-output hybrid neural network which utilizes transfer-learning, linguistics, and metadata to learn the hidden patterns. Our goal is to better classify disclosure/non-disclosure content in terms of the context of the situation. We trained and evaluated our model on a human-annotated ground truth data-set, containing a total of 5,400 tweets. The results show that the proposed model was able to identify both the information type (health, finance, relationship) of the tweets and disclosed contents with a valuable degree of accuracy (information type 99%, disclosure 76%) by jointly learning for two separate tasks.

6.1 Introduction

Over the years, with the increase in accessibility of the internet and growth of communication platforms and social networking sites, user's concern about their privacy has also increased [115, 80]. In order to provide usable tools and algorithms for users to manage the disclosure of their private information, many research has been carried out [117]. Mostly focused on understanding how users are sharing their private information through emails, text messages, and social media platforms and providing them with a clear picture of privacy threats and consequences of information sharing activities [108, 37].

Research in this area is especially important since the aggregated amount of personal information that individual shares could be exploited by the modern AI (artificial intelligence) techniques to gain meaningful insights on their private information which could lead to serious privacy violations [64]. Wang et al. argue that user-specific targeted attacks are becoming more common by exploiting the victim's private information [162]. Hence, the need to design and develop efficient tools and techniques to protect individual's privacy have resulted in researchers focusing on understanding

the individual's motive to disclose private information. [69, 103].

Since a significant amount of information is shared in textual form, researchers from the area of natural language processing (NLP) have focused on developing automated tools to detect, classify, and sanitize private information in text data [134, 27, 2]. A usable privacy-disclosure detection tool is dependant on the understanding of what constitutes as private information and what defines a disclosure for an individual user. Different information is considered as private or sensitive across different domains of human lifestyle [29]. Researchers have also intended to classify someone's private information into two main categories: objective (i.e., factual information such as age, sex, marital status, health condition, financial situation) and subjective (i.e., internal states of an individual such as interests, opinions, feelings)[156]. As per the scope of this paper, we define *privacy disclosure* as an occurrence when a piece of text, which is usually a statement/expression from an author, contains someone's private information/situation. In other words, we focus mostly on the objective disclosure where users explicitly reveal someone's privacy. We consider three types of information disclosure in this research work: health condition, financial situation, or relationship issues. For example, a disclosure occurs when a user tweets about his/her economic situation, i.e., the financial crisis he/she is going through, investment details, etc. Another example of disclosure could be when a patient tweets about his/her own physical/mental health condition, diagnosis results, medication/drug he/she is taking, etc. The intuition is similar for the Tweets that are about relationship issues. Likewise, we define non-disclosure as an event when a piece of text is not disclosing someone's health condition, financial situation, or relationship issues. Examples of non-disclosure information sharing activities are: when an activist tweets about the

national/global financial crisis, observations about the stock market, tips and tricks for the new investors, etc. Another example of non-disclosure could be when a doctor tweets about a disease, its symptoms, health care advice, etc. Therefore, a usable privacy disclosure tool is required to differentiate between public/private information and overcome the difficulties associated with the natural language processing of context-based textual data.

As part of this efforts, a wide range of proposed methodologies such as dictionary utilization, information theory, statistical model, machine learning, and deep learning have shown promising results in identifying privacy disclosure in text data [31, 59, 27]. However, most of the methods are based on the fact that they solely rely on the existence of keywords/terms/phrases and disregard meaning inference from the text. We observed through our experimentation that these limitations, in some cases, result in missclassification. This is because the existence of sensitive keywords in a piece of text does not always result in a user's privacy disclosure (please refer to examples 4,5,6 in Table 6.1). Hence, we propose a novel and hybrid multi-input multi-output neural network based model that overcomes the NLP challenges by precisely identifying privacy disclosures through tweets by combining knowledge from pre-trained language model, semantic analysis, linguistics, and the use of metadata. The multi-input, multi-output model is able to identify both the information type (health, finance, relationship) of the tweets and the disclosure occurrence by jointly learning for two separate tasks. We also trained and evaluated our model on a human-annotated ground truth dataset that contains a total of 5,400 tweets from anonymous users. Thus, our model could be implemented in the practical and usable fields of data privacy, information security, natural language processing, etc. A few of the notable

contributions of this paper includes:

- Presenting a multi-input, multi-output hybrid neural network that utilizes pre-trained language model, and still make use of traditional linguistics and structured metadata.
- Evaluating its multi-output capability that jointly learns for solving two separate NLP tasks while utilizing a pre-trained language model.
- Sharing the model performance on a ground truth dataset for benchmarking.

The rest of the paper is organized as follows: Section 6.2 describes the dataset used in this paper along with the detail of the data labeling strategy. The methodology, data pre-processing, and feature engineering techniques are described in detail in section 6.3. The detail of the deep neural network architecture is presented in section 6.4, following the experiments in 6.5. Lastly, section 6.6 represents the experimental results following the conclusion.

6.2 Dataset

The deep learning based methodology proposed in this paper consists of a supervised neural network model that requires labeled data to learn the patterns of the disclosure and non-disclosure texts. There might be several reasons why no dataset is available for this purpose in literature, i.e., the restricted access policies of such data sources (e.g., emails, SMS, chat records), lack of privacy preserving research strategies, the complexity associated with the data labeling technique, etc. Therefore, we collected, and human-annotated a ground truth dataset that contains human expressions, comprised of multiple English sentences, through which their privacy might have been

disclosed. The following two sections detail our data collection and data labeling steps.

6.2.1 Data Collection

In order to collect diverse and user-centric data from different domains, we use the online platform Twitter. People tend to prefer this platform to share their personal opinions, perceptions, issues, and observations through tweets which are comprised of a few sentences, hashtags, and emojis. We utilized Twitter search API [155] for mining the required dataset following a set of cleaning and labeling processes. We limited the data collection to those tweets that are written in the English language and from anywhere in the world. This allows us to collect a generalized set of data written in different styles. In addition, the data is limited to the tweets posted between the year 2019 and 2021. The specific crawling dates in this range are randomly chosen by the crawler for better sampling. Most importantly, we filtered out the tweets based on a set of criteria such as i) tweets that contain any links, ii) retweets, iii) replies to the tweets, iv) tweets that are from verified accounts, v) tweets that are posted by bots.

A total of 45,000 tweets is collected from three different privacy domains i) health, ii) finance, iii) relationship. The advanced search query strategies offered by the Twitter API [155] allowed us to properly identify and collect the tweets from these three categories. From these sets of tweets, we sampled a set of 6,000 random tweets based on the stratification of these three information types, selecting 2000 tweets from each category. This smaller subset of dataset is then used for human-annotation and model training. In addition, we maintained the anonymity of the tweets by removing all the metadata excepts the tweet's date-time, tweet texts, and device-type used to

Table 6.1: Example of disclosure and non-disclosure tweets (Samples are taken from the set of 5,400 tweets).

No	Text	Information Type	Disclosure?
1	Ran into two 'mean girl' ex friends today. They're still mean. I was having a bad mental health day too. But I'm choosing to look on it as a lesson that I was right to cut them off. I was having doubts about one of them. Not now.	Health	Yes
2	stop calling me a homewrecker I'm simply breaking up a relationship for my own personal gain RANBOO HELLO	Relationship	Yes
3	We all 7311 Candidates who passed Beltron Deo 2019 2020 exam want joining because our financial condition is so poor and all are workless.	Finance	Yes
4	Financial abuse is so scary amp it's very common. It's why I always discourage women from being transparent about their finances (he doesn't need to know about all your money) or merging finances with a man and not having her own private accounts.	Finance	No
5	Being self aware is sexy. Taking your mental health serious is sexy. Loving yourself sexy. Pretty face and body fades eventually but your mind will always keep developing and expanding.	Health	No
6	Shout out the teachers who talked about their divorce and personal problems and just passed us instead of teaching	Relationship	No

post these tweets. Therefore, usernames, handles, permalinks, or tweet IDs remained hidden from the human annotators. We also meet the Twitter Developer Agreement and Policy [2] by conducting only non-commercial research on this dataset.

6.2.2 Data Labeling

In each of the collected tweets, people tend to share their personal issues, opinions, perceptions, and advice, etc. It is observed that the authors intentionally or unintentionally disclose their own or someone else's private information such as health condition, financial situation, or relationship issues through their tweets. Some examples of such privacy disclosure and non-disclosure tweets can be found in Table 6.1, which are randomly sampled from the 6K dataset.

[2]You may use the Twitter API and Twitter Content to measure and analyze topics like spam, abuse, or other platform health-related topics for non-commercial research purposes.

We recruited human annotators from Amazon Mechanical Turk[3], an online crowdsourcing marketplace to label all of the tweets as either disclosure or non-disclosure. The detailed instructions along with a set of good and bad examples of labeling were provided to assist the annotators to understand the task correctly. We specifically guided them to follow the definitions of disclosure and non-disclosure. We limited the selected annotators to the USA with a good reputation (i.e., at least 95% HIT[4] approval rate) and those who are at least 18 years old. Each annotator was paid $.05 per tweet based on our pilot trials indicating workers could label each tweet within 30 seconds. It is worth mentioning that only the binary labeling of disclosure/non-disclosure was completed by the human annotators. They were not asked to label the information types, since we already assigned these labels as a bi-product while crawling the tweets using the advanced search query API of Twitter. Most importantly, we employed 3 human annotators per tweet to decide whether or not that post is a privacy disclosure. This enabled us to select the most voted label for the tweet as the ground truth.

6.2.3 Data Augmentation

We discovered a moderate level of data imbalance after annotating the dataset. A total of 807 tweets out of 2000 from the health category and 769 tweets out of 2000 from the finance category were labeled as disclosure class, whereas 799 tweets out of 2000 from the relationship category were labeled as non-disclosure. Therefore, we performed a data augmentation step to make the dataset balanced. First, we randomly sampled the candidate tweets to be augmented from each category. We

[3]A crowd sourcing website for businesses and researchers to hire remotely located "crowdworkers" to perform on-demand tasks such as surveys, data labeling, etc.

[4]Human Intelligence Task

Table 6.2: Final dataset (balanced) for model training.

Info Type	# of Disclosure Tweets	# of Non-disclosure Tweets	Total
Health	900	900	1800
Finance	900	900	1800
Relationship	900	900	1800
Total	2700	2700	5400

sampled 93 disclosure tweets from the health category, 131 disclosure tweets from the finance category, and 101 non-disclosure tweets from the relationship category. Then we applied *domain-specific paraphrasing and synonym replacement* technique to these tweets as our augmentation strategy. This simple yet effective approach of augmenting text data has been recommended by researchers and proved to be useful for getting generalized text data [164]. After the augmentation, we got 900 (800+93) disclosure tweets for the health category, 900 (769+131) disclosure tweets from the finance category, and 900 (799+101) non-disclosure tweets for the relationship category. On the other hand, we re-sampled 900 non-disclosure tweets from the health category, 900 non-disclosure tweets from the finance category, and 900 disclosure tweets from the relationship category. This resulted in a balanced dataset of 5,400 tweets where each of health, finance, and relationship categories contained 900 disclosure and 900 non-disclosure tweets (Table 6.2).

6.3 Methodology

The neural network based model proposed in this paper adopts a transformer based pre-trained model called BERT (Bidirectional Encoder Representations from Transformers). We use this state-of-the-art pre-trained model to develop our custom multi-

input multi-output model because i) it supports fine-tuning for custom NLP tasks (transfer learning), ii) it is trained on a huge corpus of unlabeled texts (3,300 million of words), ii) contains millions of parameters (110M), iv) supports parallelization for hardware acceleration, etc. The following subsections further detail on this component along with the data pre-processing and feature-engineering steps.

6.3.1 Data Preprocessing

As depicted in Table 6.1, both disclosure and non-disclosure tweets could contain similar keywords, sentence structure, and other syntactic constructs. This makes the classification problem particularly challenging, because we cannot simply rely on the lexical items and obvious keywords in the text, like bag-of-word models. Rather, we are required to discover the hidden patterns and infer the author's intentions that are embodied in the text, and to encode the underlying meaning expressed in the text to better classify the disclosure/non-disclosure activities. Therefore, unlike the traditional approaches that are mostly based on the bag-of-words technique, we kept the punctuation and stop words in the text to preserve the syntactic structure. We use NLP [63] to clean the tweets in a customized way that ignores noisy and redundant tokens such as *",,", ";–", "!!!", ":-)"* and preserves the non-redundant ones such as *",", ";", ":", ".", "he", "the", "in"* etc. This is in contrast to the traditional approach of text analysis that is based on removing all the punctuation. It is important to note that we also removed Twitter-specific tokens such as , # and non-unicode special characters which might have been added by the users' device.

6.3.2 Feature Engineering

We performed feature engineering on the dataset to produce four new features which then were fed into the neural network through its multiple input channels. Based on

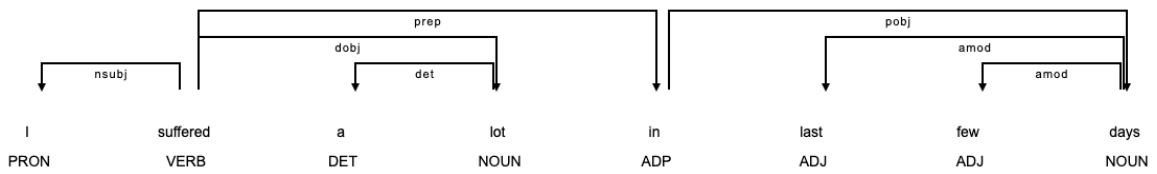


Figure 6.1: Dependency Parse Tree Information of a Sentence.

the Dependency Parse (DP) tree information of the texts, the underlying syntactic relationship of the data was generated. Additional features, i.e., date and time of the tweets, and the type of device that was used to post the tweets are also fed into the network as metadata. Below we explain these new synthetic features in more details.

Syntactic Structure

Certain formal properties of the language, such as dependency parse tree information, are known as "purely stylistic" by theoretical linguistics [132]. In other words, two English sentences might have different syntactic forms but still express a similar meaning or vice versa [49]. For example, (*I suffered a lot in last few days*) with the DP structure *nsubj ROOT det dobj prep amod amod pobj* could be semantically equivalent to another sentence (*In last few days I suffered a lot*) having the structure *prep amod amod pobj nsubj ROOT det npadvmod*, though they are syntactically different. Figure 6.1 depicts the DP information of an example sentence where the DP tags are shown on the edges. Along with the parts of the speech tags, these types of representation of the language features enable the deep-learning based models to learn about the sequential patterns of the sentence constructs along with the arrangements of the word token themselves [110]. This helps the model. Hence, we used a natural language toolkit [63] to extract the information to enrich the feature space of the dataset.

6.3.3 Transfer Learning and Fine Tuning

Most of the NLP tasks such as text classification, machine translation, text generation, language modeling, etc., are considered sequence modeling tasks. Typical machine learning models such as bag-of-words, term-frequency inverse document-frequency, and multi-layer perceptrons are not able to capture the sequential information presented in the text. Therefore, to capture this important piece of information, researchers have introduced techniques such as recurrent neural network (RNN) and long short-term memory based network. However, these types of neural networks introduce new issues in terms of performance and efficiency. For the reason that both RNN and LSTM based neural network takes one input (token in case of text sequence) at a time, they could not be parallelized. This makes the training operation, time-consuming, especially while handling a large dataset.

This was the case until 2018, when Google introduced the transformer model, which turned out to be groundbreaking [161]. It is mainly an attention mechanism for learning contextual relations between words in the text (Figure 6.2). It also introduced an architecture that supports parallelization and makes use of unlabeled text data for training. In the following year, BERT has been introduced, which makes use of the transformer architecture. It is a new language representation model published by researchers from the Google AI Language team in 2018 [43]. Since then, all the tailored solutions to various NLP tasks are being outperformed by this generic transformer based model. Most importantly, BERT supports transfer-learning, which allows us to develop domain-specific custom NLP models while utilizing the power of transformer based pre-trained models. Transfer learning is pre-training a neural network model on an informed task and then using the trained network as the basis

of a new purpose-specific model, otherwise known as fine-tuning [153]. Researchers from the area of computer vision have already shown the significance of this technique [53], and in recent years, they have been showing how a similar technique could be useful in natural language tasks as well [131]. Figure 6.3 depicts an abstract view of BERT's pre-training and fine-tuning Procedures.

Therefore, we adopt the transfer-learning technique to design and develop our hybrid multi-input multi-output neural network, which not only fine-tunes a pre-trained model but also makes use of the linguistic pattern-learning and metadata utilization. There are different ways of fine-tuning a model: i) the entire architecture could be further trained on a new dataset which allows the model to update its pre-trained weights ii) retraining only the higher layers while keeping the weights of initial layers of the model frozen iii) keeping the all the layers of the model frozen, and add one or more new neural network layers of our own, where only the weights of the new layers will be updated during the training phase. In this paper, we utilize the last technique where we import the pre-trained BERT model as a neural network layer into our custom neural network architecture. This acts as one of the three main input channels of our network. The other two input channels provide additional data to the model that we detail in the following sections.

