Measuring and Enhancing Users' Privacy in Machine Learning

by

Walaa Alnasser

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2022 by the
Graduate Supervisory Committee:

Huan Liu, Chair
Hasan Davulcu
Tiffany (Youzhi) Bao
Kai Shu

ARIZONA STATE UNIVERSITY

August 2022

ABSTRACT

With the bloom of machine learning, a massive amount of data has been used in the training process of machine learning. A tremendous amount of this data is user-generated data which allows the machine learning models to produce accurate results and personalized services. Nevertheless, I recognize the importance of preserving the privacy of individuals by protecting their information in the training process. One privacy attack that affects individuals is the private attribute inference attack. The private attribute attack is the process of inferring individuals' information that they do not explicitly reveal, such as age, gender, location, and occupation. The impacts of this go beyond knowing the information as individuals face potential risks. Furthermore, some applications need sensitive data to train the models and predict helpful insights and figuring out how to build privacy-preserving machine learning models will increase the capabilities of these applications. However, improving privacy affects the data utility which leads to a dilemma between privacy and utility. The utility of the data is measured by the quality of the data for different tasks. This trade-off between privacy and utility needs to be maintained to satisfy the privacy requirement and the result quality. To achieve more scalable privacy-preserving machine learning models, I investigate the privacy risks that affect individuals' private information in distributed machine learning. Even though the distributed machine learning has been driven by privacy concerns, privacy issues have been proposed in the literature which threaten individuals' privacy.

In this dissertation, I investigate how to measure and protect individuals' privacy in centralized and distributed machine learning models. First, a privacy-preserving text representation learning is proposed to protect users' privacy that can be revealed from user generated data. Second, a novel privacy-preserving text classification for split learning is presented to improve users' privacy and retain high utility by defend-

ing against private attribute inference attacks.

To my parents, Abdulaziz and Albandry and my husband, Meshaal

To the source of my happiness: my doughters, Reema and Alanoud

I love you and dedicate this achievement to you

ACKNOWLEDGMENTS

I have had the privilege to do my research under the supervision of Prof. Huan Liu. I am very grateful and honored to have had the opportunity. Your words of encouragement helped me move confidently in this journey. Thank you for your consistent support, guidance, and the thoughtful comments.

I would like to thank my thesis committee, Hasan Davulcu, Tiffany Bao, and Kai Shu, for their valuable interaction and constructive criticisms. I also want to thank my mentor Ghazaleh Beigi for all the help she provided when I started the Ph.D program, and for the mentoring and helpful discussion. Additionally, I would also like to thank my colleagues at the Data Mining and Machine Learning Lab (DMML) for the quality discussions and valuable feedback. In particular, I would like to thank Mansooreh Karami for being reachable and helpful. She was always there when I needed her.

It is impossible to thank my parents, Abdulaziz and Albandry, adequately for everything they have done and for always being there for me. Thank you for your unconditional love, endless support, and encouragement. I would like to especially acknowledge my husband, Meshaal Alyahya, because this journey would not have been possible without his love and support. I am very lucky to have him in my life. To my daughters, Reema and Alanoud, you are my inspiration to achieve my goal. I would like to express the deepest gratitude to my brother Riyadh and my sisters Siba, Raghad, Layan, and Leen. You have all provided support and encouragement. I would not be who I am today without you all.

Finally, a special thanks to the Ministry of Education in Saudi Arabia and the Saudi Arabian Cultural Mission (SACM) in the United States for providing me a scholarship to obtain my Ph.D. degree.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

With the bloom of machine learning, a massive amount of data has been collected, shared, and used. This raises privacy issues that impact individuals by revealing their private information such as age, gender, and location. Privacy leakage in the machine learning model arises from utilizing a vast amount of individuals' data. Although the corresponding privacy-preserving techniques have been studied and proposed in the literature, it is challenging to improve privacy while maintaining the data's utility. Improving privacy in machine learning models comes at the cost of utility, making a trade-off that needs to be maintained. The utility of the data is measured by the quality of it for different tasks.

Protecting privacy in machine learning models can be divided into two general categories: (1) protecting the privacy of published data and (2) protecting the privacy of training data while utilizing it in the machine learning model. Each category is tackled using different privacy mechanism techniques. Privacy for data publishing and sharing could be achieved by de-anonymization. The second category can be tackled using differential privacy (Dwork, 2008) or adversarial learning (Jia and Gong, 2018)(Raval *et al.*, 2019). The leakage of the individuals' privacy has been shown in different applications such as recommendation systems (Beigi *et al.*, 2020) and pre-trained language models (Carlini *et al.*, 2020b). Multiple protection techniques need to be applied to defend against different attacks in some critical applications. Recent works review different aspects of users' privacy and compare traditional privacy models for protecting users' private attributes (Alnasser *et al.*, 2020a).

Protecting individuals' privacy in machine learning is essential because of its im-

pact on individuals and societies. Utilizing datasets that contain individuals' information will cause private information leakage and abuse to arise. In addition, using personal data has been regulated by governments. Uncountable privacy regulations have been amended to regulate the use of personal data such as the California Consumer Privacy Act (CCPA) and the Health Insurance Portability and Accountability (HIPAA) in the USA, the General Data Protection Regulation (GDPR) in the European Union, the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada, and Act on the Protection of Personal Information (APPI) in Japan. Government regulations regulate the use of personal data in general and machine learning models, which primarily utilize a vast amount of personal data.

Driven by privacy concerns and toward scalability, decentralized machine learning has started in the last few years as a new mechanism of machine learning. Machine learning models can be classified to centralized and decentralized when it comes to the training process and the location of the training data. In centralized machine learning, the data is centrally pooled from different resources and the training process is centralized, while in decentralized machine learning the training process is enabled across multiple parties. Federated learning approach, which was introduced by McMahan *et al.* (2017), one of the common approaches of distributed machine learning. Then, split learning has emerged as a new distributed machine learning approach (Gupta and Raskar, 2018). In general, distributed machine leaning improves privacy by keeping the training data stored locally. However, privacy issues have been proposed in the literature, particularly when the data is very sensitive and could identify individuals' identities or attributes (Li *et al.*, 2020).

Achieving the capability to use machine learning models without revealing private information will help in critical applications such as health and finance. Predicting Alzheimer's disease or cancer types are examples of the applications that will improve

the human life if we figure out how to utilize the data privately. Providing personalized services and accurate results requires considerable personal data such as age, location, gender, and occupation. Building privacy-preserving machine learning techniques that can be trained without leaking private information will significantly improve the machine learning era.

There is vast literature on protecting the privacy of users in social media from two different perspectives: (1) identification of vulnerabilities and (2) mitigation of risks. The first group investigates the potential privacy breaches from social media user-generated data by introducing different variations of private-attribute inference attacks. The goal of these attacks is to identify possible vulnerabilities of user-generated data against leakage of private-attribute information. The second group seeks to mitigate existing privacy risks regarding the leakage of private-attribute by properly anonymizing user-generated data while preserving the utility of data. Existing private-attribute inference attacks can be classified into two classes: (1) friend-based private-attribute attacks and (2) behavior-based private-attribute attacks. Friend-based private-attribute inference attacks are based on the homophily theory, which implies that friends have more similar attributes than two random users. Behavior-based private-attribute inference attacks rely on inference using the similar users' behavior. They assume similarity between users is based on their behaviors, which further indicates they share the same attributes. Besides these two classes, other approaches use both friend and behavior information to infer users' private-attributes.

Consequently, various privacy protection models are proposed to protect users against private-attribute inference attacks. These works utilize traditional privacy preserving techniques such as k-anonymity and Differential Privacy. K-anonymity (Sweeney, 2002) is one of the traditional privacy preserving techniques which seeks to anonymize the instances in the dataset by suppression and generalization. Differential

privacy (Dwork, 2008, 2006) is another traditional technique that is applied during statistical query over a dataset and seeks to improve the privacy while preserving the accuracy of results. Figure 1.1 illustrates the attacks and the protection techniques of users' private-attributes.