6.4 Neural Network Architecture

The architecture of the proposed neural network is divided into three main segments: i) pre-trained BERT model ii) implementation of the linguistic features iii) integration of structured metadata. The output of this model consists of two different branches: i) multi-class classification of information types ii) binary classification of disclosure/non-disclosure information sharing transactions. Figure 6.4 depicts the ar-

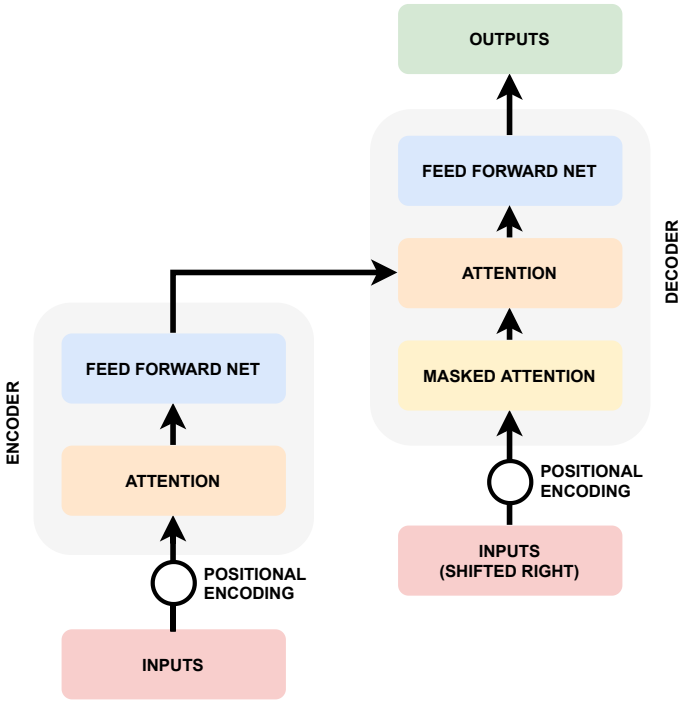


Figure 6.2: Simplified View of the Transform Architecture [161]

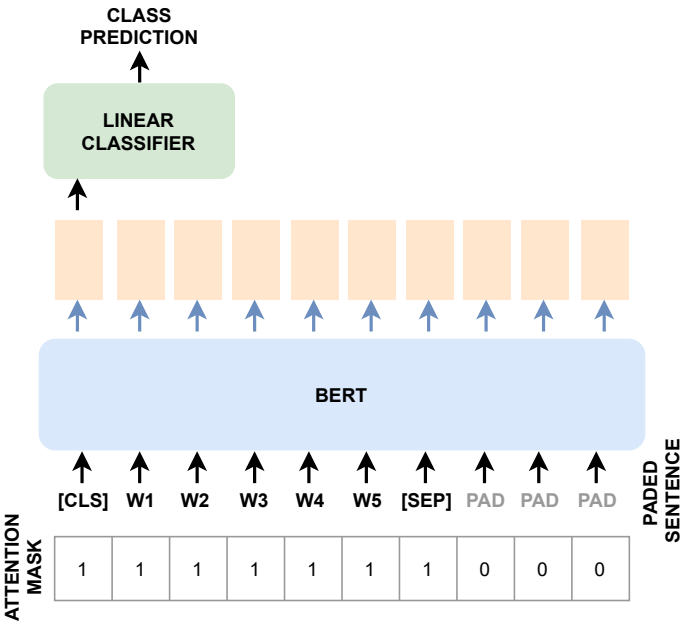


Figure 6.3: Simplified View of BERT Fine-tuning Procedures [43]

chitecture of the proposed multi-input, multi-output hybrid neural network. In the following subsections, we describe each component of the model in detail.

6.4.1 Leveraging BERT

BERT model has two inputs: first from the word tokens, and second from the segment layer following their embedding layers. BERT has a vocabulary of 30,000 distinct tokens comprised of complete English words and word piece components (e.g., embedding for both *play* and *##ing* to work with *playing*). These tokens are associated with an initial embedding space known as WordPiece embedding. The two inputs are added and summed over a third embedding known as position embedding, followed by the dropout layers and layer normalization. The resulting BERT model contains 12 multi-headed self-attention layers (encoders), which are identical to each other. BERT is trained on two NLP tasks: i) the Next Sentence Prediction (NSP), ii) Masked Language Modeling (MLM). These two tasks are informally called fake tasks. In other words, when the pre-training of BERT happens, the model learns the language patterns while solving for these two given tasks. In the end, the trained model is saved and used for further fine-tuning to solve specific NLP tasks, like one in this paper (disclosure and information type classification). Please refer to the original research paper for detailed implementation of the BERT's architecture [43].

6.4.2 Inputs to the Proposed Network

As mentioned previously, we import the pre-trained BERT model as a layer into the proposed neural network. Token-ids and attention-masks are fed into the model through two input channels of the BERT layer. Token ids are the integer encoded values for each of the tokens of the input text. Attention masks are supporting vectors that enable BERT differentiate between the actual and padding tokens. We

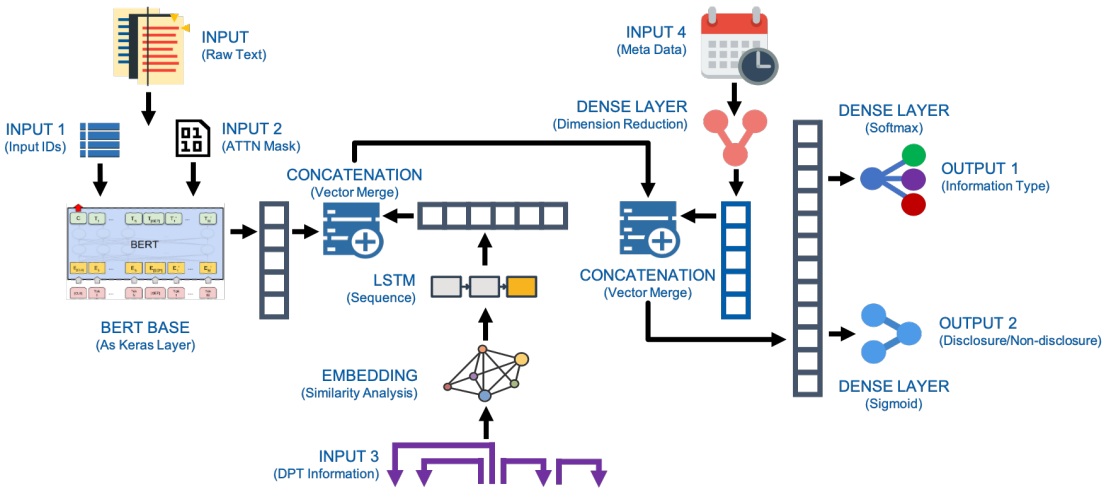


Figure 6.4: Bigger Picture of the Model

add a dropout layer after the BERT main layer as suggested by the literature [65]. In addition, a separate input channel was added to the proposed neural network through which we fed the dependency parse tree information of the same input text. This input path has its own embedding layer which gets learned during the training process. Then we added an LSTM layer that learns the sequential information of the dependency tags. The output of this LSTM layers is then concatenated with the output of the dropout layer. At this stage, we employ the metadata to the neural network through another input channel. This input takes the day of the week, hour of the day, and device type information associated with each input text. A dense layer is added to reduce the dimensionality caused by the encoding of these categorical features. This dense layer uses rectified linear unit as its activation function. Finally, we concatenated the output of this input channel with the output of the previous concatenation operation (BERT's output + DP output).

6.4.3 Outputs from the Proposed Network

Since we aim to solve two parallel tasks through a single neural network model, there are two separate output layers in the proposed model. In one output layer we add three neurons that result in a probability distribution of the information type variable. The predicted probabilities of an input text being any of the three classes: health, finance, relationship are distributed among these three neurons. The neuron with the heights probability wins and shows the information type of the input text. The other output layer is comprised of a single neuron that calculates the probability of the input text being either disclosure or non-disclosure. In other words, the model jointly optimizes for a multi-class classification task and a binary-class classification task. Therefore, we employ different loss functions for these two separate out layers. The multi-class prediction layer uses categorical cross-entropy and the binary class prediction layer uses binary cross-entropy with accuracy as the evaluation metrics.

6.5 Experiments

In this section, we describe the implementation detail of the proposed neural network architecture along with the tools we used. We also talk about the optimizer, loss functions, metrics, and a set of hyper-parameters in this section.

6.5.1 Tools and Libraries

We utilize the Huggingface's Transformers package, which is an open-source natural language processing library developed in Python programming language [169]. This library lets developers import a wide range (32+ pre-trained models in 100+ languages) of transformer-based pre-trained models such as BERT, ALBERT, XLnet, GPT-2, etc. It is also very easy to switch between different transformer based models

through Huggingface Transformers. Most importantly, it supports interoperability between PyTorch, TensorFlow, and other deep learning libraries. We use Tensorflow that comes with Keras pre-built to architect the multi-input multi-output neural network [1]. More specifically, we use the Keras functional API to create the neural network architecture [74].

We make use of the *TFBertModel* module from the Transformers package which is an interface to the Tensorflow library. We import the pre-trained BERT model called *bert-base-uncased* using this module. This is a pre-trained model on the English language, and it is uncased meaning it does not make a difference between the words playing and Playing [44]. This specific base model consists of 110 million parameters. The main layer of this pre-trained model is imported as a keras layer into our custom architecture following a dropout layer. In the other input channel of our model, an LSTM layer with *tanh* activation function is used over the dependency parse tree information by utilizing the keras *LSTM* layer [75]. Before this layer, we use the keras *Embedding* layer to learn the embedding of these dependency tags in a 16-dimensional vector space [73]. The Keras *concatenate* method then takes the output from this LSTM layer and the dropout layer from BERT to merge them into a single vector. The final input into our custom neural network makes use of a keras *Dense* layer [72], and its output is also gets concatenated with the other branch before going through the final output layers.

For text pre-processing, we applied Spacy [63] to derive the dependency parse tree information of each tweet. Spacy provides dependency parser, trainable models, tokenizer, noun chunk separator, etc., in a single toolkit. It offers the fastest syntactic parser in the world and its accuracy is within 1% of the best available natural language

toolkit [30]. To perform the data augmentation step, we used another Spacy based library called spaCy WordNet [129]. It is a custom component for using WordNet and WordNet domains with spaCy which allows users to get synsets for a processed token filtering by domain. Text encoding and padding for these tag based sequences are done using Keras text to sequence and padding methods, respectively [78]. To tokenize, pad, and prepare the raw texts for the BERT side input, we utilize the *BertTokenizerFast* that comes with the Transformer package. This tokenizer converts the raw texts into BERT compatible format such as adding special tokens ([CLS], [SEP]), truncating longer sequences, returning token ids and attention masks, etc.

6.5.2 Optimizer, Loss, and Metrics

We use the *Adam* gradient descent algorithm as the optimization method for the neural network. It is considered to be computationally efficient and has little memory requirement [79]. The separate output heads use two different logarithmic loss functions: categorical cross-entropy for information type classification, and binary cross-entropy for disclosure detection. The network uses *accuracy* as the optimization metrics for both of the output heads which is evaluated by the model during training and testing.

6.5.3 Hyper-parameters

In the case of fine-tuning based training, most of the hyper-parameters of the core model itself stay the same. Therefore, we also retain the hyper-parameters of BERT as it is. However, readers can refer to the BERT paper which gives specific suggestions on the hyper-parameters that require further tuning. In this section we only describe those hyper-parameters which we use for our custom neural network model.

First of all, we consider 55 (mode) as the maximum length of the input text se-

quences. Since the tweets in the dataset are of varying length, we use truncation and padding to make all the tweets have this same length. The first custom input that takes the dependency parse tree information learns an embedding space of length 16 with a vocabulary size of 47. The subsequent LSTM layer is comprised of 32 units which is the dimensionality of its output space. This layer uses *tanh* as the activation function with no *dropout*. All other parameters are kept default from the keras implementation [75]. The Keras *concatenate* method takes the output from this LSTM layer (32 dimensions) and the dropout layer from BERT (768 dimensions) to merge them into a single vector of 800 dimensions. The other custom input channel (metadata input) uses a dense layer with 32 neurons and rectified linear unit as their activation function which reduces its 149-dimensional input to 32. One of the final output layers that classify the information type uses 3 neurons with *softmax* activation. The other output that detects disclosure uses a single neuron with *sigmoid* activation. Both of these output layers use truncated normal distribution as the kernel initializers where the standard deviation is 0.02 for initializing all weight matrices. This value comes as default from the standard implementation of BERT by the Transformer library. The parameters for the Adam optimizer are chosen as follows: learning rate = $5e-04$, epsilon (a small constant for numerical stability) = $1e-08$, clipnorm (gradient norm scaling) = 1.0. Other parameters of this optimizer are kept as default from the Tensorflow implementation of it [76].

The whole dataset is split into a 90-10 ratio for training and testing respectively. For the validation purpose, we kept 20% from the training dataset while the model training process happens. Thus, 10% of the original dataset is used as test dataset which was never shown to the model. We feed the input data to the model with

a batch size of 64, and let the model train for 5 epochs. We achieved the best performance from the model within this amount of iterations. It's worth mentioning that, all the above-mentioned hyper-parameters are chosen based on several trials and outcomes.

6.5.4 Computing Resources

We used Google Colaboratory [52] as the experimentation platform which provided an Nvidia Tesla T4 GPU with 16GB memory. It took 15 minutes on average to run a complete training phase given the hyper-parameters that we mentioned already. Since this platform provides virtual infrastructure and sometimes shares the resources among the users, the reported time may vary.

6.6 Results

The results show that, by utilizing transfer learning and pre-trained language model, a multi-input neural network based model can be trained that learns beyond simple keyword spotting and utilizes linguistic features to classify whether or not a piece of text contains a privacy disclosure with a useful degree of accuracy. Moreover, through the experimentation, it is observed that, integration of metadata to the model increases the performance noticeably (increasing the accuracy by 1.80%). Since our dataset is balanced, we report receiver operating characteristic (ROC) curve, precision [5] and recall [6] score, f1-score [7], confusion matrix, and accuracy [8] score for both the binary and multi-class classification task.

In table 6.3 and table 6.4, we describe the classification report for both infor-

[5] What fraction of predictions as a positive class were actually positive.
[6] What fraction of all positive samples were correctly predicted as positive.
[7] The harmonic mean (average) of the precision and recall.
[8] The fraction of the total samples that were correctly classified.

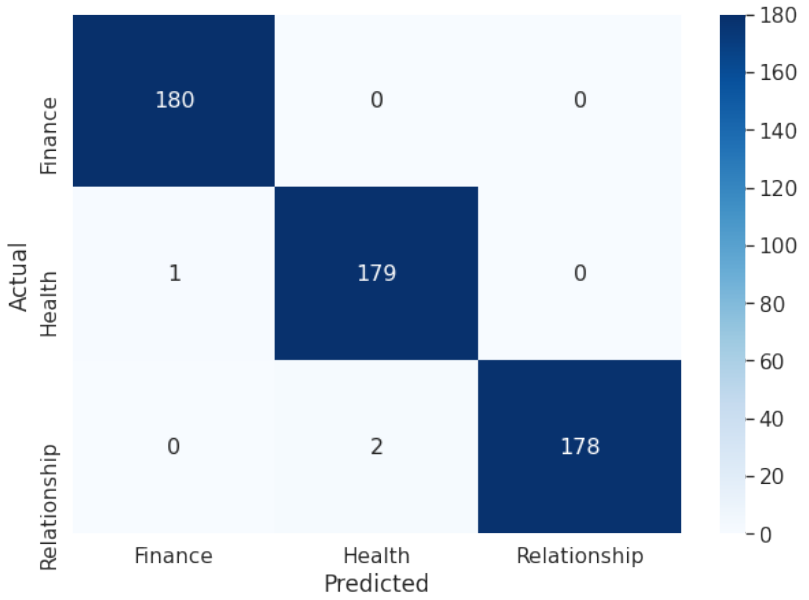


Figure 6.5: Confusion matrix for information type classification.

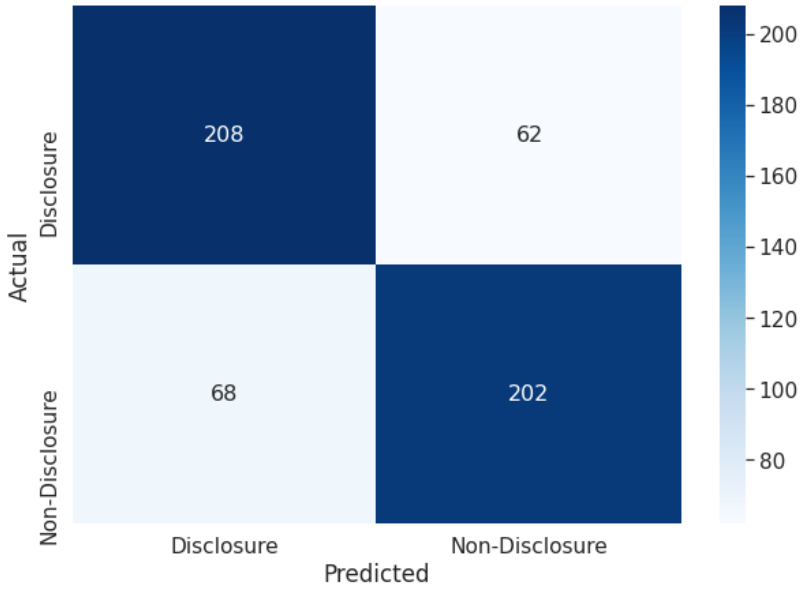


Figure 6.6: Confusion matrix for disclosure classification.