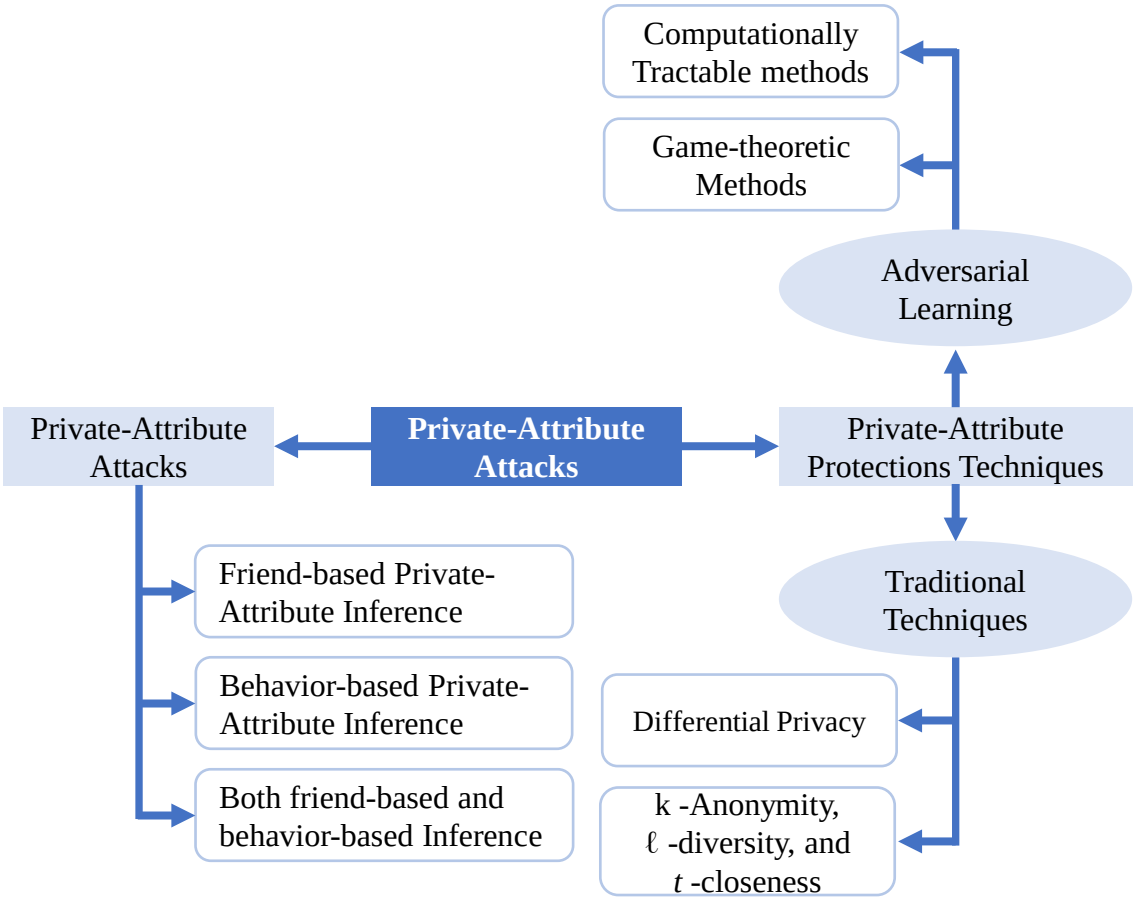

Figure 1.1: Attacks and Protection Techniques of User's Private-Attribute

1.1 Background

With the increase of information on social network platforms, a massive amount of user-generated data online is created. This user-generated data is rich in content and includes information about users' preferences and characteristics such as geo-

graphic location, gender, occupation, and age. Therefore, user-generated data has been used by researchers and service providers to better understand users' behaviors and offer them personalized services. However, publishing user-generated data causes privacy problems as this data includes information about users' private and sensitive attributes. Private attributes information is those that users do not want to explicitly disclose such as marital status, location, political view, occupation, age, and gender. Private attributes can be easily inferred by malicious adversaries from users' activities on online social networks. This privacy issue mandates social media data publishers to protect users' privacy by anonymizing user-generated social media data. Data anonymization is a challenging task and the ultimate goal of anonymization techniques is to prevent adversaries from inferring private attributes by perturbing given user-generated data. Perturbing user-generated data can affect the utility of data. This leads to a dilemma between privacy and utility and makes the problem of protecting user privacy even more challenging.

To protect users' privacy in social networks, data publishers are required to protect users' data using anonymization techniques. One technique to alleviate privacy issues is to remove the "personally identifiable information" (PII) such as name, social security number, and age (Narayanan and Shmatikov, 2009). Removing this information and keeping the graph structure has limitations in protecting users' privacy. There are several scenarios illustrating the limitation of depending on PII as an anonymization technique. For example, the anonymized dataset published for the Netflix prize challenge was not enough to preserve users' privacy.

Users' private attributes are the information and characteristics that describe individuals online. Private attributes serve different goals for the attackers; it could be for selling, delivering personalized advertisements, or performing targeted social engineering. The attacker in this context could be defined as:

Definition: ***Attacker***

The attacker could be any party that intends to disclose users' attributes using different techniques. Private-attribute inference attacks happen when the attacker intends to use the publicly available attributes to infer the missing or hidden attributes about target users. It could be formally defined as:

Definition: ***Attribute Inference Attack***

Given $T = (G, A, B)$, which illustrates a social media with a graph $G = (V, E)$ where V shows the set of users and E shows the relationship between the users, A is the users' attributes and B is the users' behavior. Attribute inference attack is to disclose the attributes a_v of v for $\forall\ v \in V_t$ where V_t is the set of targeted users using the available attribute matrix A and behavior matrix B.

1.2 Research Challenges

Measuring and enhancing users' privacy in machine learning is challenging. In this dissertation, we investigate how users' privacy is protected and measured considering these challenges:

- Trade-off between privacy and utility: Improving the privacy comes at the cost of performance or utility. To elaborate, using homomorphic encryption in machine learning models to improve the privacy has limitations regarding the performance and it can be applied in specific scenarios (Aslett *et al.*, 2015). In addition, adding noise to improve the privacy affects the utility of the data. The utility of the data means the quality of predicted results such as personalized services and accurate predictions. This challenge needs to be addressed by maintaining the trade-off between privacy and utility (or performance) as much as needed to avoid sacrificing one of them.

- De-identification is insufficient: Removing the Personally Identifiable Information (PII) from the training data is not enough and it has been shown in the literature this solution is insufficient. One example of this is the Netflix prize challenge (Beigi and Liu, 2020) where the re-identification successfully happened by linking the released dataset (Netflix) with a publicly available dataset (IMDb).

- Unscalability: Most of the privacy-preserved machine learning models need additional processing and computational cost which may affect the ability of these models to handle the massive amount of data (Al-Rubaie and Chang, 2019). This may limit real-world applications from protecting users' information in machine learning models.

- No one protection technique is able to solve all privacy issues: Protecting users' privacy and measuring the leakage can be tackled from different points of view. First, protecting the training data while utilizing it in the model using differential privacy (Dwork, 2008). Second, when the attacker has access to the input and output of the model at the test time, he can reverse engineer the model parameters and then estimate the private training data. This attack can be tackled by using differential privacy to make the reverse engineering of the parameters as hard as possible. Third, protecting the privacy of published data before sharing it using anonymization techniques such as k-anonymity (Sweeney, 2002), ℓ-diversity (Machanavajjhala *et al.*, 2006), and t-closeness (Li *et al.*, 2007). These anonymization techniques have their strengths and limitations as shown in the literature, and t-closeness is considered the strongest among these techniques. After illustrating all these categories of attacks, no one defense mechanism can be applied to improve the users' privacy in machine

learning from all these attacks. Combining multiple protection techniques to minimize the users' privacy leakage is the best practice to follow. Choosing from the privacy techniques is highly dependent on the sensitivity level of the data and the domain (Tanuwidjaja *et al.*, 2020)

1.3 Contributions

In this dissertation, we investigate how to protect individuals' privacy in centralized and distributed machine learning models while considering the trade-off between privacy and utility. In particular, we study how to maintain the dilemma so the data utility is retained and privacy is improved at the same time. The contributions of this dissertation are summarized as follows:

- Present a literature review of protecting users' private attributes and building privacy preserving machine learning models;
- Protecting users' privacy in centralized machine learning by proposing a text representation learning framework that protects individuals' privacy and retains the utility of the data;
- Protecting users' privacy in distributed machine learning by proposing a privacy preserving text classification framework on split learning setting which defends against private attribute inference and preserves data utility;
- Conduct experiments on a real world dataset to verify and demonstrate the effectiveness of our proposed frameworks.

1.4 Organization

The remainder of this dissertation is organized as follows: in Chapter 2, we review related work privacy preserving machine learning and protecting users' private

attributes. In Chapter 3, we study the privacy issues in the user generated text in centralized machine learning and propose a privacy-preserving text representation learning framework using BERT, DP_{BERT}. We use BERT to extract the sentences embedding from the textual data. Our framework, DP_{BERT}, learns the textual representation that satisfies three conditions: (1) differentially private to protect against identity leakage, (2) protects against leakage of private-attributes information, and (3) maintains the high utility for downstream tasks. In Chapter 4, we extend our scope and study users' privacy in distributed machine learning and propose a privacy preserving text classification framework on split learning setting, $PPSL$, which defends against private attribute inference and maintains the utility of the data at the same time. In Chapter 5, we conclude the dissertation and present future research directions.

Chapter 2

RELATED WORK

Online social networks enable users to participate in different activities, such as connecting with each other and sharing different contents online. These activities lead to the generation of vast amounts of online user data. Publishing user-generated data creates the problem of user privacy as this data includes information about users' private and sensitive attributes. This privacy issue mandates social media data publishers to protect users' privacy by anonymizing user-generated social media data. Existing private-attribute inference attacks can be classified into two classes: (1) friend-based private-attribute attacks and (2) behavior-based private-attribute attacks. Consequently, various privacy protection models are proposed to protect users against private-attribute inference attacks such as k-anonymity and differential privacy. This chapter will review and compare recent state-of-the-art research in terms of private-attribute inference attacks and corresponding anonymization techniques. Then, privacy in distributed machine learning will be discussed from two sides: attacks and defences.

2.1 Private Attribute Inference Attacks

Many social networks contain rich information as nodes attributes. The aim of private attribute inference attacks is to fill the missing or incomplete attribute information for a given network using different approaches. Several recent studies have demonstrated different private attribute inference attacks that could be classified as friend-based and behavior-based. This section of the chapter discusses in detail the classes of private attribute inference attacks and each class is discussed with examples.

2.1.1 Friend-based Private Attribute Inference

Friend-based private attribute inference attacks are based on the homophily theory which implies that friends have more similar attributes than two random users. The similar users seem to form communities that have a high probability for sharing the same attributes values. For instance, if the friends of a user are between 20 to 25 years old the user might also in the same age group. Initially, to illustrate the attacker process, the attacker accesses the friends' list of the target user; then does calculation to infer the hidden attributes of the target user.

Several studies have considered using the network structure to inference nodes' attributes. Ali *et al.* (2021) proposed a schema to represent the nodes in a graph as feature vectors based on its attributes and nearby nodes' attributes. These vectors will be as input for standard machine learning algorithms to predict the attributes. Chen *et al.* (2016) proposed different attribute inference models based on network characteristics and social relationships. They have implemented Naïve Bayes, Decision Tree and Logistic Regression and they found that the social relationship between users is more critical in attribute inference than network characteristics. Another work addressed the correlation between a pair of attributes that help in predicting the attributes (Rabbany *et al.*, 2017). They proposed a model called ProNe to measure the pattern and calculate the correlations of the attributes which can then be used in several applications such as prediction and privacy protection. By determining which pair of attributes can disclose users' private attribute, these attributes should be obscured.

Another direction of works in this category is to predict network structure and infer private attributes. The reason for simultaneously solving these two problems is the fact that individuals with similar attributes connect to each other and individuals who

are friends are most likely to share similar attributes (Beigi and Liu, 2020). Gong *et al.* (2014) extended the social attributed network (SAN) framework different supervised and unsupervised algorithms on attribute inference. Moreover, they observed the attribute inference could help inform the link prediction. In other words, inferring the missing attributes first will improve the accuracy of the link prediction. Their work used the network structure and node-attributes information to enhance the performance of both tow problems link-prediction and attribute inference.

2.1.2 Behavior-based Private Attribute Inference

Behavior-based private attribute inference attacks rely on inference using similar users' behavior. They assume similarity between users is based on their behaviors, which further indicates they share the same attributes. For instance, if a user liked books, shared articles, and participated in hashtags that are similar to those who that are liked and used by users majoring in computer science, the user might also major in computer science. Indeed, the attacker needs to analyze the behavior of the target user in order to infer the hidden private attributes.

A lot of research has focus on inferring the private attributes using users' behavior. Alipour *et al.* (2019) proposed gender inference attacks on Facebook users based on their published pictures using alt-text generated by Facebook and comments that are written about the picture by friends, friends of friends, or strangers. They explained how the attacker design feature sets with the public attributes to infer specific hidden attributes. Other work for gender expression classification model is called GENEC and predict the gender of users (Filho *et al.*, 2016). This classification model uses 60 textual meta-attributes such as characters, syntax, words, structure, and morphology of short length. This work is based on the relationship between gender and the used language. Gender classification can be defined as the task of identifying each user in a

network of male or female gender by analyzing the content and the behavior illustrated in their messages. Several algorithms have also been evaluated and compared in performance and the results show the use of Best First Tree (BFTree) algorithm can achieve excellent results. Another work considered many social media platforms simultaneously instead of using only a single resource to infer the users' attributes (Nie *et al.*, 2017). Since the majority of users use many social network platforms at the same time, the authors studied their behavior across multiple platforms to infer the users' attributes. They proposed a unified multi-source learning model to infer the users' occupations which treats each occupation as a task and simultaneously adjusts source consistency and task relatedness. Historical multimedia posts from multiple mobile platforms such as Twitter, Foursquare, Instagram, and LinkedIn are crawled for each user first and extract descriptive features. The proposed model jointly learns the features by lasso and graph-guided fused lasso. The graph-structure is built up by leveraging external and internal knowledge.

2.1.3 Both Friend-based and Behavior-based

Uncounted approaches combine both network links and user behavior information to infer the private attribute. Combining these two approaches can be used to improve the effectiveness of inferring the private attributes. Gong and Liu (2018) amalgamate social structures, user behaviors, and user attributes and design a social-behavior attribute network called social-behavior-attribute (SBA). After that, they proposed an attack called vote distribution attack (VIAL) under the SBA network to perform attribute inference. Jia *et al.* (2017) proposed a model called AttriInfer which is a Markov Random Field (MRF) based method. This method aims to infer users' private attribute using their public data by leverage both friends, behavior, and the label information of training users who have an attribute and who do not. Besides, they

modeled AttriInfer using Loopy Belief propagation (LBP) to compute the posterior probability, which is the probability that a specific user has the attribute giving a training dataset. Zhong *et al.* (2015) demonstrated a framework called a location to profile (L2P) to infer demographic attributes of online users such as gender, age, education level, marital status, and blood type. The proposed model incorporates three aspects embedded in the check-in data which are spatiality, temporality, and location knowledge features. Spatiality illustrates that the location check-ins are not uniformly distributed in the geospatial space. Temporality means the location check-ins are changing over time. Location knowledge describes the strong correlation between individuals with the functionality of locations that encourage individuals to travel between different places. Another work integrated social structures and attributes into a probabilistic model to predict targeted users' attributes with the assumption of powerful adversaries with background knowledge (He and Huang, 2019).

Moreover, Cai *et al.* (2018) presented a novel implementation method for collective inference which can effectively inference users' sensitive information using both available attributes and friendship information. They showed exactly how the attacker launches an inference attack to predict users' private attributes by investigating a typical inference attack called collective inference. The collective inference is using a network to propagate the current inference results iteratively to improve the accuracy. They illustrated the impact of utilizing both attribute and link information which can significantly improve the accuracy of private attributes inference attacks. Other work is proposed to predict the voting behavior of users in social media using a Bayesian network model that combines demographic information, users' behavior, and social structures. The goal of the model is to predict the user voting behavior on Facebook based on the public portions of the user's profile. Bayesian network classifiers have strong points which are the ability to support the combination of data, the

previous knowledge about the specific domain, and handling the missing values very well. These points are important with social network datasets (Idan and Feigenbaum, 2019).

Another work investigated the relevance among social attributes, which can be used to infer users' private attributes. One attribute may have relevant to others which makes the population distribution of one attribute over the second attribute unbalanced. For instance, the attribute gender may have relevant to attribute job as the population distribution of job over gender is unbalanced. The distribution can be seen in some types of jobs where most workers are male or female. That is to say, the relevant among different attributes affects the performance influences on private attribute inference (Mao *et al.*, 2019). The authors proposed a relevance attribute inference method called ReAI using random walks based on a social graph. They constructed a social graph and embedded the relevant values among attributes into the social graph as edge weights.

2.2 Protection Techniques

Online users' privacy has become a critical problem for the researchers which has led to vast solutions for mitigating the risk of inference attacks. After discussing the different approaches of private attribute inference attacks, this section will explore how researchers mitigate the risk of privacy leakage by providing different defense algorithms. Traditional privacy-preserving techniques will be discussed first, and then we will discuss how adversarial learning can preserve the users' privacy.

2.2.1 Traditional Privacy Preserving Techniques

a. Differential Privacy

Differential privacy is one of the traditional techniques applied during statistical

query over a dataset and seeks to improve the privacy while preserving the accuracy of results. The goal of differential privacy is to maximize the accuracy of the queries and minimize privacy leakage. It was developed by Dwork (2008) and it provided a guarantee that the behavior of the dataset will hardly be affected after a single instance is added or removed from the dataset. In essence, the presence or absence of a specific individual in the dataset is insignificant. For example, consider a survey about an embarrassing activity and out of 100 participants 20 participants respond "Yes". When a new participant answers the questions, his answer could be easily inferred after comparing the result of the survey before and after his answer. This example illustrates the problem that differential privacy has proposed to solve. The formal definition of differential privacy as Dwork (2008, p.2) defined it:

A randomized function K gives ϵ-differential privacy if for all data set D_1 and D_2 differing on at most one element,and all $S \subseteq Range\ (K)$,

$$Pr[K(D_1) \in S] \leq exp(\epsilon) \times Pr[K(D_2) \in S] \tag{2.1}$$

Where the probability is taken is over the coin tosses of K and ϵ is called privacy budget and it is a metric of privacy loss. The smaller value of ϵ is, the higher privacy preserved because the attacker is unable to infer the change in the database after adding or removing an instance. K is a randomized function that is independent of any knowledge the attacker may have about the database. There are two widely used approaches for adding random noises to achieve differential privacy, namely, the Laplace mechanism for numeric data and the exponential mechanism of non-numeric data. The Laplacian mechanism is a common technique for adding Laplace noise drawn from Laplace distribution. The quantity of added noise depends on the sensitivity of the query function. The sensitivity of the query describes the difference

between the value of the function on the two databases D_1 and D_2 which differ in at most one instance that needs to be hidden by adding noise. To calculate function sensitivity (Dwork, 2008, p.4):

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \tag{2.2}$$

Differential privacy introduces the least possible noise when Δf is small. For example, simple queries that need only counting such as, counting instances with attribute $x = 1$ have quite small sensitivity value.

Differential privacy was initially proposed to protect the privacy of individual instances in a statistical database. Then, it was adopted to preserve privacy in social networks. Researchers have applied differential privacy to preserve the privacy of the data in social networks using different approaches. Xiao and Xiong (2015) proposed a solution to preserve location privacy using differential privacy. They protected the exact location for the user at every timestamp by defining "δ–location set". To protect the user's exact location, they hid it in the δ –location set and as a result, any pairs of locations are not distinguishable. Moreover, an efficient location perturbation mechanism called planar isotropic mechanism (PIM) was presented in their work to achieve the differential privacy. Liu *et al.* (2020) proposed a local differential privacy model for social network publishing called DP-LUSN that preserved community structure information and provided a proof that the local differential privacy model satisfies the definition of differential privacy while preserving the utility of the community structure. In local differential privacy, the community structure is considered an independent unit and the local neighboring databases are defined as a community with one different edge.

b. k-Anonymity, ℓ-Diversity, and t-Closeness

According to Sweeney (2002), k-anonymity is one of the traditional privacy-

preserving techniques which seeks to anonymize the instances in the dataset by suppression and generalization. Suppression implies not releasing the value of an attribute. For example, instead of releasing the value of the zip code of an instance, "*" will be released. Generalization involves making the value of an attribute more general and less specific. For example, releasing the value of the zip code of an instance as "1234*" instead of "12345". These two ideas have been combined to achieve k-anonymity. The basic idea of k-anonymity is to make each instance in the dataset identical from at least k instances that share identifying attributes. The k instances of the same attributes are indistinguishable regarding their quasi-identifiers. Quasi-identifiers are defined as the attributes that are linked with an external public dataset that uniquely identifies at least one individual (Machanavajjhala *et al.*, 2006). An example of quasi-identifiers is social security numbers which can identify individuals. Furthermore, not only explicit information can identify the individuals, but also the combination between some attributes can uniquely identify the individuals.