Table 6.3: Classification report for information types.

	Precision	Recall	f1-score	Support
Health	0.99	0.99	0.99	180
Finance	0.99	1.00	1.00	180
Relationship	1.00	0.99	0.99	180
Accuracy			0.99	540
Macro Avg.	0.99	0.99	0.99	540

Table 6.4: Classification report disclosure/non-disclosure.

	Precision	Recall	f1-score	Support
Disclosure	0.75	0.77	0.76	270
Non-disclosure	0.77	0.75	0.76	270
Accuracy			0.76	540
Macro Avg.	0.76	0.76	0.76	540

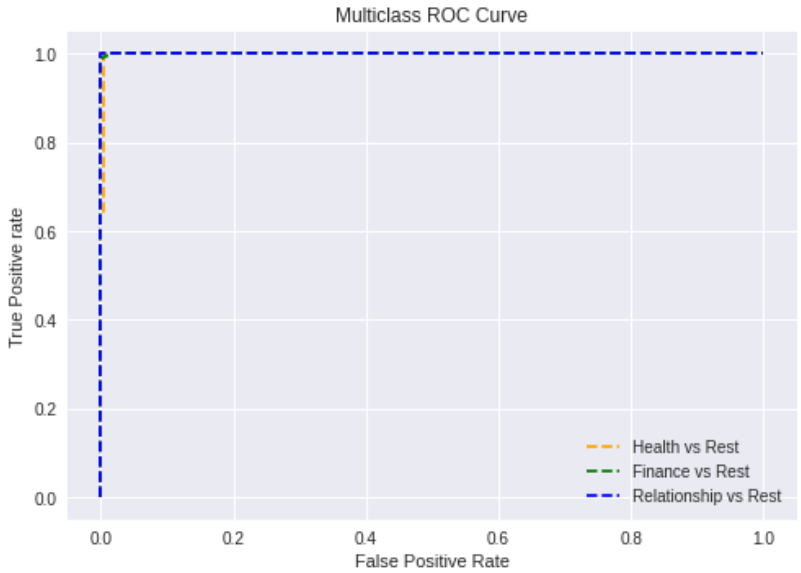


Figure 6.7: ROC curve for information type classification.

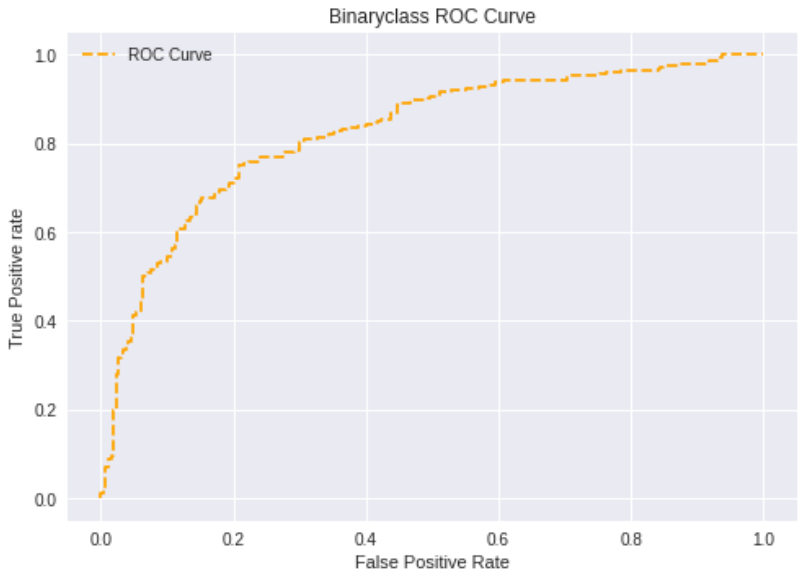


Figure 6.8: ROC curve for disclosure classification.

mation type and disclosure detection respectively. As can be seen from these tables, the information type classifier achieves an impressive accuracy of 99%. The disclosure/non-disclosure classifier reaches up to 76% which is 9% more than bag-of-words and RNN based baseline models. We can also see a good recall score for the binary classifier which depicts its capability to detect most of the disclosure texts. In other words, 77% of all the disclosure texts have been identified successfully. Figure 6.5 depicts the confusion matrix for information type classification. It can be seen that, only a few miss-classifications had occurred, especially when the information type of the texts was *Relationship*. Likewise, figure 6.6 depicts the confusion matrix for disclosure/non-disclosure classification. From this figure, we can see that the event of miss-classifying non-disclosure texts as disclosure happened more than the event of miss-classifying disclosure texts as non-disclosure. This insight is also supportive of our model since the problem domain is privacy, and it's considered safe when there

are more false positive alarms than false negatives. In figure 6.7, we show the ROC curve for information type classification, and in figure 6.8 we show the ROC curve for disclosure/non-disclosure classification. The binary classifier shows an area under curve (AUC) score of 0.82. Unlike the binary class ROC curve, we render the multi-class ROC curve by using the one-vs-all technique to properly represent its performance.

The performance of our model is not directly comparable with other similar approaches proposed in the literature because of the lack of common and shared datasets with similar properties. However, the closest and recent work of detecting self-disclosure on the #OffMyChest dataset, which contains Reddit comments, is worth comparing [39]. In their work, they achieved an accuracy of 74.12% and 74.20% on two different classes of the dataset: information disclosure and emotional disclosure respectively. Also, the precision and recall scores were 0.710, 0.551, and 0.636, 0.510 respectively. In comparison, the performance of our model is noticeably better in all the metrics.

6.7 Conclusion

In this paper we have proposed a multi-input, multi-output hybrid neural network that utilizes the state-of-the-art transformer based pre-trained model called BERT along with language features and metadata to precisely detect privacy disclosure in text data. We also evaluate our model on a ground truth dataset that contains a total of 5,400 tweets from three different privacy domains: health, finance, and relationship. Unlike the traditional text classification techniques that primarily rely on keyword spotting, this model focus on underlying meaning and hidden patterns by leveraging pre-trained language model and classical linguistics. Additionally, our

proposed architecture shows the capability of solving two separate text classification tasks within a single model that provides new insights which can help build practical NLP models. We want to collect a diverse dataset on various privacy domains in the future, using more sources such as forums, emails, text messages, and so on. We also plan to integrate the explainability features into the model for its fairness and trustworthiness.

CHAPTER 7:

MODELING OF PERSONALIZED PRIVACY DISCLOSURE BEHAVIOR: A FORMAL METHOD APPROACH[1]

In order to create user-centric and personalized privacy management tools, the underlying models must account for individual users' privacy expectations, preferences, and their ability to control their information sharing activities. Existing studies of users' privacy behavior modeling attempt to frame the problem from a request's perspective, which lacks the crucial involvement of the information owner, resulting in limited or no control of policy management. Moreover, very few of them take into the consideration the aspect of correctness, explainability, usability, and acceptance of the methodologies for each user of the system. In this paper, we present a methodology to formally model, validate, and verify personalized privacy disclosure behavior based on the analysis of the user's situational decision-making process. This work reuses the same dataset from the chapter 3 survey. We use a model checking tool named UPPAAL to represent users' self-reported privacy disclosure behavior by an extended form of finite state automata (FSA), and perform reachability analysis for the verifi-

[1]Nuhil Mehdy, H. Mehrpouyan, "Modeling of Personalized Privacy Disclosure Behavior: A Formal Method Approach," 2021 *4th International Workshop on Behavioral Authentication for System Security (BASS 21) @ARES 21*

cation of privacy properties through computation tree logic (CTL) formulas. We also describe the practical use cases of the methodology depicting the potential of formal technique towards the design and development of user-centric behavioral modeling. This paper, through extensive amounts of experimental outcomes, contributes several insights to the area of formal methods and user-tailored privacy behavior modeling.

7.1 Introduction

Privacy in the information domain refers to the right of a person to monitor and control the processing, exposition, and preservation of information about themselves. [98]. Accordingly, the responsibility is on the user themselves to take control of what kind of information should be shared with whom, when, and how [166, 109, 70]. However, for an individual, it is quite cumbersome and difficult to manage and control their information sharing preferences [170]. This is because different devices, applications, and software require different privacy configurations from the users, and most importantly, they are not designed to be personalized or assistive. Therefore, it is important than ever before to develop and provide suitable tools and algorithms to the users so that they can define, manage, and make the best use of their privacy preferences with ease. Existing methodologies and protocols intend to tackle this problem by employing techniques such as access control policies [121, 138], machine-readable privacy policy languages [38, 10], formal methods [11, 23], machine learning [36, 152, 108], etc. However, most of the works attempt to frame the problem from a request's perspective which lacks the crucial involvement of the information owner, resulting in limited or no control of policy adjustment. Moreover, very few of them take into consideration the aspect of personalization and explainability of such tools. Most importantly, while there is a significant amount of research aimed at design and

development of privacy management tools and techniques, 'their practical usability and acceptance remain an important challenge' [90].

Therefore, this paper applies model-based analysis to personalize privacy behavior which answers two key research questions: how to model privacy behavior and how to use this privacy behavior model for analysis. We decomposed this problem into three subcategories: (I) Identification of relevant privacy behavior and situational factors, (II) applying proper modeling techniques, (III) validating the models.

As part of model-based approach, we focus on formal methods that are concerned with modeling, specifying, and verifying any systems using mathematical techniques otherwise known as model checking [33]. A system could be physical or conceptual comprised of interconnected components such as processes, states, nodes, etc. Model checking is an automated approach to verify that a model of a system, usually a finite-state machine, satisfies a set of desired properties (i.e., requirement specifications) written in a temporal logic [57]. This is achieved by exhaustively searching a system's state space in order to determine if these criteria hold. If there is a violation, an error trace is produced (i.e., a counterexample). Model checkers take system description (i.e., formal model) and a set of requirements as input and reason whether the requirements are satisfied or not. In privacy literature, human decision-making, in other words, an individual's intention to disclose private information is also considered as a process which involves different components, otherwise known as influential factors [7]. When the number of factors is large, doing manual specifications and testing of the privacy policies is difficult. It is also possible that subtle conditions get unnoticed. Again, a way to tackle this problem to a certain extent, is the use of mathematically-based techniques. Hence, we adopt the analogy of finite state ma-

chines from the theory of computation and aim to model human privacy disclosure behavior based on this specific formalism technique.

That being said, to learn user's privacy behavior towards the development of user-specific models, it is important to investigate the factors and parameters that influence users to make dynamic privacy decisions [7, 126, 92, 99]. The decision to exchange private information, as well as the risk perceptions that drive this decision, differs from situation to situation. Various considerations, such as the type of information, the receiver of the information, and the source of confidence underlying the reason for sharing, all play a role in the decision-making process [68, 141]. Moreover, risk assessment, potential risks consideration, and alternate exploration are all part of the process of deciding what to do in a specific situation [165, 5]. Additionally, individual variations in demographics, personality traits, and decision-making styles as well as their effect on users' privacy-related habits must be studied before developing any behavioral model. Therefore, we work on the dataset from chapter 3 which was obtained by conducting a custom-designed survey on Amazon Mechanical Turk [2] (N=401) based on the theory of planned behavior (TPB) to measure the way users' perceptions of privacy factors and intent to disclose information are affected by three situational factors embodied hypothetical scenarios: information type, recipients' role, and trust source.

In this work, we chose to focus on the user's situational decision-making process and represent our approach to formally model, validate, and verify personalized privacy behavior. We represent a scaled-down version of our proposed methodology where we model each individual's privacy disclosure behavior where their disclosure

[2] A crowd sourcing website for businesses and researchers to hire remotely located "crowdworkers" to perform on-demand tasks such as surveys, data labeling, etc.

decision merely rely on three factors— information type, recipients' role, and trust source. Even though human decisions depend on many more factors, we chose this level of abstraction because the dataset in hand captures the users' privacy behavior based on these three factors. On the other hand, we wanted to evaluate our approach on top of a ground-truth dataset. Nevertheless, the methodology presented in this paper depicts the potential of formalism towards the development of privacy management tools. This paper is the first to our knowledge to leverage an extended version of automata-based transitioned systems towards modeling individual's privacy behavior. This work provides insight into:

- Model-based analysis of personalized privacy behavior
- Formulate personalized privacy policies
- Detect and reason about unwanted disclosure behavior
- Validate the proposed model-based approach and demonstrate its practicality

7.2 Learning Privacy Preference

In this work, we represent and evaluate our formal method approach to model users' privacy disclosure behavior based on a dataset that we obtained through a survey. We captured users' situational privacy decisions, through a custom scenario-based survey with 401 participants, each responding to a subset of 48 total unique scenarios. Every data point is referred to the responses to a series of questionnaires that assess participants' attitudes toward each situation, as well as their expectations of and willingness to reveal personal information in the given situation. By manipulating three situational factors: information type, recipient's role, and trust source,

we use path analysis to model participants' privacy perceptions and plans, taking into account their assessments on subjective norm, perceived behavioral control, and attitude. This choice of factors is partly inspired by the theory of contextual integrity (CI) [118, 16]. The findings show how users make privacy decisions in a variety of contexts, as well as how situational factors influence users' views of privacy factors and their willingness to share private information. Most importantly, the results also reveal how every individual has their own preferences and concerns about disclosing their private information in certain situations. Therefore, this dataset best suit our personalized behavioral modeling experiment. The following sections describe the survey strategy and the data set in more detail.

7.2.1 Survey

After agreeing to participate in the survey, a person is given a series of eight hypothetical scenarios and asked to answer them one by one. Each scenario places the subject in a position where he or she must choose whether or not to reveal the information embodied in that scenario. This includes the situational factors on which participants can place a high degree of confidence in their interpretation and decision on whether or not to disclose. We manipulate three situational factors to see how they affect participant responses:

Information Type (IT) The type of the information that is illustrated in the scenario. Each scenario is about one of three information types: health, finance, or relationship.

Recipient's Role (RR) The type of the recipient, based on the relationship to the survey participant, to whom the information may be disclosed. We take into

account four such recipient roles: family, friend, colleague, and online service (e.g., facebook, twitter, discussion forum, etc.).

Trust Source (TS) From whom the participant got the motivation of disclosing the information to the recipient. We consider four trust sources: family, friend, expert (e.g., physician, counselor, financial adviser, etc.), and self (i.e., searching the internet).

Different combinations of these factors yield a total of 48 (3*4*4) unique scenarios. For each combination, we prepare a scenario where a trust source encourages the participant to share the information with a recipient. We made every scenario as similar as possible to minimize extraneous variability while incorporating the factors in a natural and coherent manner in the hypothetical scenario. In other words, we made sure the framing of the scenarios does not become significantly different from each other so that only the factors get changed, and a proper parametric analysis is justified. An example scenario with *health* as information type, *friend* as trust source, and *family member* as recipient's role could be:

> Your doctor called and told you that your lab results came back positive for a disease. One of your friends suggested discussing the situation with a family member and asking their support, saying it could be helpful.

Another unique scenario could be generated by changing the trust source from friend to family and recipient's role from family to online:

> Your doctor called and told you that your lab results came back positive for a disease. A family member suggested asking other patients and doc-

tors on an online discussion forum, saying they have found it helpful for dealing with their similar condition.

Every participant is assigned a set of 8 random scenarios with associated questionnaires. A participant has to read a given scenario and respond to all of the corresponding questions before proceeding to the next assigned scenario. We used rejection sampling to ensure that each user's 8 scenarios covered all 11 distinct factor levels at least once, ensuring a minimal degree of heterogeneity between their circumstances and, as a result, responses. To minimize order effects, we also randomly order the set of 8 scenarios for each participant. In the end, the individual completes a brief survey in which we intend to capture their general privacy attitudes regardless of any specific situation. This move is intended to capture expectations that are believed to be constant over time and do not alter in response to changing circumstances. Participants are asked to optionally enter their ethnicity, age group, country of origin, and period of residence in that country in the final phase of the survey for accumulating demographic information.

There are two sets of questions in the survey: i) scenario-specific questions (12 total) and ii) general attitude questions (4 total). For each of the eight scenarios allocated to each person, the first set of 12 questions is repeated. At the end of the survey, the second set of questions is presented. The scenario-specific questionnaire is inspired by [60], and the second set of questions is inspired by prominent privacy research [24, 3]. Appendix C.1 shows a screenshot of the survey system representing 1 of 8 random scenarios given to a participant, and appendix C.2 shows the screenshot representing the general attitude questions given to a participant at the end of the survey. Before the main survey, we conducted a pilot test with six of our research lab's

colleagues. Their feedback was instrumental in resolving problems with the survey interface, user experience, and clarity of the scenarios and questionnaires. Later we used Amazon Mechanical Turk, an online crowd-sourcing marketplace, to find participants for the final survey. We looked for workers from the United States who are at least 18 years old with at least 95% HIT (Human Intelligence Task) acceptance rate[3] and 50 hits approved.