Even though k-anonymity has protected the privacy of the released dataset, it cannot protect against private attribute inference attacks. k-Anonymity is susceptible to some types of attacks which are homogeneity attack and background knowledge attack (Machanavajjhala *et al.*, 2006). Homogeneity attack is defined as the ability to infer an instance's private attributes when sensitive values are in an equivalent class that lack diversity. Background knowledge attack presumes the attacker knows a piece of information about one of the instances in the dataset. Because of these attacks, k-anonymity is insufficient to prevent private attribute inference attacks. In addition, k-anonymity cannot protect against private attribute inference.

To solve k-anonymity's limitations, ℓ-diversity was introduced by Machanavajjhala *et al.* (2006) and ensures the diversity of the sensitive attributes at each equivalence class. A dataset is said to have ℓ-diversity if there are at least ℓ well-represented values

for the sensitive attributes. Two instantiations of the ℓ-diversity are introduced that both result in diversity in the sensitive attributes: entropy ℓ-diversity and recursive (c, ℓ)-diversity. Entropy ℓ-diversity is defined as the entropy of the distribution of the sensitive attributes in each equivalence class to be at least $log(\ell)$. Recursive (c, ℓ)-diversity means the most frequent value should appear frequently enough in the dataset. It is clear that ℓ-diversity has several advantages against the previous attacks that risk k-anonymity. First, ℓ-diversity does not require knowledge of sensitive and non-sensitive attributes. Second, it protects against all different background knowledge attacks. The larger the value of parameter ℓ, the more information is required to eliminate the values of the sensitive attributes.

Even though ℓ-diversity achieved higher privacy than k-anonymity, it has negative aspects and vulnerabilities. First, ℓ-diversity is considered difficult to achieve. Second, ℓ-diversity is vulnerable to two types of attacks: (1) skewness attacks and (2) similarity attacks. Skewness attacks are defined as gaining information about a sensitive attribute when the global distribution of this attribute is available. ℓ-Diversity is vulnerable to another attack called similarity attack which happens when the sensitive attributes are distinct but are similar semantically. This because ℓ-diversity does not consider the semantical closeness of the values (Li *et al.*, 2007).

A new privacy concept was introduced, t-closeness which requires the distribution of a sensitive attribute to be close to the distribution of the attribute in the whole dataset. To satisfy t-closeness, the distance between the distribution of a sensitive attribute in an equivalence class and the distribution of the attribute in the whole dataset is no more a threshold t. t-closeness is calculated for a dataset based on Earth Movers Distance (EMD) (Li *et al.*, 2007).

To discuss these three models and their ability to protect the users' private attributes, k-anonymity does not protect against private attribute attacks. That hap-

pens when the variability of sensitive attributes in the equivalence class is low which gives the attacker the ability to decide the equivalent class of a user that will disclose information about the sensitive attributes of that user (Soria-Comas *et al.*, 2015). ℓ-Diversity is insufficient to prevent the private attributes attacks as discussed before; it is vulnerable to two types of attacks. The reason for this limitation in ℓ-diversity is because the semantical closeness is ignored (Li *et al.*, 2007). t-Closeness is considered the strongest among the three models by limiting the number of users' private attributes that can be observed by the attacker.

These traditional privacy techniques focus on securing the released dataset and do not consider the inference attacks using machine learning techniques (Han *et al.*, 2019). The following section discusses the adversarial learning techniques to preserve users' privacy and protect users' private attribute. To conclude, this section discussed the traditional techniques to preserve users' privacy by protecting the private attributes.

2.2.2 Adversarial Learning

Adversarial learning is the state of the art approach for privacy techniques to preserve users' privacy by defending against machine learning models that target to infer private attributes. Recently, uncountable works have started defending against private attributes inference attacks.

a. Game-theoretic Methods (Intractable)

Game theory is a mathematical discipline that studies the strategic interaction among rational individuals. It was established based on the assumption that there are players and a strategy that will make one player win the game (Wang *et al.*, 2016). Game theory is wildly used to solve different problems in many fields. The purpose of using game theory to defend against private attribute inference is to propose strategies

to keep an equilibrium between data utility and privacy preservation. In this method, the attacker will perform the inference attack based on the previous knowledge of the defense algorithm. On the other hand, the defense system will protect against the optimal inference attack. The dilemma between data utility and privacy preservation can be illustrated using data user, data collector, and data provider as follows: the data user performs different algorithms to extract the knowledge on the data offered by the data collector. It will be better for the data user to have less anonymized data so that more relevant knowledge and patterns can be extracted. On the other hand, data providers prefer to keep private data secured, so they want to anonymize it. The data collector is responsible for deciding how much anonymization the data needs (Shah *et al.*, 2019).

Several works solved the problem of the trade-off between privacy and utility and modeled the interactions between different parties as a game. Xu *et al.* (2015) discussed this concept by modeling the interactions among data providers, data collectors, and data users as a game. A general approach has been proposed in this work to find the Nash equilibriums of the game, which is the optimal level of anonymization that need to be applied on the data. They also presented a specific game formulation that takes k-anonymity as the anonymization method. These defenses are considered computationally intractable when applied to attribute inference attacks.

There have been noteworthy works to solve different problems in privacy using game theory approaches. Shah *et al.* (2019) reviewed the application of game theory in privacy preservation. A comparative study of the uses of different game-theoretic models in privacy preservation was proposed. One example of applying game theory is for defending against location inference attacks (Shokri *et al.*, 2012). The authors proposed an analytical framework that enables system designers to find the optimal location-privacy preserving mechanism (LPPM) for location-based services

against the optimal inference algorithm. They used Bayesian Stackelberg games to find the solution and the scenario is as follows: a user and an adversary interact based on a strategy where each one's gain is the loss of the other. The user plays first by choosing a location-privacy preserving mechanism (LPPM) and running it on his exact location. Then, the adversary will play by predicting the user's location, given the location-privacy preserving mechanism that the user has run. This is the Bayesian game because the adversary has incomplete information about the user's exact location and he is playing the game depending on his hypothesis about the user's location. As a result, the authors found the optimal point in the trade-off which satisfies both user privacy and service quality based on users' information. Applying game-theoretic methods have theoretically proved privacy preservation and defending against inference attacks, but they have several limitations. The computational cost is exponential because the public data vector in the real dataset has high dimensionality. In addition, to defend against private attribute inference, noise has been added to a user's public data which could affect the utility (Jia and Gong, 2018).

b. Computationally Tractable Methods

As a solution to the computationally intractable problem, different researchers have proposed many traceable approaches. Salamatian *et al.* (2015) developed Quantization Probabilistic Mapping (QPM) which relies on a general statistical inference framework. They reduced the amount of data by clustering the users' public data and take the cluster centroid to represent each cluster. Another work proposed practical adversarial machine learning to defend against attributes inference attacks called AttriGuard. It finds a minimum noise for each attribute value then randomly selects one of the previously founded noises to mislead the attacker. More simply, the goal is to add random noise to the attribute to reduce the attacker's inference accuracy with minimum utility loss (Jia and Gong, 2018). Another work approached adver-

sarial learning differently to guarantee that the private attributes are protected from all machine learning attackers (Raval *et al.*, 2019). The authors proposed a privacy framework called Olympus to eliminate the risk of inferring private attributes by obfuscating the data and preserving the utility of the data. Their proposed framework used a generative adversarial network (GAN) to solve the problem between the obfuscator and the attacker.

Several methods are reported in the literature to address how to preserve the textual information privacy. Beigi *et al.* (2019) used adversarial learning to anonymize the textual information by proposing a privacy-preserving text representation learning framework called DPText. The final output of this model is the text that obscures the private attribute information by making sure the sensitive attributes are not captured in the latent representation so the adversary will not be able to infer these attributes. To create private representation in the text field, Li *et al.* (2018) proposed a novel approach for privacy-preserving learning based on generative adversarial network (GAN) to train deep models with adversarial learning in order to improve the robustness and privacy of the neural representation. Their method has been evaluated on the tasks of part of speech tagging (POS), sentiment analysis, and protecting several demographic private attributes such as gender, age, and location.

A series of recent successful studies used deep reinforcement learning to preserve privacy. By leveraging reinforcement learning in controlling privacy-utility balance, feedback from the attacker is included in a reward function. Mosallanezhad *et al.* (2019) proposed a novel reinforcement learning-based text Anonymizer, RLTA, which addresses the problem of private attribute inference while preserving the utility. RLTA consists of two components: an attention-based task aware text representation learner and a deep reinforcement learning-based privacy and utility preserver. The goal of these components is to extract the embedded representation of the text by minimizing

the loss and then manipulate the embedded text by learning the optimal strategy that preserves both privacy and utility. The incorporation of deep reinforcement learning in this model works to anonymize the text embedding by receiving privacy and utility feedback and learning the optimal balance for proper manipulation of text embeddings.

From the data provider point of view, preserving the privacy in textual information is challenging in two ways. The first is the trade-off between privacy and the semantic of the text which will be affected when the text is converted to the new form. For example, this sentence, "I'm a math teacher with 17 years' experience" will discloses one of the user's private attributes, age, since it clearly shows that the user is over 40-years-old. The second challenge is the disclosure of a private attribute is indirect, such as how some words are highly used by females (Xu *et al.*, 2019). Several works have addressed these challenges and preserved the private attributes in textual information. Xu *et al.* (2019) developed a tool to rewrite the users' text into less sensitive text to preserve the users' privacy. Their proposed model is based on back translation to reduce the exposure of private information.

The directions of other works focused on preserving the users' privacy in the context of the recommendation system. Recommendation systems build a profile for each user which include user's public information and interests then recommend relevant products to each user based on these profiles. Despite the capability of the recommendation system, they could be a source of private attributes inference attacks when the attackers have access to the output of the recommendation system and information about the targeted users. Beigi *et al.* (2020) proposed an adversarial learning-based recommendation with an attribute protection model called Recommendation with Attribute Protection (RAP). RAP protects users against private attributes inference attacks and maintains utility. It consists of two integrated components: Bayesian

personalized recommender and the private attributes attacker and it has shown the effectiveness in both protecting users' privacy and preserving the quality of the recommended items. This proposed model preserves the utility by offering relevant products and, at the same time, preserves the privacy by making the private attributes inference challenging for the adversary.

A few other works have proposed methods to estimate online users' private attributes inference risk. Different works have been designed to measure a specific kind of attack, but in reality, the behavior of the attacks is unpredictable and various adversaries will perform different inference attacks. To address this limitation, Han *et al.* (2019) proposed a general framework for private attribute disclosure estimation, called F-PAD, which can estimate inference risk for users given a basket of different inference attack models. F-PAD consists of three steps: 1) attack model simulation to simulate adversaries' attack that predicts private attributes of users, 2) disclosure model training to study the inference results in order to learn from the correct and wrong predictions, and 3) disclosure risk estimation to build up the disclosure estimator that integrates the result of multiple models within a high confidence interval in terms of disclosure probability and risk level. The goal is to generalize the privacy estimation function to cover many attributes. Moreover, F-PAD offers some suggestion to increase the privacy preservation for the users who have a high disclosure risk.

2.3 Privacy in Distributed Machine Learning

Although distributed machine learning improves privacy by storing the training data locally, privacy issues have arisen and different solutions have been proposed in the literature. The shift from centralized machine learning to distributed machine learning can be classified to: federated learning (McMahan *et al.*, 2017) and split learning (Gupta and Raskar, 2018).

2.3.1 Privacy Threats in Distributed Machine Learning

From a privacy perspective, the privacy threats in distributed machine learning that affect individuals' private information can be classified based on the attack source as follows:

1. **Attacks from the server:** When the server is a malicious server that infers private information about the participants' training data. In federated learning, Melis *et al.* (2019) illustrated how the attack happens when the server is an adversary and his goal is to infer information about the participants' training data. The adversary analyzes the participants' updates during the training process, and he can do either passive or active attacks. In split learning, Pasquini *et al.* (2021) demonstrated a malicious server which recovers the private training data of the participants. The malicious server hijacks the model's learning process to do an inference attack.

2. **Attacks from a client:** When one of the clients in the distributed environment is a malicious client who infers other participants' training data. In federated learning, Melis *et al.* (2019) illustrated the case when a malicious client can infer private information from the training set, such as specific locations. In split learning, Pasquini *et al.* (2021) proposed a malicious client, which recovers the private training data from other participants in the distributed learning process.

3. **External attacks:** When the source of the attack happens from the communication between entities. In federated learning, Singh *et al.* (2019) proposed a reconstruction attack which is when the attacker has access to the updates and communication between the participants. In split learning, Vepakomma *et al.*

(2020) proposed a reconstruction attack by utilizing the intermediate updates between the client and the server. They showed how the private data leaked through the communication between the entities even after passing some layers on the client side.

2.3.2 Privacy Protections in Distributed Machine Learning

In the last few years, we have seen a growing body of works proposing privacy-preserving distributed machine learning that utilizes different protection techniques. In the light of protecting individuals' private information in distributed machine learning, several mitigation techniques have been added to federated learning and split learning. Table 2.1 shows the different protection techniques to improve individuals' privacy in federated and split learning.

Geyer *et al.* (2017) proposed an enhancement in federated learning by applying differential privacy as a protection technique. They improved the clients' level of privacy by defending against membership attacks, which makes it harder for the attacker to infer whether a client participated in the training process or not. Another protection technique is Secure Multi-party Computation (SMC), which collectively computes a function over multiple parties. Bonawitz *et al.* (2017) used Secure Multi-party Computation with federated learning to enhance privacy. In addition to SMC, another protection technique called homomorphic encryption has been used with federated learning to improve privacy (Cheng *et al.*, 2019). The proposed privacy-preserved system, SecureBoost, improves privacy while maintaining performance.

In split learning, Abuadbba *et al.* (2020) applied differential privacy on the split layer to protect each entity's training data. On each client, a specific amount of noise is added to the activation parameter before sending it to the server. As a second mitigation, they have increased the model complexity by adding more hidden

layers on the client-side. Another work demonstrated how to reduce the information leakage by minimizing the distance correlation between the training data and the intermediate parameters across the clients and the server. They applied their ideas by adding the distance correlation as an additional loss term to the classification loss term (Vepakomma *et al.*, 2020).

Table 2.1: Privacy Protection Techniques in Distributed Machine Learning

	Federated Learning	Split Learning
Differential Privacy	(Geyer *et al.*, 2017), (Choudhury *et al.*, 2019)	(Abuadbba *et al.*, 2020)
Homomorphic Encryption	(Cheng *et al.*, 2019)	(Pereteanu *et al.*, 2022)
Secure Multi-party Computation (SMC)	(Bonawitz *et al.*, 2017)	-
Other Protection Techniques	-	Minimizing distance correlation (Vepakomma *et al.*, 2020), increased the model complexity (Abuadbba *et al.*, 2020)

2.4 Conclusion

The amount of online data has increased. Social media platforms are one of the rich areas of user-generated data which contain sensitive information. User-generated data has been studied from researcher and service providers to better understand users' needs. However, sharing this information may risk the users' privacy since it includes sensitive information or private attributes. Research has shown that private attributes inference is one of the privacy risks that concern online users. A private

attributes inference attack discloses private attributes using the available public attribute by a malicious adversary. Most of the existing works are divided into two groups: proposing new attacks to infer the users' private attributes and defending against that risk to minimize the privacy leakage. In this chapter, existing private attribute inference attacks have been explained in both classes: friend-based private attribute attacks and behavior-based private attribute attacks. After that, the protection techniques against private attribute inference attacks were discussed. There are various protection models that were classified into traditional techniques and adversarial learning. I reviewed, categorized, and compared the existing works in terms of users' private attributes. Finally, I reviewed the privacy in distributed machine meaning from two sides: attacks and defences.

Chapter 3

PRIVACY PRESERVING TEXT REPRESENTATION LEARNING USING BERT

3.1 Introduction

User generated textual data is rich of information that can be used in different tasks such as understanding users' behavior and recommendation systems. From the privacy perspective, user generated textual data can cause a privacy leakage since it contains sensitive information about the individuals. There are several privacy issues related to textual data such as re-identification and private-attributes inference.

Online users who publish textual data may not be aware that their private information can be easily inferred by malicious adversaries. Many research studies have shown that the user generated textual data may reveal private information about the users. The following table 3.1 shows examples of how user-generated textual data reveal private information. From the first example, it is clear that one can infer the user's gender (female), which is one of the private attributes that the user may not want to share publicly. Example 2 shows the leakage of private health status that can be inferred from a tweet that contains disease symptoms. Individual's location and age group can be inferred from user-generated textual data as shown in examples 3 and 4, respectively.

Table 3.1: Examples of Private Information Leakage in User Generated Textual Data

User generated textual data	Revealed information
"...I should receive my shoes at the end of next week so I waited by the end of the week there was no shoes so ***my husband*** *called and was told the order never shipped out..."*(Hovy *et al.*, 2015)	Gender (Female)
"Dr.appt Tuesday morning was told I need to lose 30 pounds by X-Mas, ***have high cholesterol****, and* ***high blood pressure****. Today starting counting calories #myfitnesspal and juicing for dinner"* (Beigi *et al.*, 2019)	Disease symptoms
"Bravofly is a rip off. Reimbursed me 97,76 Euros out of 265,06. I will report them to ***the French Interior Ministere Service*** *des Fraude/Escroquerie etc."*(Hovy *et al.*, 2015)	Location (France)
"Well what can I say The show was Brilliant and well worth going to watch We took out ***2 granddaughter's*** *ages 5 and 13 "* (Hovy *et al.*, 2015)	Age group (above 45)

Two categories in general of information leakages have been studied: identity disclosure and private-attributes disclosure. Identity disclosure happens when a targeted instance is mapped to an instance in a publicly released dataset while private-attributes leakage happens when the adversary is able to infer some of the sensitive information such as age, gender, and location (Beigi and Liu, 2020). To protect user's privacy, various protection techniques have been proposed such as k-anonymity

(Sweeney, 2002) and differential privacy (Dwork, 2008) which used to tackle the identity disclosure attack. However, these techniques have shown inefficiency to protect textual generated data for several reasons such as the data being unstructured and contains a huge number of short and informal post (Fung *et al.*, 2010). Besides, these techniques do not protect textual information against private-attributes leakage and also may have a negative impact on the utility as they do not take it as a part of the solution.

Our main contribution is proposing a framework, called DP_{BERT}, which learns a privacy preserved text representation that is differentially private to protect against identity leakage (if a target instance is available in the data or not), does not leak private-attributes information (age, gender, location, etc.), and preserves the semantic of the text.

3.2 Related Work

This section describes related work on the following areas: (1) Sentence embedding; (2) Textual Data Privacy; and (3) Protecting Private-Attributes Information.