7.2.2 Dataset

We employed a number of filters to ensure the quality of the data. First, we capture the time a participant spent on each scenario step and removed the data points from our analysis if the spent time was too low. Second, we randomly placed attention search questions in between survey questions. We also restricted repeated submissions from the same participant by setting a browser cookie for 3 days after a satisfactory submission. The answers to the questions were translated into a numeric format (1 to 5) from the 5-point scale (ranging from Strongly Disagree to Strongly Agree). For the final decision question, we represent the Share and Not-Share options in logical numeric form, 1 and 0. In the end, we get 3208 data points, grouped by 401 participants, containing their information disclosure decisions based on different situational factors.

7.2.3 Path Model for Privacy Behavior Analysis

In one of our earlier works [107], we leveraged the data to measure users' behavioral intention and their situational perception of three constructs: attitude, subjective norm, and perceived behavioral control. These constructs and the path model are inspired by the theory of planned behavior (TPB) [7]. We also, incorporated the

[3]whose previous works got approved by 95% of the requesters.

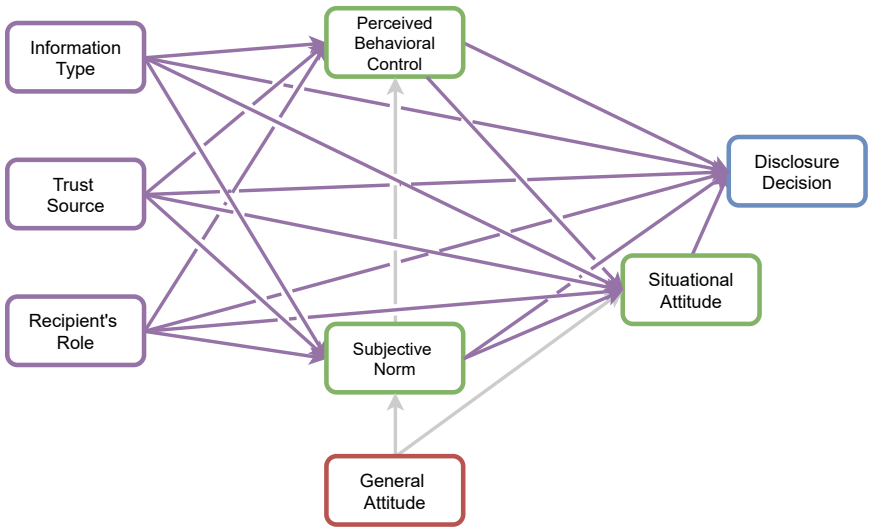


Figure 7.1: The Path Model for Analyzing Users' Privacy Decision-making Process.

scenario factors— information type, recipient's role, and trust source in our path analysis to measure the correlation of these factors with the information disclosure decision of the user. Figure 7.1 depicts the path model. The analysis results show that the path model fits the data very well with $\chi^2_{11} = 12.017$, $p = 0.3623$, $CFI = 1.0$, $TLI = 0.99$, $SRMR = 0.008$, $RMSEA = 0.005$, 90% $CI = 0.000$ to 0.020. Also, the comparative fit index (CFI) and Tucker-Lewis index (TLI) values which range from 0 to 1 show near-perfect scores.

Among all the path analysis results published in our work, one of the findings shows that there exist significant (indirect) effects of the scenario factors on the users' disclosure decisions. In Figure 7.1, they refer to the paths from the purple leftmost boxes to the blue rightmost box via the mediator green boxes in between. These total effects describe *how* users' intention changes from one scenario to another; the mediating TBP factors provide an explanation for *why*. A few important findings include but not limited to— with regard to the recipient's role in the scenario, compared

to the recipient "online service", the odds of disclosure were estimated to be 16.6% higher when the recipient was a family member and 12.9% higher when the recipient was a friend; with regard to the type of information, compared to relationship information, the odds of disclosure were estimated to be 3.1% lower when the scenario involved financial information and recipient was a family member and 5.1% higher when the scenario involved health information, etc. These results indeed proof the influence of the situational factors towards users' disclosure decisions and therefore act as the basic components of our formal privacy behavioral model.

7.3 Formal Modeling

This section describes the approach of developing the formal model of a user's privacy disclosure behavior by taking into account the privacy decisions made by that user. Our approach aims to address the issue of formally modeling the privacy behavior of a user which could be eventually utilized to develop a personalized privacy management system. The whole approach is divided into four main stages: i) observing user's historical sharing activity, ii) modeling users' personalized privacy behavior, iii) validating the model, iv) verifying the model given the privacy properties of the user. We have already detailed the survey and the dataset in the earlier sections which refer to the first stage.

7.3.1 Model Assumptions

In this work, one of the main assumptions of the users' disclosure behavior is that the user decides to share/not-share a specific type of information with a certain type of recipient(s) after being advised by a specific trust source. We represent this knowledge and the decision made by the user in the form of a state model. Transitions between

states occur with respect to a specific information type, trust source, and recipient's role. We also assume that there are no other factors/components involved in the user's decision-making process. Additionally, we assume that the user's behavior could conceivably be modeled as a finite state machine. This research utilizes finite state automata (FSA) extended with data variables to model the privacy disclosure behavior of the users.

7.3.2 Model Paradigm

FSA as a chosen formalism allows for a design and development of well-structured tools to conduct an automated analysis during the early stages of studying user's privacy behavior. Accordingly, there are various tools for designing and verifying such FSA based formal models, i.e., NuSMV, PVS, Z3, and UPPAAL, are a few examples. We choose UPPAAL because of its ability to support model checking over a network of automata using temporal logic [19]. UPPAAL also supports formalism through parallel compositionality among the automata. This modeling paradigm helps us to retrieve the traces of the transition while checking for a given query. Therefore, this modeling paradigm enables us to execute the requirements as temporal logic queries which in turn exhaustively check the satisfaction of the privacy properties. On the other hand, counterexamples are provided to reason about privacy properties that are violated.

7.4 Modeling in UPPAAL

The reason for selecting UPPAAL is because UPPAAL provides a better graphical user-interface that allows for the development, modification, validation, and verification of any system model with drag and drop interface [91]. In UPPAAL, a system is

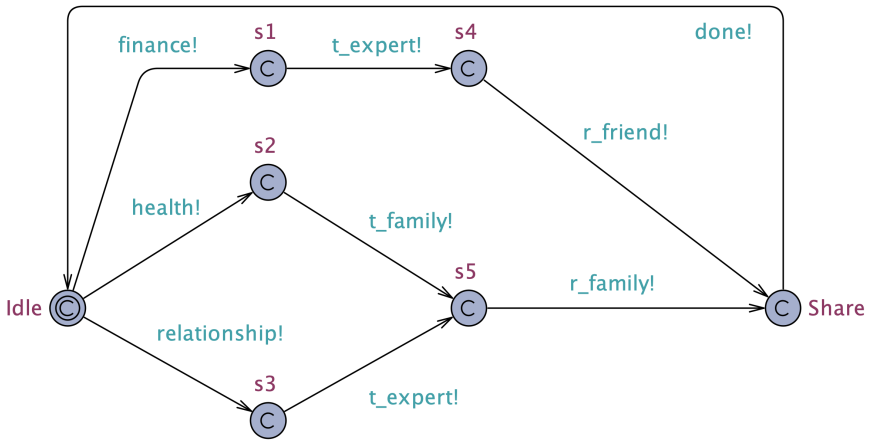


Figure 7.2: The Behavioral Model of User 89 Created in UPPAAL

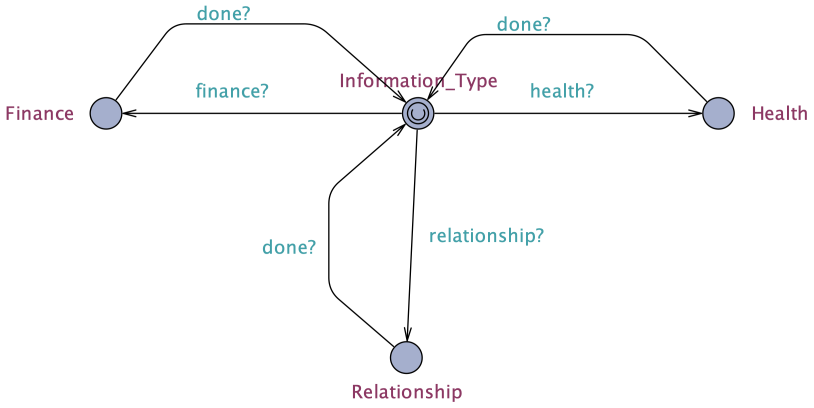
made up of several concurrent processes, each of which is modeled as an automaton. Each automaton has a set of locations otherwise known as states. Transitions between these states could be managed by guard and synchronization. A guard imposes conditions on variables and clocks ensuring when the transition is enabled. Synchronization in UPPAAL enables two or more processes to communicate with each other based on a hand-shaking synchronization. Two actions are possible while a transition happens—assignment of variables or reset of clocks. UPPAAL further extends timed automata with other types of data variables such as integer and Boolean towards developing a modeling language that is as close as a high-level programming language [91].

7.4.1 Behavioral Analysis and Personalization

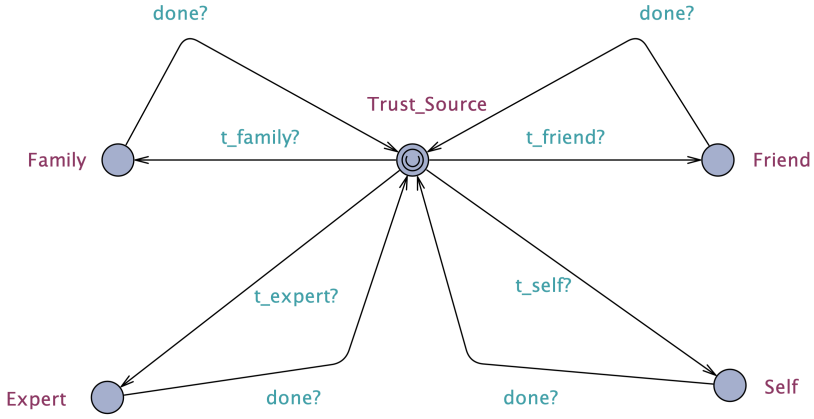
To model the privacy disclosure behavior of a specific user, we collect the user's responses to the survey questionnaire and observe the information sharing behavior in different scenarios. For this, we randomly pick a user, for example, number 89 in our tabular dataset. Table 7.1 contains the 8 random scenarios which were assigned to this user. Table 7.1 represents that the user agreed to share information in 3 out of 8

Table 7.1: Disclosure Decisions by the User 89 Captured by the Survey.

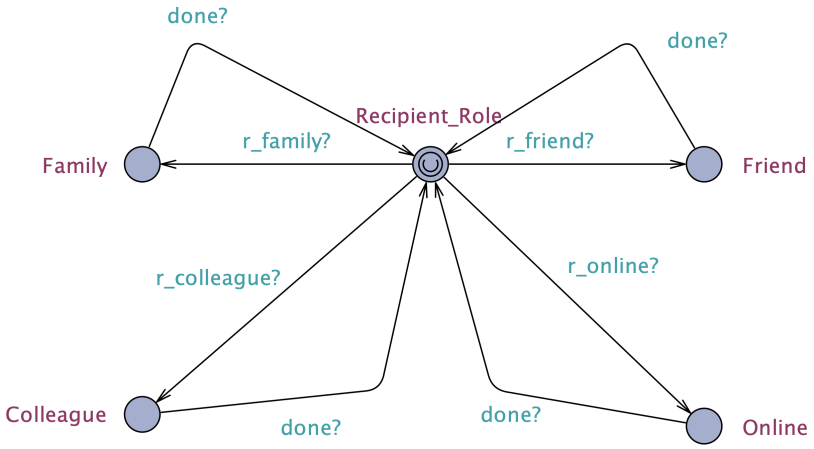
No	Scenario	IT	TS	RR	Share
1	You recently had a very bad argument with your partner. Your counsellor suggested sharing and discussing this matter with a family member, saying they could support you.	Rel	Exp	Fam	Yes
2	Your doctor called and told you that your lab results came back positive for a disease. A family member suggested discussing the situation with a family member and asking their support, saying it could be helpful.	Hea	Fam	Fam	Yes
3	Your doctor called and told you that your lab results came back positive for a disease. A family member suggested asking other patients and doctors on an online discussion forum, saying they have found it helpful for dealing with their similar condition.	Hea	Fam	Onl	No
4	You recently had a very bad argument with your partner. One of your friends suggested asking on an online discussion forum they use to get support from others, saying they have found it helpful for dealing with their situation.	Rel	Fri	Onl	No
5	Your doctor called and told you that your lab results came back positive for a disease. You did some research and found that people often find it helpful to get support from a colleague.	Hea	Sel	Col	No
6	You received a notice from a collection agency saying you have a debt which needs immediate attention. A family member suggested asking on an online discussion forum they use to get support from others, saying they have found it helpful for managing a similar situation.	Fin	Fam	Onl	No
7	You received a notice from a collection agency saying you have a debt which needs immediate attention. Your financial advisor suggest discussing the situation with a friend and asking their support, saying it could be helpful.	Fin	Exp	Fri	Yes
8	You recently had a very bad argument with your partner. A family member suggested sharing and discussing this matter with a colleague, saying they could support you.	Rel	Fam	Col	No



(a) Information Type Observer



(b) Trust Source Observer



(c) Recipient Role Observer

Figure 7.3: Observer Models Created in UPPAAL

given situations (scenarios 1, 2, and 7). Based on that, we model the privacy behavior by composing them into a data-dependent transition graph (Figure 7.2). This graph contains a set of states and synchronization operations. When a transition happens from one state to another, a message is emitted to one or more observer processes through the synchronization channel. For example, when a transition happens from the *Idle* state to state *s1*, it emits a message titled *finance* to any listening processes. This is one of the many useful features of UPPAAL which allows to design network of FSMs (i.e., parallel composition). The start and end states are marked as "committed states", which means there would be immediate transitions from these two states as soon as the transitions are enabled. In UPPAAL, the committed states take prompt transitions when the simulation or exhaustive search happens. This feature allows us to simulate the transitions spontaneously without waiting for any external inputs. It is worth mentioning that, we only model the positive sharing behavior of each user. In other words, figure 7.2 only contains a composition of 3 different scenarios where this user agreed to share the information with the recipients. Hence, if an information sharing attempt, described as a query, fails to comply with the model in figure 7.2, then the model checker tells that the corresponding query was not satisfied and also shows a counter-example trace (if available).

7.4.2 Observer Models

An observer is an add-on automaton that without perturbing the observed system can detect events. We use 3 such models along with the user's behavioral model (Figure 7.2) to keep track of the transitions and associated factors. This eventually helps to prepare and employ descriptive queries for the verification of the model. Figure 7.3 depicts those 3 separate observer models. Figure 7.3 (a) represents the observer

which keeps track of the information types. It listens for the messages— *finance*, *health*, and *relationship* whenever a transition in the behavioral model emits one of these values. For example, if a transition happens from *Idle* state to the *s1* state in the behavioral model (Figure 7.2) then this observer model transitions from the *Information_Type* state to the *Finance* state. The activities of the other two observer models (Figure 7.3 (b) and (c)) are similar. Model 7.3 (b) listens for the messages *t_family*, *t_friend*, *t_expert*, and *t_self* to keep track of the trust source. Likewise, model 7.3 (c) listens for the messages *r_family*, *r_friend*, *r_colleague*, and *r_online* to keep track of the recipient's role. All the observer models return to their initial state once they get a specific message - *done* from the behavioral model.

7.4.3 Behavior as Systems

The user-specific behavior model along with the observer models creates the network automata otherwise known as a concurrent system in UPPAAL. This type of composition is also known as a parallel composition of processes made of automaton. In our setup, the user model synchronizes data between itself and the observer models by leveraging the channel features in UPPAAL. The formal definition of the system model could be defined as follows:

$$User||Information\_Type||Trust\_Source||Recipient\_Role$$

7.4.4 Validation

UPPAAL uses graphical simulation as the model validation strategy [19]. Therefore, we conduct a simulation step to validate our models by running the system automatically which ensure that the models behave as intended, without any unexpected

crash or deadlock. By utilizing the simulation feature of UPPAAL, we manually conduct some transitions in the behavioral model, and also utilize the random simulation feature to make sure the transitions are taken as expected. Figure 7.4 shows the UPPAAL simulation control panel where, the button *Reset* and *Next* are used to manually perform some transition operations, and the button *Random* is used to start an automatic simulation that can run indefinitely. The simulation also allows us to make sure that the concurrency operation between the behavioral and the observer processes is taking place without any system breakdown.

7.5 Verification with Model Checking

In this section, the verification of the user's privacy disclosure model is explained. Figure 7.5 depicts a high-level abstraction of the model checking process. In this approach, a set of desired properties (i.e., specifications) are checked against a model of a system [21, 47].