3.2.1 Sentence Embedding

is the process of converting a linguistic sentence to a numerical representation taking into account its meaning. The aim of sentence embedding is being able to use the logistic features for downstream tasks. There are two categories for sentence embedding techniques:non-parameterized and parameterized models (Wang and Kuo, 2020). Non-parameterized techniques such as tf-idf and uSIF (Ethayarajh, 2018) which depend on high-quality pre-trained word embedding techniques. On the other hand, parameterized techniques are more convoluted than non-parameterized techniques. SBERT(Reimers and Gurevych, 2019) technique is based on BERT foundation. The

work of (Wang and Kuo, 2020), called SBERT-WK, proposed a new sentence embedding method by using geometric analysis of the space learned by deep contextualized models.

3.2.2 Textual Data Privacy

User-generated data has been used by researchers and service providers to better understand users' behaviors and offer them personalized services. However, publishing user-generated data may cause the problem of user privacy as this data includes information about users' private information. Several methods are reported in the literature to address how to preserve the textual information privacy. Beigi *et al.* (2019) used adversarial learning to anonymize the textual information by proposing a privacy-preserving text representation learning framework, called DPText. The final output of this model is the text that obscures the private-attribute information by making sure the sensitive attribute does not capture in the latent representation so the adversary will not be able to infer these attributes. The previous study on creating private representation in the text field by (Liu *et al.*, 2020) proposed a novel approach for privacy-preserving learning based on generative adversarial network GAN to train deep models with adversarial learning to improve the robustness and privacy of the neural representation. Their method has been evaluated on the tasks of part of speech tagging (POS) and sentiment analysis, protecting several demographic private-attributes such as gender, age and location.

3.2.3 Protecting Private-Attributes Information

Private-attribute information can be defined as the information that individuals do not want to explicitly disclose such as marital status, location, occupation, age, and gender. Numerous research studies have been mitigated the problem of privacy

leakage. Recent works (Alnasser *et al.*, 2020b) identify different aspects of user privacy. In particular, Alnasser *et al.* (2020b) illusterate the privacy risks and compare different traditional privacy models for protecting user private-attributes.

Our work is distinct from the previous works in that we use BERT to extract the sentence embedding in our proposed model. Our model can be useful in preserving privacy in published datasets that are used for different tasks.

3.3 Problem Statement

PROBLEM 1. *Given a set of documents X, a set of sensitive attributes P, and a task T, learn a function f that can anonymize the text embedding representation $\tilde{Z}_i$ for each document x_i in X so that, 1) the adversary cannot infer the targeted user's private-attributes P from the privacy-preserving text representation $\tilde{Z}_i$ and 2) the generated private representation $\tilde{Z}_i$ is preserving the utility for a downstream task T.* The problem can be mathematically defined as Beigi *et al.* (2019):

$$\tilde{Z}_i = f(x_i, P, T) \tag{3.1}$$

Note that our work aims to protect against external attackers who have access to the released dataset. We assume the system to be trusted.

3.4 The proposed Framework

In this section, we discuss the details of the proposed model framework which is an extension to a novel double privacy preserving text representation learning framework, DBT_{EXT}, (Beigi *et al.*, 2019). The illustration of the entire model is shown in Fig. 3.1. Our proposed model framework uses BERT for text representation. BERT (Devlin *et al.*, 2019) is a language representation model which is developed to pre-train deep bidirectional representations from a given text by jointly conditioning on both left

and right context in all layers. Then, the differential-privacy-based noise adder adds random noise, e.g., a Laplacian noise, to the original text representation. Since adding noise affects the semantic meaning and may destroy the utility of the data, we added two discriminators for the semantic meaning D_S and private-attributes D_P to infer the proper amount of the added noise. The semantic meaning D_S verifies that the added noise does not destroy the semantic meaning given the sentiment classification task. The private attribute discriminator D_P controls the amount of added noise to make the manipulated representation does not leak users' private information. The final output of the proposed model is the manipulated embedding text $\tilde{Z}$ which is deferentially private, hides the private attributes, and preserves semantic meaning.

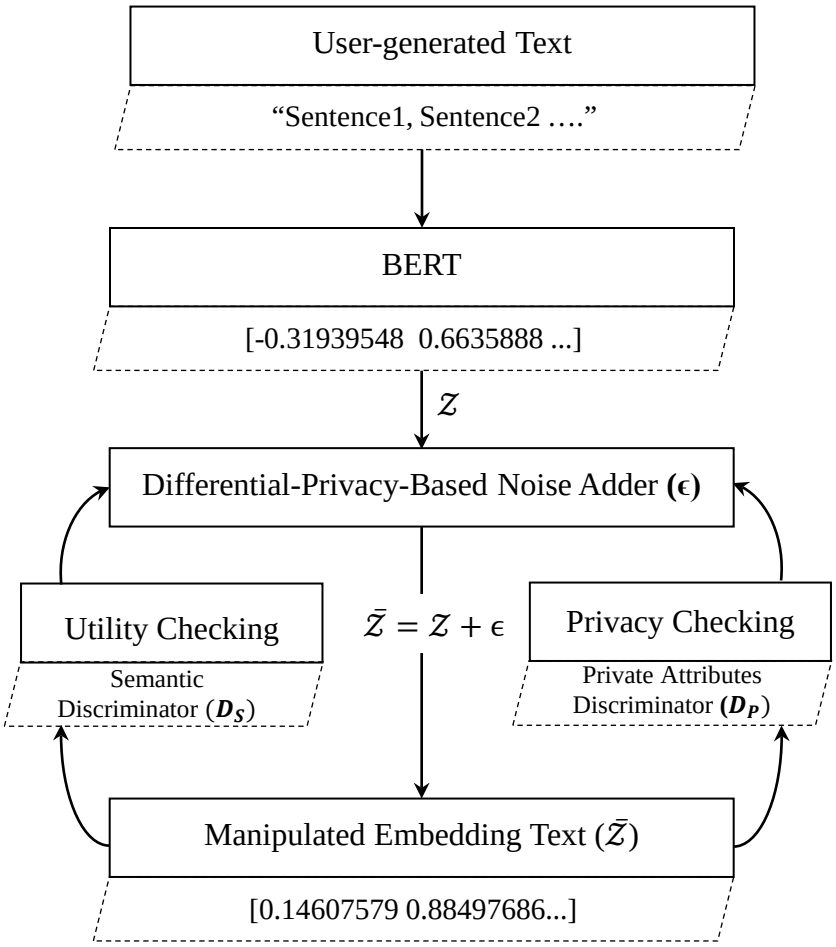

Figure 3.1: The Framework of DP_{BERT} Architecture. It Consist of Four Components, BERT, a Differential Privacy- based Noise Adder, a Semantic Discriminator D_S and a Private-attributes Discriminator D_P. The Manipulated Embedding Text $\tilde{Z}$ Is a Noisy Representation Which is Deferentially Private, Hides Private Information and Has Semantic Meaning

3.4.1 Sentence Representation using BERT

Here, we illustrate how to extract the sentence embedding for a given text. Let $X = \{x^1, ..., x^m\}$ be a document that contains m sentences. We use BERT to extract the textual embedding because BERT has shown to be significantly efficient

when modeling textual embedding (Devlin *et al.*, 2019; Reimers and Gurevych, 2019; Wang and Kuo, 2020). In particular, we use Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), which achieves state of the art performance for various sentence embeddings task, to extract the sentence embedding. SBERT is a modification of BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings.

3.4.2 Perturbing Text by Adding Noise

Textual data is rich in content, leading to privacy issues by revealing information about individuals. The privacy issues can be identity of private attribute attacks. Furthermore, when the attacker accesses the text latent representation, he can reverse engineer to the original input. Thus, it is crucial to protect textual information to improve individuals' privacy. From the literature, we have seen that differential privacy is an early and powerful protection technique to improve the users' privacy by providing a privacy guarantee. Adding noise to the textual representation will prevent text re-identification. In this work, we follow (Beigi *et al.*, 2019), where differential privacy technique is used for preserving the privacy of users' data. By adding noise to the latent representation, we improve the privacy by making it harder for the attacker to do re-identification of learned text representation and preventing the attacker from recovering the raw textual data. Output perturbation mechanism is used which is achieving the differential privacy by adding Laplacian noise to the output of an algorithm $\tilde{Z}$ (Chaudhuri *et al.*, 2011). We added Lablacian noise to perturb the output Z as follows:

$$\tilde{z}_i = z(i) + s(i), s(i) \sim Lap(b), b = \frac{\triangle}{\epsilon}, i = 1, \ldots, d \tag{3.2}$$

where s the noise vector, $s(i)$ and $z(i)$ are the $i-th$ element for vectors s and z, respectively, $\triangle$ is the L_1-sensitivity of the latent representation z, ϵ is the privacy budget and d the dimension of z.

Instead of directly sampling noise $s(i)$ using Laplacian mechanism to learn the value of the privacy budget ϵ, we use reparameterization trick which was introduced first in a work done by Kingma and Welling (2014). It first samples a value r from a uniform distribution and then rewrites $s(i)$ as follows:

$$\mathbf{s}(i) = -\frac{\Delta}{\epsilon} \times \text{sgn}(r) \ln(1 - 2|r|), \quad i = 1, \ldots, d \tag{3.3}$$

3.4.3 Preserving Text Utility

As discussed previously, adding noise (Eq. 3.2) to the textual representation will prevent privacy leakage. However, adding noise comes at the cost of text utility loss. We measure text utility by its semantic meaning. In order to preserve the text's semantic meaning, the optimal amount of noise needs to be added to ensure not too much noise has been added to the textual embedding as it can reduce the utility of the textual information. We need to add an optimal amount of noise which does not destroy the semantic meaning of the text data and in the same time ensuring data privacy. This addresses the trade-off between preserving privacy and maintaining utility. We train a classifier to learn the amount of added noise with the privacy budget ϵ as:

$$\hat{y} = softmax(\tilde{z}; \theta_{D_S}) \tag{3.4}$$

where $\hat{y}$ represents the inferred label for the classification and θ_{D_S} are the weights associated with the softmax function.

For the semantic meaning of the text representation, we define a semantic discrim-

inator D_S to assign a correct class label to the perturbed representation as follows:

$$\min_{\theta_{D_S},\epsilon} \mathcal{L}(\hat{y}, y) = \min_{\theta_{D_S},\epsilon} \sum_{i=1}^{C} -y(i) \log \hat{y}(i) \tag{3.5}$$

where $\mathcal{L}$ is the cross entropy loss function, C is the number of classes, y is the ground truth label for the classification task and $y(i)$ represents the $i-th$ element of y.

3.4.4 Protecting Private Information

As we discussed, adding noise to the textual representation will prevent adversaries from inferring the user information. The other side of preserving text representation privacy is to ensure that the sensitive information of the individuals is not captured in the text representation. In our proposed model, we follow the idea of adversarial learning by training a private-attributes discriminator D_P that identifies the private information from the text representation (Beigi *et al.*, 2019). At the same time, learning a representation that minimizes the leakage of private information by fooling the discriminator. The adversarial learning can be formally written as:

$$\min_{\left\{\theta_{D_P^t}\right\}_{t=1}^{T}} \max_{\epsilon} \mathcal{L}_{D_P} = \min_{\left\{\theta_{D_P^t}\right\}_{t=1}^{T}} \max_{\epsilon} \frac{1}{K.T} \sum_{t=1}^{T} \sum_{k=1}^{K} \mathcal{L}_{D_P^t}\left(\hat{p}_t^k, p_t\right), \text{ s.t. } \epsilon \leq c_1 \tag{3.6}$$

where $\theta_{D_P^t}$ is demonstrates the parameters of discriminator model D_P, $\mathcal{L}_{D_P^t}$ represents the cross entropy loss function, $c1$ is a predefined privacy budget constraint, T is the number of private-attributes, $\hat{p}_t^k$ is the predicted $t-th$ private-attribute using $K-th$ sample and it is defined as follows:

$$\hat{p}_t^k = \text{softmax}\left(\tilde{\mathbf{z}}^k, \theta_{D_P^t}\right) \tag{3.7}$$

to calculate the predicted $t-th$ sensitive attribute using the $k-th$ sample. The outer minimization of equation (Eq.4.7) finds the strongest private-attributes inference at-

tack and the inner maximization seeks to fool the discriminator by obscuring private information.

3.4.5 DP_{BERT} - Learning the Text Representation

In the previous sections, we show how we: (1)add noise to prevent the adversary from regenerate the original text from the embedding representation and minimize the chance of privacy leakage by achieving differential privacy (Eq. 3.2), (2) control the amount of the added noise to preserve the semantic meaning of the textual information(Eq. 4.1), and (3)protect the private-attributes (Eq. 4.7). Inspired by the idea of adversarial learning, we model the objective function as a minmax game among the two discriminators D_P and D_S. Assume that $\mathcal{L}_{D_P^t}$ and $\mathcal{L}_{D_S}$ denotes cross entropy loss function for the private-attributes discriminator and cross entropy loss function for semantic task, respectively. Our goal is to maximization $\mathcal{L}_{D_P^t}$ that finds the strongest private-attributes inference attack and minimize $\mathcal{L}_{D_S}$ that measured by the incorrect label in the sentiment prediction. We can write the objective function as follows:

$$\min_{\theta_{D_S},\epsilon} \max_{\left\{\theta_{D_P^t}\right\}_{t=1}^{T}} \mathcal{L}_{D_S} - \alpha\mathcal{L}_{D_P^t} + \lambda\Omega(\theta) \quad \text{s.t.} \quad \epsilon \leq c_1 \tag{3.8}$$

where α controls the contribution of the private-attributes discriminator D_P in the learning process, and $\Omega(\theta)$ is the parameters regularizer.

The goal of this objective function is to learn the private text representation by adding a proper amount of noise to the latent representation to prevent the membership attack when the attacker is able to infer the existence of one instance in the training dataset. In addition, to prevent the reconstruction of the original text and inferring users' private information.

3.5 Experiments

3.5.1 Data

We use a dataset from TrustPilot from Hovy et al.Hovy *et al.* (2015). In this collected dataset, there are reviews on different products and ratings from one to five star. Each review is associated with three private-attributes: gender (male, female), age, and location (Denmark, France, United Kingdom, and the United States). The same approach of (Beigi *et al.*, 2019) is followed in this work. We follow the setting of (Hovy and Søgaard, 2015) by categorizing age attributes into three groups, above 45 years, under 35 years, and between 35 years and 45 years. We subsample 10k reviews for each location to balance the five locations. For the sentiment ground truth, we consider the review's rating score as a sentiment class.

3.5.2 Experimental Design

We compare DP_{BERT} with its variant $ORIGINAL_{BERT}$ which uses the original text representation without adding noise to it. We report accuracy score to evaluate the utility of the given text for a sentiment analysis (dos Santos and Gatti, 2014). Specifically, we use the rating score of each review for the sentiment prediction task. In addition, we report the examination of the text representation using F1 score for predicting the private-attributes. It is worth mentioning that the higher accuracy score for the semantic discriminator shows high utility for the sentiment task and a lower F1 score for the private-attributes discriminator demonstrates high privacy in the given text.

Since the maximum length of text for BERT is 512. We follow in this work the head-only method which keeps the first 512 tokens of the given text (Sun *et al.*, 2020). This head-only method considers that the important information that we need

to capture from the text will be in beginning of a document.

3.5.3 Experimental Result

We have conducted experiments to answer three questions:

- **Q1-** *Utility:* Does the learned text representation preserve the utility of the original text by keeping the same sentiment preserved?
- **Q2-** *Privacy:* Does the learned text representation hide the private information?
- **Q3-** *Utility-Privacy trade-off:* Does the trade-off between the utility and privacy reach the optimal point without sacrificing any of them?

The experimental results are demonstrated in Table 3.2. We compare DP_{BERT} with $ORIGINAL_{BERT}$, which is a variant of DP_{BERT} that publishes the original representation z without adding noise or utilizing the two discriminators.

Table 3.2: Accuracy for Sentiment Prediction and F1 for Evaluating Private Attribute Prediction Task. Higher Accuracy Shows Higher Utility, While Lower F1 Demonstrates Higher Privacy.

Model	Sentiment (ACC)	Private Attribute (F1)		
		AGE	Location	Gender
$ORIGINAL_{BERT}$	0.3870	0.4330	0.3072	0.5623
DP_{BERT}	0.2446	0.195	0.0671	0.4047

To emphasize, semantic discriminator D_S is applied to test data to predict the sentiment meaning where rating score is used as a label. Likewise, we apply the private-attributes discriminator D_P to mimic the attacker behavior who is trying to infer the private information about the individuals from the textual representation. For evaluation, a higher accuracy score for semantic discriminator D_S indicates high

utility for the given task, and lower F1 score for private attribute discriminator D_P demonstrates high privacy for individuals.

To answer the first question (Q1), we report experimental results for our proposed model using the sentiment analysis task. We predict sentiment of the textual information and measure the performance using the metric accuracy. The result of sentiment prediction for DP_{BERT} shows that the representation preserves the sentiment meaning of the textual data which means high utility. $ORIGINAL_{BERT}$ performs better than DP_{BERT} and the reason is that the first one uses the original text representation without adding noise to it which means high utility.

To answer the second question (Q2), we consider three different private information,i.e., age, location, and gender. We examine efficiency of the model in privacy protection using its performance in predicting values of different private-attributes. We measure performance of private-attributes predictor using F1 metric. Our proposed model has a remarkably lower F1 score which indicates higher privacy in terms of hiding the private-attributes. In addition, DP_{BERT} exceeds $ORIGINAL_{BERT}$ in terms of hiding the private information and does not sacrifice the utility significantly.

To answer the third question (Q3), we evaluate the utility loss against privacy improvement of the given text. DP_{BERT} has achieved a better trade-off results which are shown in high privacy and low utility loss comparing with $ORIGINAL_{BERT}$. $ORIGINAL_{BERT}$ achieves high accuracy in the sentiment task while suffering from significant privacy loss.

The results have shown that DP_{BERT} learns the textual representation of a given text that does not leak private information and preserve the semantic meaning of the text by achieving higher accuracy score for semantic discriminator D_S which indicates that representation has high utility for the semantic meaning, and lower F1 score for private-attributes discriminator D_P which demonstrates that the textual

representation has higher privacy for users due to obscuring their private information.

3.5.4 Parameter Analysis

DP_{BERT} has a parameter α which controls the contribution from the private attribute discriminator D_P. In this section, we investigate the effect of this parameter by varying it as: 0.125, 0.25, 1, 2, 4. We do the experiments and check the change of the accuracy for sentiment prediction for different values of α. Results are shown in the Fig 3.2. We can see the increase of α will decrease the accuracy of sentiment prediction task for DP_{BERT} and $ORIGINAL_{BERT}$. This shows that increasing the contribution of the privacy component leads to a decrease in the quality of the main task, i.e., sentiment classification. Moreover, The performance of private attributes discriminator for age, location, and gender is shown in Fig 3.3 using F1 score. Another observation is choosing $\alpha = 1$ will improve the accuracy of the sentiment prediction task and preserve privacy by keeping the F1 score low. This result shows the importance of the DP_{BERT}s privacy component in preserving users' privacy.