7.5.1 Specification Language in UPPAAL

The set of privacy properties (i.e., requirement specifications), which we expect the formal model to verify, are formulated based on the conducted survey in section 2.1. The specification languages that could be used to express these types of privacy properties are Linear Temporal Logic (LTL), Computation Tree Logic (CTL), and Timed Computation Tree Logic (TCTL)[12]. In UPPAAL, the process of verification operates with a simplified version of TCTL which is a subset of CTL. In TCTL, temporal connectiveness is expressed as pairs of symbols where the first element represents one of the path quantifiers and the second element represents one of the state quantifiers. Likewise, UPAAL query language consists of path formulae and

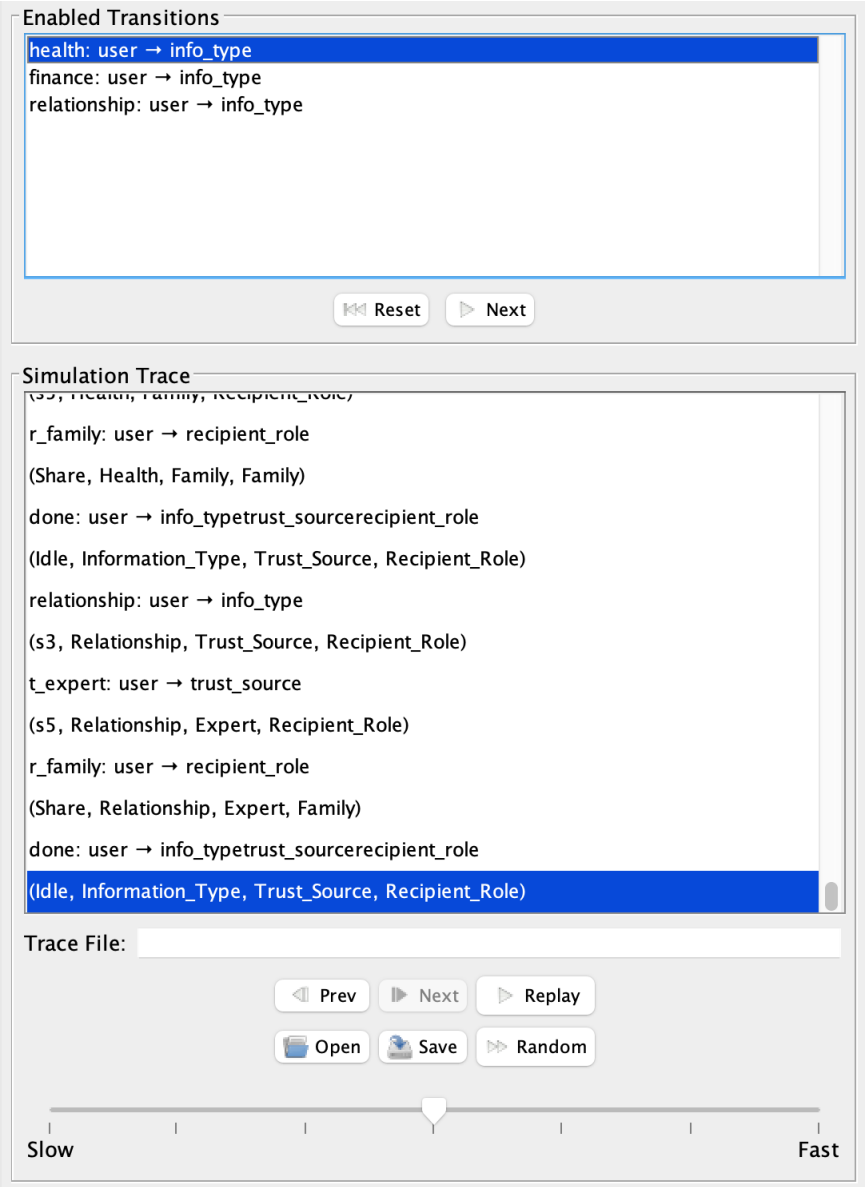


Figure 7.4: Part of the Simulation Window Containing the Control Buttons for Automatic and Manual Transition.

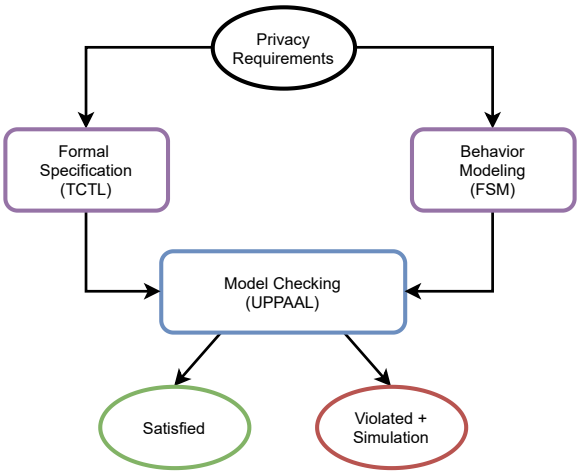


Figure 7.5: Model checking approach

state formulae [19]. The path formulae quantify over paths (traces) of the model, whereas state formulae describe individual states. In UPPAAL, these quantifiers are expressed as follows:

E = exists a path (E in UPPAAL),

A = for all paths (A in UPPAAL),

F = some state in a path (<> in UPPAAL),

G = all states in a path ([] in UPPAAL),

Example queries could be written as $A[]p$, $A <> p$, $E <> p$, $E[]p$, and $p \rightarrow q$ where p and q are local properties. In other words, the query $E <> p$ tells that, 'it is possible to reach a state in which p is satisfied' or 'p is true in at least one reachable state. $E <> Process.End$ is the UPPAAL notation for the same temporal logic formula $\exists\Diamond Process.End$ and is understood as 'it is possible to reach the location End in automaton $Process$'.

7.5.2 Personalized Privacy Properties

In order to formulate the privacy properties of user 89, we translate the user's disclosure decisions that are represented in Table 7.1 in to the following statements: 'if the information type is *health* and the trust source is a *family* member and the recipient of the information is also a *family* member, then the user *share* the information. Similar to this specific criteria, every user has their own requirements when they agree to share the private information based on the situational factors. For each user, we translate their own privacy disclosure criteria into UPPAAL specification formulas. These formulas are then checked against his/her behavioral model to ensure the correctness of it. Since we use observer models (Figure 7.3) along with the behavioral model (Figure 7.2) to create a concurrent system model, the observers have their own formal specifications. In Table 7.3, we represent the equivalent expressions of the scenario factors in UPPAAL's specification language, while Figure 7.3 visualizes the state transition graphs of those factors. Thus, the privacy disclosure properties for user 89 are represented in Table 7.2 that is the transformation of his/her responses based on the scenarios 1,2, and 3 from Table 7.1. Therefore, property number 1 from Table 7.2 expresses: there exists a path, eventually where the properties enclosed in the parenthesis are true.

7.5.3 Reachability Analysis

There are three types of properties that are commonly checked against a formal model— safety, liveness, and reachability properties. Reachability properties are used in state-transition systems which helps to examine the type and number of states that can be accessed through a particular system model [85]. It is the simplest form of properties that determines whether a given state formula, Φ, possibly could be

Table 7.2: Requirement Specifications or Privacy Properties of User 89

No	Privacy Property
1	$E <>$ (user.share and information_type.Health and trust_source.Family and recipient_role.Family)
2	$E <>$ (user.share and information_type.Relationship and trust_source.Expert and recipient_role.Family)
3	$E <>$ (user.share and information_type.Finance and trust_source.Expert and recipient_role.Friend)

Table 7.3: Scenario Factors' Properties

No	Knowledge Base Property
1	$E <> (information\_type.Health)$
2	$E <> (information\_type.Finance)$
3	$E <> (information\_type.Relationship)$
4	$E <> (trust\_source.Family)$
5	$E <> (trust\_source.Friend)$
6	$E <> (trust\_source.Expert)$
7	$E <> (trust\_source.Self\_Search)$
8	$E <> (recipient\_role.Family)$
9	$E <> (recipient\_role.Friend)$
10	$E <> (recipient\_role.Colleague)$
12	$E <> (recipient\_role.Online\_Service)$

satisfied by any reachable state. In this work, we verify whether or not the user-specific privacy properties holds in any, some, or all state of that user's privacy behavior model. We prefer reachability analysis over other similar methods (e.g., graph matching approach) because it allows us to search all potential paths in which the properties may or may not be satisfied, in a thorough and automated manner. Using UPPAAL, we applied reachability analysis to check which privacy properties were satisfied and which were not. UPPAAL performs the reachability analysis using either Breadth-First-Search or Depth-First-Search for checking whether a state is reachable or not. We preferred BFS of DFS to verify our reachability properties because it is a complete algorithm, ends within a finite time, and considers the fewest edges while searching. The results of this procedure allow us to examine a user's privacy disclosure behavior, and whether or not a new sharing attempt complies with her existing privacy policies.

Table 7.4 contains a few verification queries that we check against the privacy disclosure model of user 89. Query 1 indeed gets satisfied since there is a valid transition in the FSM model (Idle $->$ s2 $->$ s5 $->$ Share) as well as in its CTL version which is verified by the TCTL formula. Query 2 does not get satisfied since this user had no history of sharing his *Health* information to either *Friend* or *Online* even when the trust source was *Family*. Query 3 does not get satisfied because there is indeed one path where the property is true, (in Figure 7.2, Idle $->$ s1 $->$ s4 $->$ Share). We can even see the diagnostic trace when this query is executed (Figure 7.6). Additionally, we can verify that the model will not face any deadlock in it's lifespan by executing queries like #4.

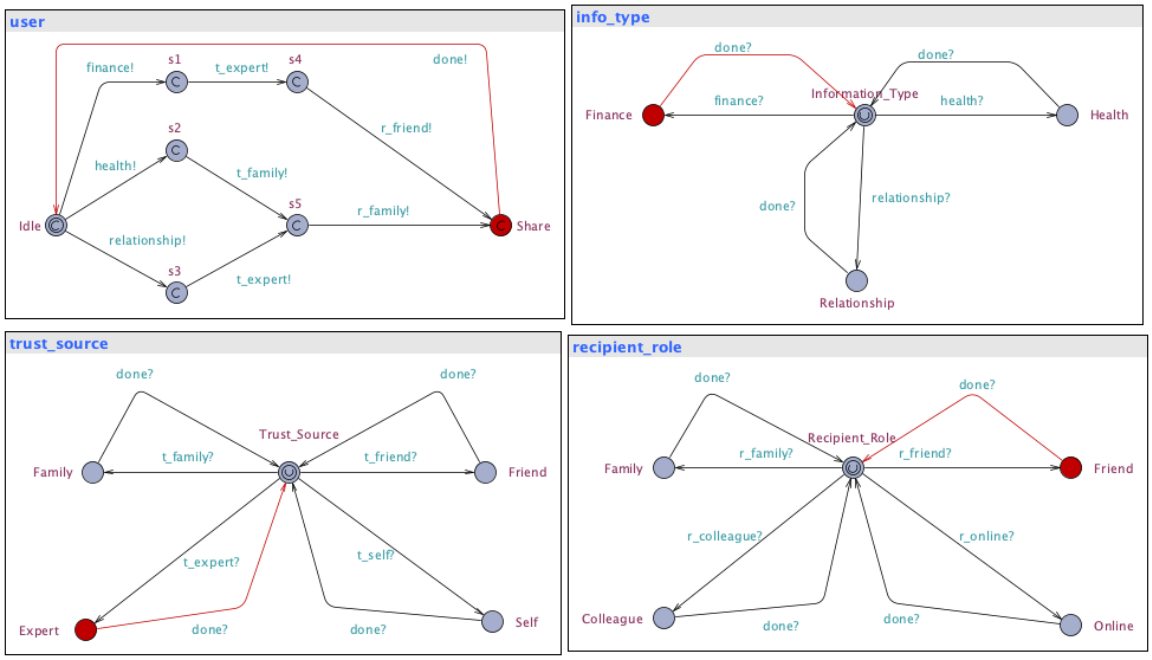


Figure 7.6: Diagnostic Trace of Query 3

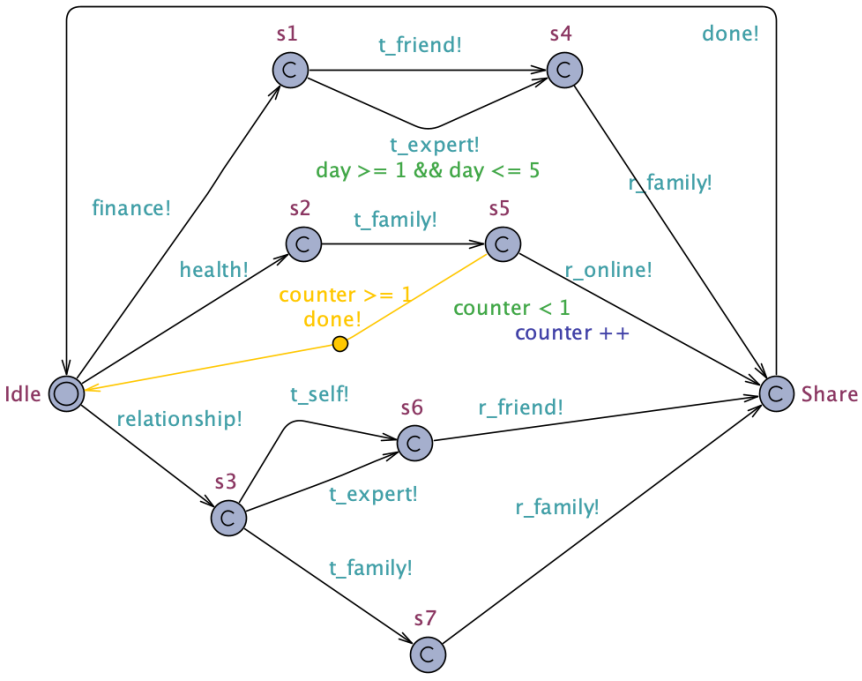


Figure 7.7: The Model of User 242 Created in UPPAAL

Table 7.4: Example of Some Queries to the User 89's Model and the Verification Results

No	To Verify	UPPAAL Query	Verification
1	There exists a path where eventually the user is in *Share* state and the information type was *Health*, trust source was *Family*, and recipient's role was *Family*	$E <>$(user.Share and info_type.Health and trust_source.Family and recipient_role.Family)	Satisfied
2	There exists a path where eventually the user is in *Share* state and the information type was *Health*, trust source was *Family*, and recipient's role was either *Friend* or *Online*	$E <>$(user.Share and info_type.Health and trust_source.Family and (recipient_role.Friend or recipient_role.Online))	Not Satisfied
3	For all paths, it should never be the case that the user is in *Share* state and the information type was *Finance*, trust source was *Expert*, and recipient's role was *Friend*	$A[]$ not (user.Share and info_type.Finance and trust_source.Expert and recipient_role.Friend)	Not Satisfied
4	There should not be any states without successors	$E <>$ not deadlock	Satisfied

7.6 Different Use Cases

In this section, we represent the privacy disclosure model of a different user. Similarly, a user is selected from the dataset randomly and holds the ID 242.In this case, in order to demonstrate the potential of the proposed behavioral model approach to include complex privacy properties with additional constraints, we imposed limitations on the days of the week or the number of times specific information could be disclosed. For this user, the responses that were received to the randomly assigned scenarios, we observed that this user agreed to share the information in 6 out of 8 situations. Therefore, we model his/her disclosure behavior in terms of a transition system (i.e., finite state machine) which is depicted in Figure 7.7. As mentioned already, we added two guard conditions on two edges of the FSM: I) the day of the week for information sharing has to be between Monday to Friday (encoded as 1-5) to make the path- *Idle* $->$ *Expert* $->$ *Family* $->$ *Share* enabled, II) *Health* type information could be shared *Online* no more than twice through the path *Idle* $->$ *Health* $->$ *Family*

$- > Online$. However, while verifying the model, we find a deadlock by querying *E¡¿ not deadlock* to the model checker. This property does not get satisfied depicting that there is indeed a deadlock. This happens because of the *counter* guard which was imposed on the path, $Idle - > Health - > Family - > Online$. Since a query is checked exhaustively, by running/simulating the model for hundreds of iterations, UPPAAL reaches this deadlocked state after simulating through this path twice. In other words, the *counter* becomes 1 and the path becomes disabled from the state *s5* to *Share*. However, we then resolve that deadlock in the model by creating a path from *s5* to *Idle* (colored in yellow). Thus, whenever the model checker tries to go through this path more than twice and faces a guard in state *s5*, it can then safely get back to the initial state without blocking the simulation operations.

In some other cases, incorporating additional decision-making factors or adding subcategories to the existing ones may result in a more complex network of automata with added granularity. For example, the information type *health* could have two subcategories: mental health and physical health. A user might want to share *physical health* condition with *family* but *mental health* condition with both *family* and *friends*. Representing this sort of scenario is also quite feasible in our proposed technique.

7.6.1 Syntax and Semantics of the Models

Each of the models in a system consists of a set of control nodes otherwise known as states. In addition to these control states, a composed model uses integer variables, simple channels, and broadcast channels. The edges of the automata contain two types of labels: guards and synchronization. The guards express the conditions on the values of the integer variables. These conditions need to be satisfied in order for the edges to be taken for transitions. In our models (e.g., Figure 7.7), we add guards on

transitions to ensure the traversal of the paths that represent the desired information sharing activity of the user. We also add synchronization variables in the models which enable the communication between the behavioral model and the observer models. In Figure 7.2 and Figure 7.7, all the variables marked with an exclamation character "!" represent message transmission. Similarly, the observer models contain synchronization variables marked with a question mark "?" (Figure 7.3) that represent message reception. For example, whenever a transition happens from the state *Idle* to the state *s1*, on the behavioral model (Figure 7.2), it transmits a message *finance* which is then received by the "Information Type" observer model through the *finance?* path. The simple channels (e.g, finance!, finance? t_family!, r_friend!) help the observer models to keep track of the scenario factors and the broadcasting channel (i.e., *done!*) helps the observers to get back to the start position once an iteration (i.e., sharing activity) is completed. The model also consists of urgent (as soon as the transition is enabled, the current state will change to the next state) and committed locations. Since the information sharing behavior is assumed to be a non time-dependent process, we make the states urgent so that the transitions happen as soon as the flags are available. It's worth mentioning that, in time-dependent systems, the use of urgent locations reduces the complexity of the analysis by reducing the number of clocks.

7.7 Application and Usability

In this section, we briefly describe the application and usability of our proposed methodology. In the following section, we describe a process of creating the baseline model from the user's historical sharing activities.

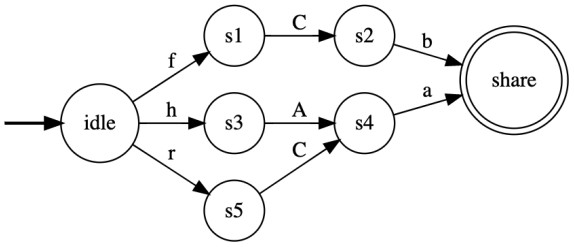


Figure 7.8: The DFA of the User 89

7.7.1 Automatic Translation of Activities to DFA

Formal modeling, formal verification and validation approaches are mostly used in the area of physical systems (e.g., industrial control systems, and cyber-physical systems). In this context, the process is mostly completed by human domain experts. Once a model is developed, validated, and verified; it is then used as the foundation of any downstream tasks such as hardware assembly, resiliency test, etc. In contrast, in the user-specific behavioral study, the formal modeling part has to be an automatic process that translated the user's desired privacy properties into formal specifications. This is because the end-users of an application will not have sufficient expertise to create the mathematically based model which is a core requirement to the model checking technique. Therefore, we utilize an existing tool [119] to automatically translate a user's historical sharing activities, manifested as regular expression, into deterministic finite automata (DFA).

First, we generate a regular expression string for every shared activity. We get three such strings— *rCa*, *hAa*, and *fCb* from the sharing activity (i.e., survey response) of the user 89 (Table 7.1). Where, h,f, and r represent the information types— $Health$, $Finance$, $Relationship$ respectively. A,B, C, and D represent the trust source— $Family$, $Friend$, $Expert$, and $SelfSearch$ respectively. a,b, c, and d

represent the recipient's role— $Family$, $Friend$, $Colleague$, and $OnlineService$ respectively. Then we combine the strings together with the regular expression's choice character + to get

$$(rC + hA)a + fCb$$

Finally, we use the tool [119] to generate a minimized DFA that accepts the regular expression (i.e., model the shared activities in terms of finite automata). Below is the formal Definition of the DFA:

$$\text{Set of state, } Q = \{idle, s_1, s_2, s_3, s_4, s_5, Share\}$$

$$\text{Alphabet, } \Sigma = \{h, f, r, A, B, C, D, a, b, c, d\}$$

$$\text{Initial state, } q_o = idle$$

$$\text{Set of final states, } F = \{share\}$$

$$\text{Transition function, } \delta = Q \times \Sigma \rightarrow Q$$

Figure 7.8 is the result of the translation which acts as the foundation of the UPPAAL model depicted in Figure 7.2. This preliminary automation step could be later taken over by another downstream automation tool (discussed later) to eventually develop a UPPAAL acceptable formal model.

7.7.2 Standalone Privacy Management Tool

Normally, the verification engine of UPPAAL is by default executed on the same computer as the user interface, but it can also run on a more powerful server which

allows hosting a complex behavioral model. Another supporting utility named *verifyta* is able to accept *.ta*, *.xta*, and *.xml* files as an input and use high-level programming language (e.g., Java)) API to perform the modeling, simulation, and model checking through a pragmatically native environment. This API makes it possible to interpret user's historical sharing activities and then develop a UPPAAL compatible formal model. The API could additionally be utilized to validate the developing model and verify it against a set of queries.

7.7.3 In Software Design and Development

One of the many advantages of formal modeling is its ability to allow for an early assessment of the model [173, 8]. In other words, it is possible to design, validate, and exhaustively verify user's privacy behavior model and think of it as the algorithm of his allowed behavior. Later on, programmers can leverage this model as the template for coding a function (e.g., shouldShare()) for that user in their software system. This process will enable the programmers to write a function which is already exhaustively tested, and therefore, no need to conduct typical unit testing on the program. An existing software can also integrate the privacy management tool by interacting with the verification engine through high level API. Thus the software can achieve a proper privacy management component inside its ecosystem.

7.7.4 User-Interface (UI) of the Privacy Settings

The user-interface of any software, mobile-app, or web-app plays an important role in providing its users with more flexible privacy settings. Users of the communication platforms are found to be less careful about properly setting their privacy preferences offered by the apps [105, 96, 97]. This is because of the generic and 'one fits all' nature of the privacy preference pages. Therefore, user-specific formal modeling can help with

the UI/UX designers and programmers are better equipped to provide personalized privacy settings pages to their users by utilizing their underlying behavioral model.

7.8 Conclusion

Users' ability to better manage their data-sharing practices is limited due to the lack of suitable user-centric privacy management tools and techniques. Moreover, very few of the existing methodologies take into consideration the aspect of personalization, correctness, and explainability. Most importantly, their practical usability and acceptance remain a significant challenge. In this paper we have presented an approach to formally model, validate, and verify personalized privacy disclosure behavior based on the analysis of user's situational decision-making process. The proposed methodology demonstrates a privacy formalism and verification technique based on UPPAAL which is a tool for modeling, validation, and verification of automata-based systems. Most importantly, the methodology depicts the potential of formalism towards the development of user-centric privacy management tools. In future work, we plan to extend the user's privacy behavior model to incorporate additional decision-making factors towards more granularity. We also plan to develop an end-to-end framework on top of UPPAAL to fully automate the process of transforming the historical sharing activities into a UPPAAL compatible network of automata.

CHAPTER 8:

LIMITATIONS

This dissertation has potential limitations that might have an impact on the results and findings. Therefore in this chapter, we identify and discuss the limitations of the methodologies and the research questions. In the subsequent chapter, we talk about the alternatives and potential approaches to minimize those limitations.

The work in chapter 3 is based on a survey that uses hypothetical scenarios. We acknowledge that those scenarios did not measure the actual disclosure behavior but rather users' *intention* to disclose their private information. This is certainly a limitation of our work, especially in light of the "privacy paradox" which shows a discrepancy between disclosure intentions and actual behaviors. In other words, the survey might not have captured the actual privacy behavior of the participants since they were asked to only share their intent, not actually share their private information. We also acknowledge that we only manipulate a few levels per factor in the study, and there could be much more granularity in the information type, recipient's role, and trust source factor. Moreover, in this quantitative analysis the situational factors are assumed to be the most influential ones while analyzing users' privacy decision-making process. However, there could be other factors with greater correlations with the disclosure intention. This is another limitation worth mentioning. Lastly, our analysis method (path analysis) which is often referred to as a causal inference

technique, reveals the predictive properties between the factors and constructs. These properties are measured in terms of path beta coefficients. Therefore, readers should be advised that this model only shows how the factors and constructs are correlated with the disclosure intention, not the actual causation.

The NLP models from chapter 4, 5, and 6 also have potential limitations in terms of assumptions, dataset, and evaluations. We collected the data from online platforms such as forums and social media. Users of these platforms might already have the mindset about the potential reach of their posts. Therefore, these texts might not represent all the different writing styles and ambiances that people follow while communicating through private mediums such as emails and text messages. Also, the human-annotated ground truth dataset from chapter 6 contains shorter texts (tweets) which might have an impact on the models' learning goals. In other words, the NLP models could have easily found the predictive ingredients in these short-length texts. Whereas, in the case of long-length texts (emails), a disclosure could happen at anywhere of the text. Therefore, our models may struggle to learn and predict from the whole body of the text. Most importantly, the diversity of the data source is subject to improvement.

Later in this dissertation (chapter 7), we represented the formal method based technique to develop the privacy verification engine, which also has notable limitations. Firstly, the approach depicts a scaled-down version of our proposed methodology. Therefore, the core design principle that relies on deterministic finite automata is not well evaluated. In other words, the potential state explosion problem is not investigated in this work. Additionally, since this work also uses the dataset from chapter 3, it inherits the limitations that are associated with the dataset. Most im-

portantly, the assumption of "human as a finite being" is worth mentioning, which we believe is another limitation of the proposed methodology.

CHAPTER 9:

CONCLUSION AND FUTURE WORK

In this chapter, we discuss the implications and summary of the results from each of the earlier chapters of this dissertation. Additionally, we discuss the future work directions.

9.1 Summary and Conclusion

The research works in this dissertation focus on three main objectives; first we analyze and model users' situational privacy decision-making process based on survey data, then we develop and evaluate several deep-learning models that are capable of identifying privacy disclosure occurrences in human-annotated text data, finally we propose and demonstrate a methodology to formally model, validate, and verify personalized privacy disclosure behavior based on the analysis of the users' situational decision-making process.

The work in chapter 3 presented the results of a scenario-based survey to understand users' *situational* privacy perceptions and disclosure intentions. Through path analysis, the work shows how users make privacy decisions in various situations. It also reveals how the situational factors have significant effects on users' perceptions of privacy factors, which in turn have an effect on their intention to disclose their private information. A few notable findings include but not limited to - i) people

are estimated to perceive a higher level of behavioral control when the recipient is a family member, a friend, or a colleague than when the recipient is online platforms, ii) people are estimated to have a more positive attitude toward disclosure when the recipient is a family member or a friend than the online platform, iii) people believe that individuals close to them would be most likely to agree with the scenario when it involves health information, followed by relationship information, and finally financial information. These results constitute a contextualized understanding of users' privacy behaviors, connected to the Theory of Planned Behavior. These results also provide new insights that can help build future user-tailored privacy models. As a matter of fact, these findings and observations are the basis of the privacy verification engine which we presented in chapter 7.

The work in chapter 7 presented a practical approach to formally model, validate, and verify personalized privacy disclosure behavior based on the dataset from chapter 3. It demonstrated a privacy formalism and verification technique based on UPPAAL, a model checking tool, and depicts the potential of formalism towards the development of personalized privacy management tools. A scaled-down version of the proposed methodology shows the process of modeling individual's privacy disclosure behavior where their disclosure decision relies on various situational factors. The proposed methodology also shows how users' privacy behavioral model could be represented as a network of automata which in turn acts as a privacy verification engine to govern the flow of information. The work also describes various ways of implementing the verification engine for practical usage and explains its correctness, explainability, usability, and acceptance. In practice, the verification engine along with a text analysis tool can warn the users while they attempt to share private in-

formation in the form of texts. Consequently, this dissertation also contains relevant research works that represent the novel architectures of multiple NLP models.

In chapters 4, 5, and 6, we represented the neural network based architectures of three different NLP models and their evaluation reports on a ground-truth dataset. The models developed on these architectures can precisely recognize privacy disclosures through text by utilizing state-of-the-art semantic and syntactic analysis, the hidden pattern of sentence structure, tone of the author, and metadata from the content. They show that, despite a lack of large amounts of labeled data, we can train neural network based models that go beyond simple keyword spotting and use linguistic features to determine if a text contains a disclosure or not with a useful degree of accuracy. Unlike the traditional text classification techniques that primarily rely on keyword spotting, the model in chapter 6, for instance, focus on underlying meaning and hidden patterns by combining pre-trained language model and classical linguistics.

Altogether, this dissertation contributes several insights to the area of user-tailored privacy modeling and personalized privacy systems through extensive amounts of experimental results. The dissertation also lays the foundation for developing a new set of verification tools, algorithms, and interfaces that enable secure, effective, and unobtrusive management of users' private information.

9.2 Future Work

The research presented in this dissertation has the potential to lead to new research directions. Also, addressing the limitations discussed in chapter 8 could help improve the methodologies presented in the dissertation. Currently, the work in chapter 3 relies on users' intended behavior based on hypothetical scenarios, failing to capture

the actual privacy disclosure behavior. Future work can bridge this gap between intention and behavior by incorporating reported or actual behavior in the model. This can be done by developing and deploying a communication app for a certain group of participants and observing the sharing behavior. Also, the sample size of the participants and their diversity could also be increased for more generalized insights. Likewise, the number of manipulating factors and their levels could be increased to more granularity. Additionally, the predictive power of the current path analysis model can be investigated by surveying a new sample of users. Most importantly, the quantitative research from chapter 3 can lead to a qualitative study and modeling of privacy disclosure behavior.

Likewise, the users' privacy behavior model presented in chapter 7 can be extended by incorporating additional decision-making factors, constructs, and constraints resulting in a robust privacy verification engine. Currently, the proposed methodology contains manual translation of users' historical privacy preferences into a form of deterministic finite automata. Hence, there are scopes to fully automate the process of transforming users' information sharing activities into network of deterministic finite automata. An interesting extension of the proposed formalism technique could be using probabilistic finite state machines, which in a sense lie between deterministic and non-deterministic state machines. Similarly, there are improvement scopes in the works presented in chapters 4, 5, and 6. The learning and predictive performance of the NLP models can be evaluated on a diverse dataset by taking samples from different data sources. In addition, introducing model explainability, performing privacy-preserving text analysis, testing the integration with end products could also be future works.

Most importantly, an immediate future direction of the research could be developing a privacy management software by integrating all the components (e.g., NLP models, verification engine) and deploying it to a group of participants. Later, the users' interactions with the privacy framework in terms of accepting or rejecting the warnings could be analyzed to evaluate the practicality and usability of such privacy tools.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Daniel Abril, Guillermo Navarro-Arribas, and Vicenç Torra. On the declassification of confidential documents. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 235–246. Springer, 2011.

[3] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 1–8, 1999.

[4] Idris Adjerid, Eyal Peer, and Alessandro Acquisti. Beyond the privacy paradox:

Objective versus relative risk in privacy decision making. *Available at SSRN 2765097*, 2016.

[5] Walid A Afifi and Laura K Guerrero. Motivations underlying topic avoidance in close relationships. *Balancing the secrets of private disclosures*, pages 165–180, 2000.

[6] Rodrigo Agerri, Xabier Artola, Zuhaitz Beloki, German Rigau, and Aitor Soroa. Big data for natural language processing: a streaming approach. *Knowledge-Based Systems*, 79:36–42, 2015.

[7] Icek Ajzen et al. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

[8] Vangalur S Alagar and Kasilingam Periyasamy. *Specification of software systems*. Springer Science & Business Media, 2011.

[9] Reza Ghaiumy Anaraky, Bart P Knijnenburg, and Marten Risius. Exacerbating mindless compliance: The danger of justifications during privacy decision making in the context of facebook applications. *AIS Transactions on Human-Computer Interaction*, 12(2):70–95, 2020.

[10] Paul Ashley, Satoshi Hada, Günter Karjoth, Calvin Powers, and Matthias Schunter. Enterprise privacy authorization language (epal). *IBM Research*, 30:31, 2003.

[11] Guillaume Aucher, Guido Boella, and Leendert Van Der Torre. A dynamic logic for privacy compliance. *Artificial Intelligence and Law*, 19(2-3):187, 2011.

[12] Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT press, 2008.

[13] JinYeong Bak, Suin Kim, and Alice Oh. Self-disclosure and relationship strength in twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 60–64, 2012.

[14] JinYeong Bak, Chin-Yew Lin, and Alice Oh. Self-disclosure topic model for classifying and analyzing twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996, 2014.

[15] Kheana Barbeau, Kayla Boileau, Fatima Sarr, and Kevin Smith. Path analysis in mplus: A tutorial using a conceptual model of psychological and behavioral antecedents of bulimic symptoms in young adults. *The Quantitative Methods for Psychology*, 15(1):38–53, 2019.

[16] Adam Barth, Anupam Datta, John C Mitchell, and Helen Nissenbaum. Privacy and contextual integrity: Framework and applications. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, pages 15–pp. IEEE, 2006.

[17] Roy F Baumeister and Kenneth J Cairns. Repression and self-presentation: When audiences interfere with self-deceptive strategies. *Journal of Personality and Social Psychology*, 62(5):851, 1992.

[18] Lisa Beck and Icek Ajzen. Predicting dishonest actions using the theory of planned behavior. *Journal of research in personality*, 25(3):285–301, 1991.

[19] Gerd Behrmann, Alexandre David, and Kim G Larsen. A tutorial on uppaal 4.0. *Department of computer science, Aalborg university*, 2006.

[20] Zakariya Belkhamza, Mohd Niasin, and Adzwin Faris. The effect of privacy concerns on smartphone app purchase in malaysia: Extending the theory of planned behavior. *International Journal of Interactive Mobile Technologies*, 11(5), 2017.

[21] Matthew L Bolton, Noelia Jiménez, Marinus M van Paassen, and Maite Trujillo. Automatically generating specification properties from task models for the formal verification of human–automation interaction. *IEEE Transactions on Human-Machine Systems*, 44(5):561–575, 2014.

[22] Paula M Brauer, Rhona M Hanning, Jose F Arocha, Dawna Royall, Richard Goy, Andrew Grant, Linda Dietrich, Roselle Martino, and Julie Horrocks. Creating case scenarios or vignettes using factorial study design methods. *Journal of advanced nursing*, 65(9):1937–1945, 2009.

[23] Travis D Breaux, Hanan Hibshi, and Ashwini Rao. Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. *Requirements Engineering*, 19(3):281–307, 2014.

[24] Tom Buchanan, Carina Paine, Adam N Joinson, and Ulf-Dietrich Reips. Development of measures of online privacy concern and protection for use on the internet. *Journal of the Association for Information Science and Technology*, 58(2):157–165, 2007.

[25] Sarah Burns and Lynne Roberts. Applying the theory of planned behaviour to

predicting online safety behaviour. *Crime Prevention and Community Safety*, 15(1):48–64, 2013.

[26] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 35–46. ACM, 2014.

[27] Venkatesan T Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania. Efficient techniques for document sanitization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 843–852. ACM, 2008.

[28] Gerald Cheshun Chao. Adaptive and scalable method for resolving natural language ambiguities, January 6 2009. US Patent 7,475,010.

[29] Dongjin Choi, Jeongin Kim, Xeufeng Piao, and Pankoo Kim. Text analysis for monitoring personal information leakage on twitter. *J. UCS*, 19(16):2472–2485, 2013.

[30] Jinho D Choi, Joel Tetreault, and Amanda Stent. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396, 2015.

[31] Richard Chow, Philippe Golle, and Jessica Staddon. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD*

international conference on Knowledge discovery and data mining, pages 893–901. ACM, 2008.

[32] Emily Christofides, Amy Muise, and Serge Desmarais. Information disclosure and control on facebook: Are they two sides of the same coin or two different processes? *Cyberpsychology & behavior*, 12(3):341–345, 2009.

[33] Edmund M Clarke and Jeannette M Wing. Formal methods: State of the art and future directions. *ACM Computing Surveys (CSUR)*, 28(4):626–643, 1996.

[34] Mark Conner, Sara FL Kirk, Janet E Cade, and Jennifer H Barrett. Environmental influences: factors influencing a woman's decision to use dietary supplements. *The Journal of nutrition*, 133(6):1978S–1982S, 2003.

[35] Wikipedia contributors. Word embedding — Wikipedia, the free encyclopedia, 2018. [Online; accessed 7-May-2018].

[36] Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry Den Hartog. A machine learning solution to assess privacy policy completeness: (short paper). In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, pages 91–96, 2012.

[37] Paul C Cozby. Self-disclosure: a literature review. *Psychological bulletin*, 79(2):73, 1973.

[38] Lorrie Cranor. *Web privacy with P3P*. " O'Reilly Media, Inc.", 2002.

[39] Tanvi Dadu, Kartikey Pant, and Radhika Mamidi. Bert-based ensembles for modeling disclosure and support in conversational social media text. *arXiv preprint arXiv:2006.01222*, 2020.

[40] Datafiniti. Hotel Reviews — Kaggle. `https://www.kaggle.com/datafiniti/hotel-reviews`, 2018. [Online; accessed 01-May-2018].

[41] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[42] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*, 2014.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[44] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[45] Tobias Dienlin and Sabine Trepte. Is the privacy paradox a relic of the past? an in-depth analysis of privacy attitudes and privacy behaviors. *European journal of social psychology*, 45(3):285–297, 2015.

[46] Tamara Dinev and Paul Hart. An extended privacy calculus model for e-commerce transactions. *Information systems research*, 17(1):61–80, 2006.

[47] G Eleftherakis and P Kefalas. Towards model checking of finite state machines extended with memory through refinement. *Advances in signal processing and computer technologies*, pages 321–326, 2001.

[48] Pardis Emami Naeini, Martin Degeling, Lujo Bauer, Richard Chow, Lorrie Faith Cranor, Mohammad Reza Haghighat, and Heather Patterson. The influence of friends and experts on privacy decision making in iot scenarios. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–26, 2018.

[49] David A Evans and Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 17–24. Association for Computational Linguistics, 1996.

[50] Stack Exchange. Stack Exchange Data Dump : Stack Exchange, Inc. : Free Download, Borrow, and Streaming : Internet Archive. `https://archive.org/details/stackexchange`, 2018. [Online; accessed 01-May-2018].

[51] Kavita Ganesan and ChengXiang Zhai. Opinion-based entity ranking. *Information retrieval*, 15(2):116–150, 2012.

[52] Google. Colaboratory - Google Research. [Online; accessed 01-May-2021].

[53] Kasthurirangan Gopalakrishnan, Siddhartha K Khaitan, Alok Choudhary, and Ankit Agrawal. Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. *Construction and Building Materials*, 157:322–330, 2017.

[54] Paul Grace and Mike Surridge. Towards a model of user-centered privacy preservation. In *Proceedings of the 12th International Conference on Availability, Reliability and Security*, pages 1–8, 2017.

[55] Jamal Greene. The so-called right to privacy. *UC Davis L. Rev.*, 43:715, 2009.

[56] Tina Groves. Why is analyzing text so hard? `http://www.ibmbigdatahub.com/blog/why-analyzing-text-so-hard`, 2018. [Online; accessed 01-February-2018].

[57] Orna Grumberg, Doron A Peled, and EM Clarke. Model checking, 1999.

[58] Jerold L Hale, Brian J Householder, and Kathryn L Greene. The theory of reasoned action. *The persuasion handbook: Developments in theory and practice*, 14:259–286, 2002.

[59] Michael Hart, Pratyusa Manadhata, and Rob Johnson. Text classification for data loss prevention. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 18–37. Springer, 2011.

[60] Wannes Heirman, Michel Walrave, and Koen Ponnet. Predicting adolescents' disclosure of personal information in exchange for commercial incentives: An application of an extended theory of planned behavior. *Cyberpsychology, Behavior, and Social Networking*, 16(2):81–87, 2013.

[61] Alex Hern. Far more than 87m Facebook users had data compromised, MPs told. `https://www.theguardian.com/uk-news/2018/apr/17/facebook-users-data-compromised-far-more-than-87m-mps-told/-cambridge-analytica`, 2018. [Online; accessed 01-May-2018].

[62] Shirley S Ho, May O Lwin, Andrew ZH Yee, and Edmund WJ Lee. Understanding factors associated with singaporean adolescents' intention to adopt privacy protection behavior using an extended theory of planned behavior. *Cyberpsychology, Behavior, and Social Networking*, 20(9):572–579, 2017.

[63] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[64] Han Hu, NhatHai Phan, Soon A Chun, James Geller, Huy Vo, Xinyue Ye, Ruoming Jin, Kele Ding, Deric Kenne, and Dejing Dou. An insight analysis and detection of drug-abuse risk behavior on twitter with self-taught deep learning. *Computational Social Networks*, 6(1):1–19, 2019.

[65] Huggingface. Huggingface Transformers. Online; accessed 01-May-2021].

[66] IBM. IBM Watson - Tone Analyzer. https://www.ibm.com/watson/services/tone-analyzer/, 2019. [Online; accessed 01-December-2019].

[67] Prateek Jindal, Carl A Gunter, and Dan Roth. Detecting privacy-sensitive events in medical text. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 617–620, 2014.

[68] Leslie K John, Alessandro Acquisti, and George Loewenstein. Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of consumer research*, 37(5):858–873, 2011.

[69] Adam N Joinson, Ulf-Dietrich Reips, Tom Buchanan, and Carina B Paine Schofield. Privacy, trust, and self-disclosure online. *Human–Computer Interaction*, 25(1):1–24, 2010.

[70] Rezvan Joshaghani, Stacy Black, Elena Sherman, and Hoda Mehrpouyan. Formal specification and verification of user-centric privacy policies for ubiquitous systems. In *Proceedings of the 23rd International Database Applications & Engineering Symposium*, pages 1–10, 2019.

[71] Rezvan Joshaghani and Hoda Mehrpouyan. A model-checking approach for enforcing purpose-based privacy policies. In *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, pages 178–179. IEEE, 2017.

[72] Keras. Dense Layres - Keras Documentation. Online; accessed 01-May-2021.

[73] Keras. Embedding Layres - Keras Documentation. Online; accessed 01-May-2021.

[74] Keras. Guide to the Functional API - Keras Documentation. Online; accessed 01-May-2021].

[75] Keras. LSTM layer - Keras Documentation. Online; accessed 01-May-2021.

[76] Keras. Text Preprocessing - Keras Documentation. [Online; accessed 01-May-2021].

[77] Keras. Convolutional Layres - Keras Documentation. `https://keras.io/layers/convolutional/`, 2018. [Online; accessed 01-February-2018].

[78] Keras. Text Preprocessing - Keras Documentation. `https://keras.io/preprocessing/text/#tokenizer`, 2018. [Online; accessed 01-February-2018].

[79] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[80] Shulamit Sara Klinger. *" Are they talking yet?": online discourse as political action in an education policy forum.* PhD thesis, University of British Columbia, 2002.

[81] Bart P Knijnenburg. Privacy? i can't even! making a case for user-tailored privacy. *IEEE Security & Privacy*, 15(4):62–67, 2017.

[82] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):441–504, 2012.

[83] Bart Piet Knijnenburg and Alfred Kobsa. Increasing sharing tendency without reducing satisfaction: finding the best privacy-settings user interface for social networks. 2014.

[84] Bart Piet Knijnenburg, Alfred Kobsa, and Hongxia Jin. Counteracting the negative effect of form auto-completion on the privacy calculus. 2013.

[85] Soonho Kong, Sicun Gao, Wei Chen, and Edmund Clarke. dreach: δ-reachability analysis for hybrid systems. In *International Conference on TOOLS and Algorithms for the Construction and Analysis of Systems*, pages 200–205. Springer, 2015.

[86] Moshe Kravchik and Asaf Shabtai. Anomaly detection; industrial control systems; convolutional neural networks. *arXiv preprint arXiv:1806.08110*, 2018.

[87] Balachander Krishnamurthy and Craig E Wills. On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 7–12. ACM, 2009.

[88] Padmanabhan Krishnan and Kostyantyn Vorobyov. Enforcement of privacy requirements. In *IFIP International Information Security Conference*, pages 272–285. Springer, 2013.

[89] Priya Kumar, Shalmali Milind Naik, Utkarsha Ramesh Devkar, Marshini Chetty, Tamara L Clegg, and Jessica Vitak. 'no telling passcodes out because they're private' understanding children's mental models of privacy and security online. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–21, 2017.

[90] O Rivera Kurkovsky, Oscar Rivera, and Jay Bhalodi. Classification of privacy management techniques in pervasive computing. *International Journal of u-and e-Service, Science and Technology*, 11(1):55–71, 2007.

[91] Kim G Larsen, Paul Pettersson, and Wang Yi. Uppaal in a nutshell. *International journal on software tools for technology transfer*, 1(1-2):134–152, 1997.

[92] Robert S Laufer and Maxine Wolfe. Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of social Issues*, 33(3):22–42, 1977.

[93] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[94] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[95] Scott Lederer, Jennifer Mankoff, and Anind K Dey. Who wants to know what when? privacy preference determinants in ubiquitous computing. In *CHI'03*

extended abstracts on Human factors in computing systems, pages 724–725, 2003.

[96] Heather Richter Lipford, Andrew Besmer, and Jason Watson. Understanding privacy settings in facebook with an audience view. *UPSEC*, 8:1–8, 2008.

[97] Yabing Liu, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. Analyzing facebook privacy settings: user expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 61–70, 2011.

[98] Jiajun Lu, Zhiqiu Huang, and Changbo Ke. Verification of behavior-aware privacy requirements in web services composition. *JSW*, 9(4):944–951, 2014.

[99] May O Lwin and Jerome D Williams. A model integrating the multidimensional developmental theory of privacy and theory of planned behavior to examine fabrication of information online. *Marketing Letters*, 14(4):257–272, 2003.

[100] Octav-Ionuţ Macovei. Determinants of consumers' pro-environmental behavior–toward an integrated model. *Journal of Danubian Studies and Research*, 5(2), 2015.

[101] Mary Madden. Privacy management on social media sites. *Pew Internet Report*, pages 1–20, 2012.

[102] Mary Madden, Amanda Lenhart, Sandra Cortesi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. Teens, social media, and privacy. *Pew Research Center*, 21:2–86, 2013.

[103] Naresh K Malhotra, Sung S Kim, and James Agarwal. Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information systems research*, 15(4):336–355, 2004.

[104] Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12. ACM, 2011.

[105] Alessandra Mazzia, Kristen LeFevre, and Eytan Adar. The pviz comprehension tool for social network privacy settings. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, pages 1–12, 2012.

[106] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908. ACM, 2013.

[107] AK Mehdy, Michael D Ekstrand, Bart P Knijnenburg, and Hoda Mehrpouyan. Privacy as a planned behavior: Effects of situational factors on privacy perceptions and plans. *UMAP'21, June 21–25, 2021, Utrecht, Netherlands© 2021 Association for Computing Machinery.*, 2021.

[108] AKM Nuhil Mehdy and Hoda Mehrpouyan. A user-centric and sentiment aware privacy-disclosure detection framework based on multi-input neural network. In *PrivateNLP@ WSDM*, pages 21–26, 2020.

[109] Nuhil Mehdy, Casey Kennington, and Hoda Mehrpouyan. Privacy disclosures detection in natural-language text through linguistically-motivated artificial

neural network. In *2nd EAI International Conference on Security and Privacy in New Computing Environments*. EAI, 2019.

[110] Nuhil Mehdy, Casey Kennington, and Hoda Mehrpouyan. Privacy disclosures detection in natural-language text through linguistically-motivated artificial neural networks. In *International Conference on Security and Privacy in New Computing Environments*, pages 152–177. Springer, 2019.

[111] Hoda Mehrpouyan, Ion Madrazo Azpiazu, and Maria Soledad Pera. Measuring personality for automatic elicitation of privacy preferences. In *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, pages 84–95. IEEE, 2017.

[112] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[113] Sandra J Milberg, Sandra J Burke, H Jeff Smith, and Ernest A Kallman. Values, personal information privacy, and regulatory approaches. *Communications of the ACM*, 38(12):65–74, 1995.

[114] David N Milne, Glen Pink, Ben Hachey, and Rafael A Calvo. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 118–127, 2016.

[115] Eni Mustafaraj and Panagiotis Takis Metaxas. What edited retweets reveal about online political discourse. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[116] Linda K Muthén and Bengt O Muthén. Mplus user's guide (version 7). *Los Angeles, CA: Author*, 1998.

[117] Melanie Nguyen, Yu Sun Bin, and Andrew Campbell. Comparing online and offline self-disclosure: A systematic review. *Cyberpsychology, Behavior, and Social Networking*, 15(2):103–111, 2012.

[118] Helen Nissenbaum. Privacy as contextual integrity. *Wash. L. Rev.*, 79:119, 2004.

[119] Noam. Noam is a JavaScript library for working with automata and formal grammars for regular and context-free languages. https://github.com/izuzak/noam, 2015. [Online; accessed 10-May-2021].

[120] Judith S Olson, Jonathan Grudin, and Eric Horvitz. A study of preferences for sharing and privacy. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1985–1988, 2005.

[121] Sylvia Osborn, Ravi Sandhu, and Qamar Munawer. Configuring role-based access control to enforce mandatory and discretionary access control policies. *ACM Transactions on Information and System Security (TISSEC)*, 3(2):85–106, 2000.

[122] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, 2010.

[123] Sameer Patil and Jennifer Lai. Who gets to know what when: configuring privacy permissions in an awareness application. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 101–110, 2005.

[124] Theodore Patkos, Giorgos Flouris, Panagiotis Papadakos, Antonis Bikakis, Pompeu Casanovas, Jorge González-Conejero, Rebeca Varela Figueroa, Anthony Hunter, Gujón Idir, George Ioannidis, et al. Privacy-by-norms privacy expectations in online interactions. In *2015 IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshops*, pages 1–6. IEEE, 2015.

[125] Elazar J Pedhazur. *Multiple regression in behavioral research: Explanation and prediction*. Wadsworth Publishing Company, 1997.

[126] Sandra Petronio. Communication privacy management theory. *The international encyclopedia of interpersonal communication*, pages 1–9, 2015.

[127] Joseph Phelps, Glen Nowak, and Elizabeth Ferrell. Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy & Marketing*, 19(1):27–41, 2000.

[128] Amir H Razavi and Kambiz Ghazinour. Personal health information detection in unstructured web documents. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pages 155–160. IEEE, 2013.

[129] recognai. spaCy WordNet. Online; accessed 01-May-2021].

[130] Alexander Rosenberg. Privacy as a matter of taste and right. *Social Philosophy and Policy*, 17(2):68–90, 2000.

[131] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, 2019.

[132] Jacqueline Strunk Sachs. Recopition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, 2(9):437–442, 1967.

[133] Alexander K Saeri, Claudette Ogilvie, Stephen T La Macchia, Joanne R Smith, and Winnifred R Louis. Predicting facebook users' online privacy protection: Risk, trust, norm focus theory, and the theory of planned behavior. *The Journal of social psychology*, 154(4):352–369, 2014.

[134] David Sánchez, Montserrat Batet, and Alexandre Viejo. Detecting sensitive information from textual documents: an information-theoretic approach. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 173–184. Springer, 2012.

[135] Ashley Savage and Richard Hyde. Using freedom of information requests to facilitate research. *International Journal of Social Research Methodology*, 17(3):303–317, 2014.

[136] Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583, 2015.

[137] Rinal B Shah. A multivariate analysis technique: Structural equation modeling. *Asian Journal of Multidimensional Research (AJMR)*, 1(4):73–81, 2012.

[138] Hai-bo Shen and Fan Hong. An attribute-based access control model for web services. In *2006 Seventh International Conference on Parallel and Distributed*

Computing, Applications and Technologies (PDCAT'06), pages 74–79. IEEE, 2006.

[139] Arnon Siegel. In pursuit of privacy: Laws, ethics, and the rise of technology. *The Wilson Quarterly*, 21(4):100, 1997.

[140] Herbert A Simon. Theories of bounded rationality. *Decision and organization*, 1(1):161–176, 1972.

[141] Itamar Simonson and Amos Tversky. Choice in context: Tradeoff contrast and extremeness aversion. *Journal of marketing research*, 29(3):281–295, 1992.

[142] Jayveer Singh and Manisha J Nene. A survey on machine learning techniques for intrusion detection systems. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(11):4349–4355, 2013.

[143] Olivia Solon. Facebook says Cambridge Analytica may have gained 37m more users' data. `https://www.theguardian.com/technology/2018/apr/04/facebook-cambridge-analytica-user-data-latest-more-than-thought`, 2018. [Online; accessed 01-May-2018].

[144] Spacy. Linguistic Features - Named Entities. `https://spacy.io/usage/linguistic-features#section-named-entities`, 2018. [Online; accessed 01-February-2018].

[145] Spacy. Named Entity Recognition. `https://prodi.gy/features/named-entity-recognition`, 2018. [Online; accessed 01-February-2018].

[146] Statista. Number of U.S. data breaches 2014-2018, by industry. https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/, 2019. [Online; accessed 01-April-2019].

[147] David L Streiner. Finding our way: an introduction to path analysis. *The Canadian Journal of Psychiatry*, 50(2):115–122, 2005.

[148] Latanya Sweeney. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association, 1996.

[149] Symantec. 10 cyber security facts and statistics for 2018, 2019. [Online; accessed 01-April-2019].

[150] Symantec. 10 cyber security facts and statistics for 2018. https://us.norton.com/internetsecurity-emerging-threats-10-facts-about-todays-cybersecurity-landscape-that-you-should-know.html, 2019. [Online; accessed 01-April-2019].

[151] Herman T Tavani. Philosophical theories of privacy: Implications for an adequate online privacy policy. *Metaphilosophy*, 38(1):1–22, 2007.

[152] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In *Companion Proceedings of the The Web Conference 2018*, pages 163–166, 2018.

[153] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[154] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

[155] Twitter. Tweets Search API Reference. Online; accessed 01-May-2021].

[156] Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. Detection and analysis of self-disclosure in online news commentaries. In *The World Wide Web Conference*, pages 3272–3278, 2019.

[157] Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. A study of self-privacy violations in online public discourse. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1041–1050. IEEE, 2020.

[158] Robert J Vallerand, Paul Deshaies, Jean-Pierre Cuerrier, Luc G Pelletier, and Claude Mongeau. Ajzen and fishbein's theory of reasoned action as applied to moral behavior: A confirmatory analysis. *Journal of personality and social psychology*, 62(1):98, 1992.

[159] Paul Van Schaik. Involving users in the specification of functionality using scenarios and model-based evaluation. *Behaviour & Information Technology*, 18(6):455–466, 1999.

[160] Asimina Vasalou, Alastair J Gill, Fadhila Mazanderani, Chrysanthi Papoutsi, and Adam Joinson. Privacy dictionary: A new resource for the automated

content analysis of privacy. *Journal of the Association for Information Science and Technology*, 62(11):2095–2105, 2011.

[161] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[162] Ding Wang, Zijian Zhang, Ping Wang, Jeff Yan, and Xinyi Huang. Targeted online password guessing: An underestimated threat. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 1242–1254. ACM, 2016.

[163] Samuel Warren et al. Louis brandeis. the right to privacy. *Harvard Law Review*, 4(5):1, 1890.

[164] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

[165] Ryan West, Christopher Mayhorn, Jefferson Hardee, and Jeremy Mendel. The weakest link: A psychological perspective on why users make poor security decisions. In *Social and Human elements of information security: Emerging Trends and countermeasures*, pages 43–60. IGI Global, 2009.

[166] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

[167] Tiffany Barnett White. Consumer disclosure and disclosure avoidance: A motivational framework. *Journal of consumer psychology*, 14(1 & 2):41–51, 2004.

[168] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[169] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[170] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 229–240, 2006.

[171] Christopher C Yang and Xuning Tang. Estimating user influence in the medhelp social network. *IEEE Intelligent Systems*, 27(5):44–50, 2012.

[172] Mike Z Yao and Daniel G Linz. Predicting self-protections of online privacy. *CyberPsychology & Behavior*, 11(5):615–617, 2008.

[173] Junbeom Yoo, Eunkyoung Jee, and Sungdeok Cha. Formal modeling and verification of safety-critical software. *IEEE software*, 26(3):42–49, 2009.

[174] Jun Zhao, Ge Wang, Carys Dally, Petr Slovak, Julian Edbrooke-Childs, Max Van Kleek, and Nigel Shadbolt. I make up a silly name' understanding children's perception of privacy risks online. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

APPENDIX A:

CHAPTER 3

A.1 Model Fitness

Table A.1: R^2 values of the fitted model.

Observed Variable	Estimate	S.E.	Est./S.E.	P-Value
Disclosure Intention	0.384	0.022	17.454	0.000
Attitude	0.352	0.029	12.101	0.000
Subjective Norm	0.065	0.012	5.602	0.000
Perceived Behavioral Control	0.401	0.032	12.472	0.000

A.2 Path Analysis Output

Table A.2: Output of the model's effects.

Constructs	Estimate	S.E.	Est./S.E.	P-Value
Attitude				
Recipient's Role (Baseline: Online)				
Colleague	-0.113	0.043	-2.645	0.008
Family	0.201	0.051	3.928	0.000
Friend	0.138	0.046	3.022	0.003
Subjective Norm	0.301	0.039	7.819	0.000
Perceived Behavioral Control	0.248	0.040	6.181	0.000
General Attitude	0.320	0.026	12.372	0.000
Subjective Norm				
Information Type (Baseline: Relationship)				
Finance	-0.131	0.048	-2.740	0.006
Health	0.202	0.047	4.338	0.000
Recipient's Role (Baseline: Online)				
Colleague	-0.109	0.049	-2.243	0.025
Family	0.327	0.048	6.793	0.000
Friend	0.256	0.046	5.587	0.000
General Attitude	-0.119	0.041	-2.921	0.003
Perceived Behavioral Control				
Recipient's Role (Baseline: Online)				
Colleague	0.201	0.037	5.409	0.000
Family	0.298	0.045	6.659	0.000
Friend	0.257	0.043	5.972	0.000
Subjective Norm	0.610	0.026	23.264	0.000
Disclosure Intention				
Attitude	0.151	0.011	13.780	0.000
Subjective Norm	0.111	0.013	8.822	0.000
Perceived Behavioral Control	0.108	0.012	9.350	0.000

APPENDIX B:

CHAPTER 5

B.1 Model Hyperparameters

Some hyperparamters worth mentioning are: pre-trained embedding with glove 100 dimensional embedding matrix having the capability of adjusting weights through the training iteration. Convolution with 32 filters with kernel size of 4. These layers have rectifier linear unit as activation function and followed by global max pooling technique. The LSTM layer contains 32 neurons with all the default settings as per the keras documentation. The first stage of dense layers after the first concatenation contains 128 and 64 neurons with rectifier linear unit as activation function. The second stage of dense layers contains 64, 32, and 16 neurons with same kind of activation function following a single output neuron with sigmoid as activation function. We train the model for 20 epochs providing the batch size of 32. The model also uses binary cross entropy as the loss function and rmsprop as the optimizer.

B.2 Neural Network Architecture

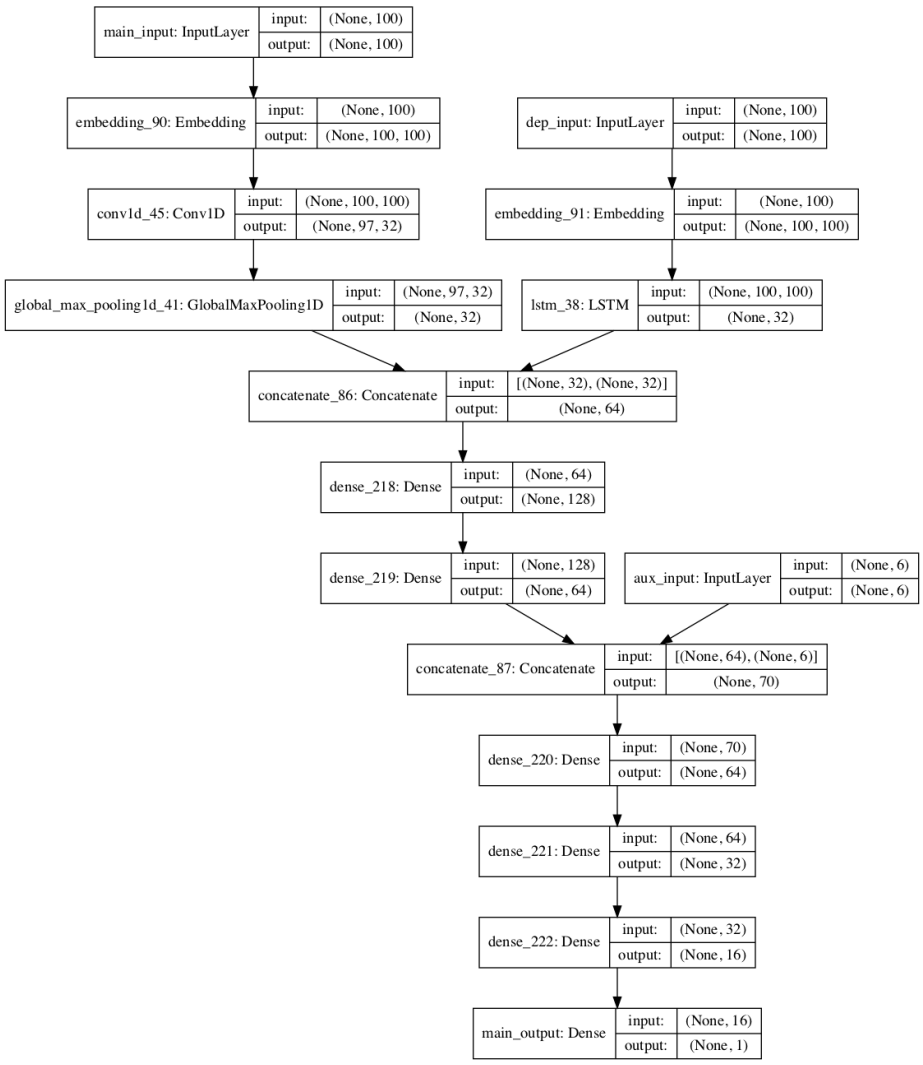


Figure B.1: Architecture of the Neural Network (Automatically Rendered by the Keras Plotter).

APPENDIX C:

CHAPTER 7

C.1 Survey Interface 1

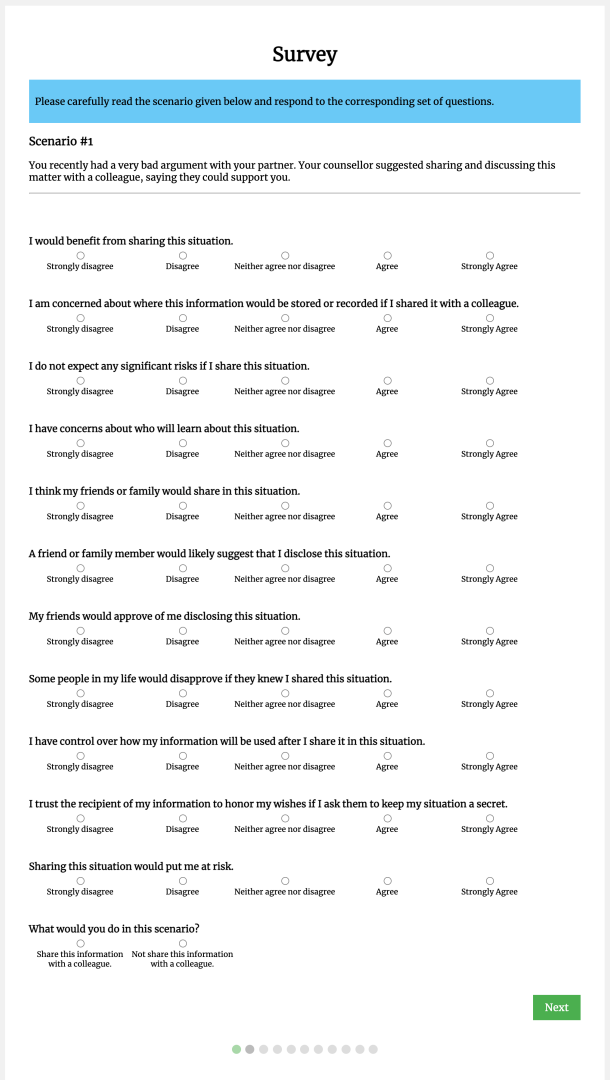


Figure C.1: Screenshot of the Survey System Representing 1 of 8 Random Scenarios Given to a Participant.

C.2 Survey Interface 2

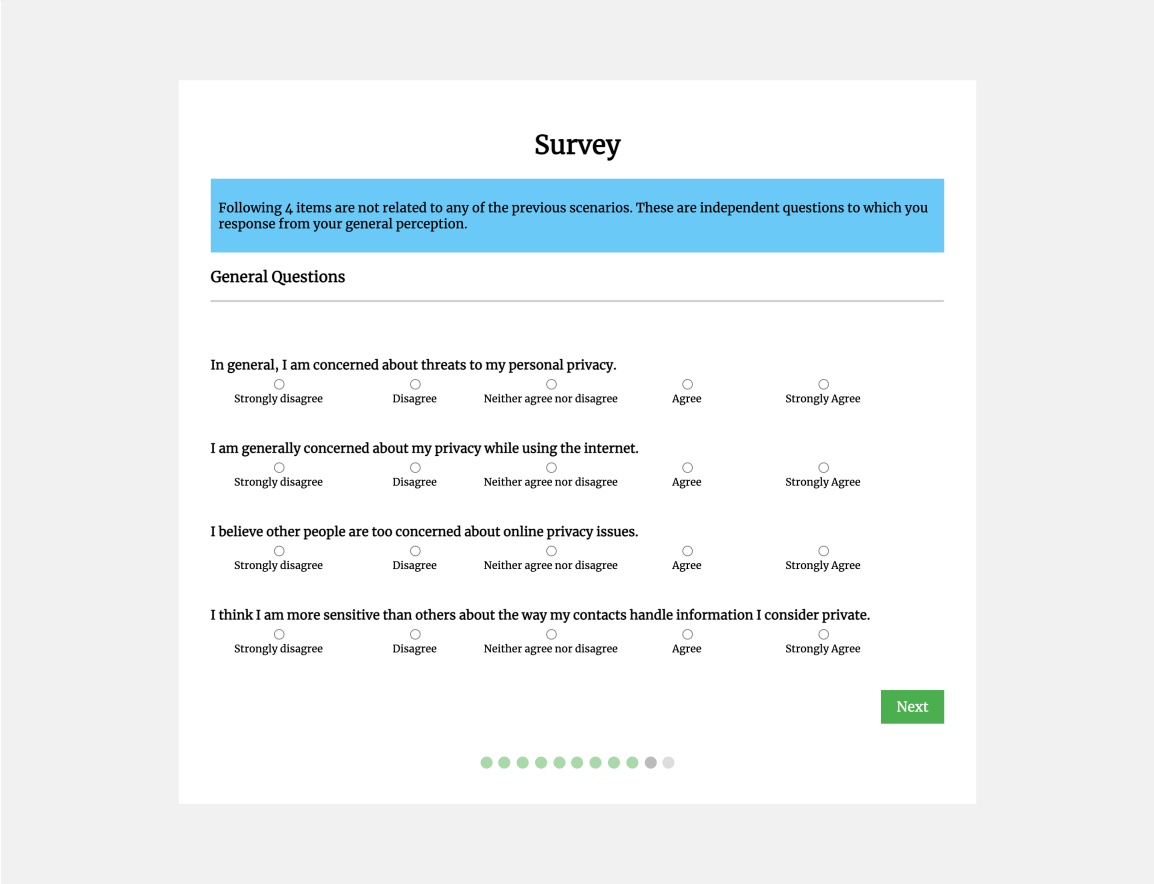


Figure C.2: Screenshot of the survey system representing the general attitude questions given to a participant at the end of the survey.