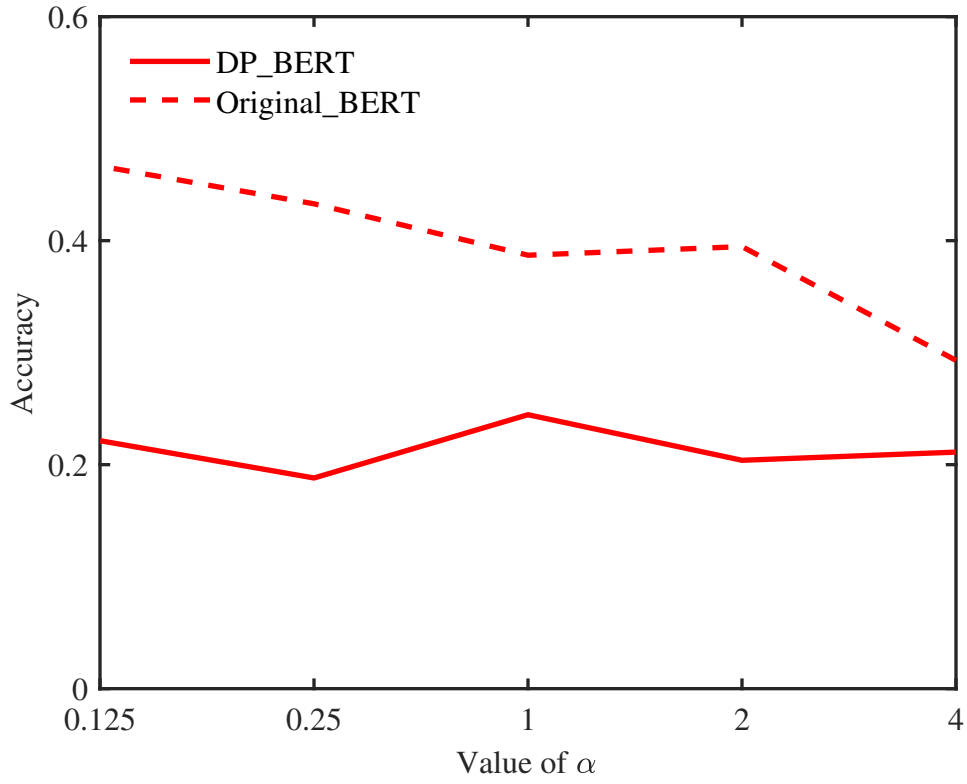

Figure 3.2: Performance Results For Sentiment Prediction Tasks For Different Values of α.

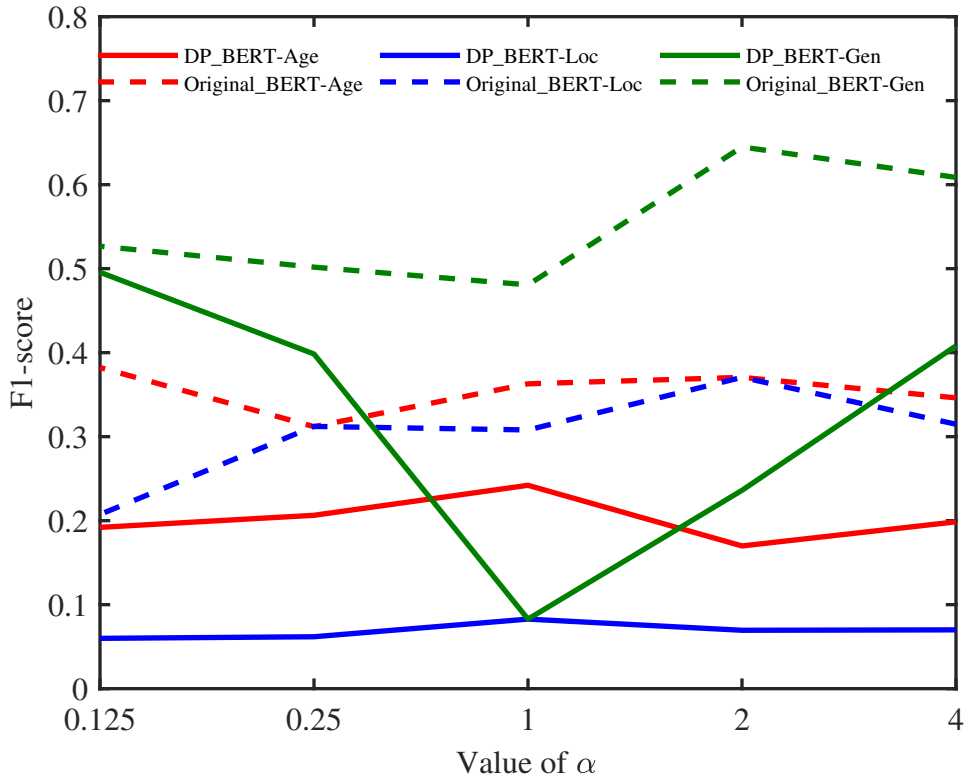

Figure 3.3: Performance Results For Private Attribute For Different Values of α. Higher Accuracy Shows Higher Utility, While Lower F1 Demonstrates Higher Privacy.

3.6 Conclusion

In this chapter, we proposed a privacy-preserving text representation learning framework, DP_{BERT}, which learns a text representation that is differential private, preserves users' private information, and maintains high utility by keeping the sentiment meaning of the given text. It has four main components which are BERT, differential-privacy-based noise adder, utility checking which is semantic meaning discriminator, and privacy checking which is private-attributes discriminator. Improving privacy comes at the cost of data utility which leads to trade-off between privacy and utility needs to be maintained. Our results showed the effectiveness of

our proposed framework, DP_{BERT}, in minimizing chances of privacy leakage of the private-attributes while preserving text semantic meaning in the same time.

Chapter 4

PPSL: PRIVACY-PRESERVING TEXT CLASSIFICATION FOR SPLIT LEARNING

4.1 Introduction

Distributed Collaborative Machine Learning (DCML) has become one of the trend research topics for its importance and impact on individuals and organizations. It has been applied in various domains such as healthcare and finance. In contrast with the centralized approach, where the data is centrally pooled from different resources and the training process is centralized, distributed collaborative machine learning enables the training process across different parties. DCML improves the privacy in different applications and under privacy regulations such as HIPPA and GDPR.

One of the latest popular DCML approaches is split learning Gupta and Raskar (2018) which divides a neural network into two or more sub-networks and trains them separately on different parties as illustrated in Figure 1.1, where the whole network is divided between Client A and the server. The first part of the network is trained on Client A on the raw data locally, and there is no need to share them with any other parties. The second part of the network will continue the training on the server-side. Accordingly, the server and other parties have no access to the client's raw data. Recently, the interest in split learning is growing (Gao *et al.*, 2020) (Poirot *et al.*, 2019) (Singh *et al.*, 2019) (Vepakomma *et al.*, 2018). Split learning achieves three main advantages among DCML approaches, which are (1) privacy protection by keeping the raw data on the client-side; (2) comparable model accuracy with centralized machine learning models; and (3) computational work is reduced on the

client side since it needs to train some layers of the network Gupta and Raskar (2018). Due to these features, there have been various research works on split learning from different aspects, including attacks (Pasquini *et al.*, 2021), defenses (Li *et al.*, 2021), and performance (Madaan *et al.*, 2021).

Split learning comes into the picture of DCML approaches to overcome the weakness of federated learning in some applications(McMahan *et al.*, 2017). In order to achieve optimal accuracy in some applications that require large computational resources and to obey the ethics and the regulation regarding sharing and utilizing the training data, split learning was proposed to enable the training in multiple entities (Gupta and Raskar, 2018). However, privacy issues need to be addressed in split learning, particularly when the data is very sensitive and could identify individuals' identities or attributes. With the split learning approach, privacy issues are caused by exchanging the intermediate parameters among the parties. In the last few years, there has been significant researches toward improving the privacy of split learning while maintaining the utility for different tasks by combining different privacy-enhancing mechanisms such as differential privacy(Abuadbba *et al.*, 2020) and distance correlation minimization(Vepakomma *et al.*, 2020).Split learning has been applied to different kinds of datasets such as sequential data (Abuadbba *et al.*, 2020), image data(Vepakomma *et al.*, 2020), and users behavior dataset(Li *et al.*, 2021). In this work, we apply split learning to a text dataset to build privacy-preserving text classification.

Text data is rich in the information that used in different applications such as recommendation systems and understanding users' behavior. However, it might cause privacy leakage since it contains implicit sensitive information about individuals (Beigi *et al.*, 2019)(Alnasser *et al.*, 2021) (Mosallanezhad *et al.*, 2019). Several privacy issues related to text data have been addressed, such as re-identification and private

attributes inference attacks. Re-identification could be tackled by traditional protection techniques such as k-anonymity (Sweeney, 2002) and differential privacy (Dwork, 2008). Despite that, these techniques have shown inefficiency in preventing private attributes leakage because of the nature of the user-generated text data, which is unstructured and informal. The risk of leakage of private information is even higher in decentralized NLP in split learning if the attacker has access to the intermediate exchangeable parameters. This arises the need to protect users against leakage of private information in this setting. Note that in this paper we assume that the user's client-side environment is trusted and an attacker can only have access to the intermediate parameters.

As a first study towards exploring the feasibility of split learning to deal with text data, we adopt a sentiment classification approach using a text dataset collected from social networks. Figure 4.1 illustrates split learning with multiple entities contributing to the training process. Considering the fact that text data may reveal individuals' private attributes that they do not want to share (Alnasser *et al.*, 2020a), training the model on split learning setting and using adversarial learning as a protection technique would improve the privacy of the classifier.

This work is devoted to investigating and answering the following research questions:

Q(1): Can split learning be applied to text data in sentiment classification to achieve comparable model accuracy as a centralized classification model?

To answer this question, we explore the feasibility of split learning to deal with text data by building a sentiment classification model. To the best of our knowledge, this is the first study on split learning using text data.

Q(2): How does adversarial learning minimize the private information

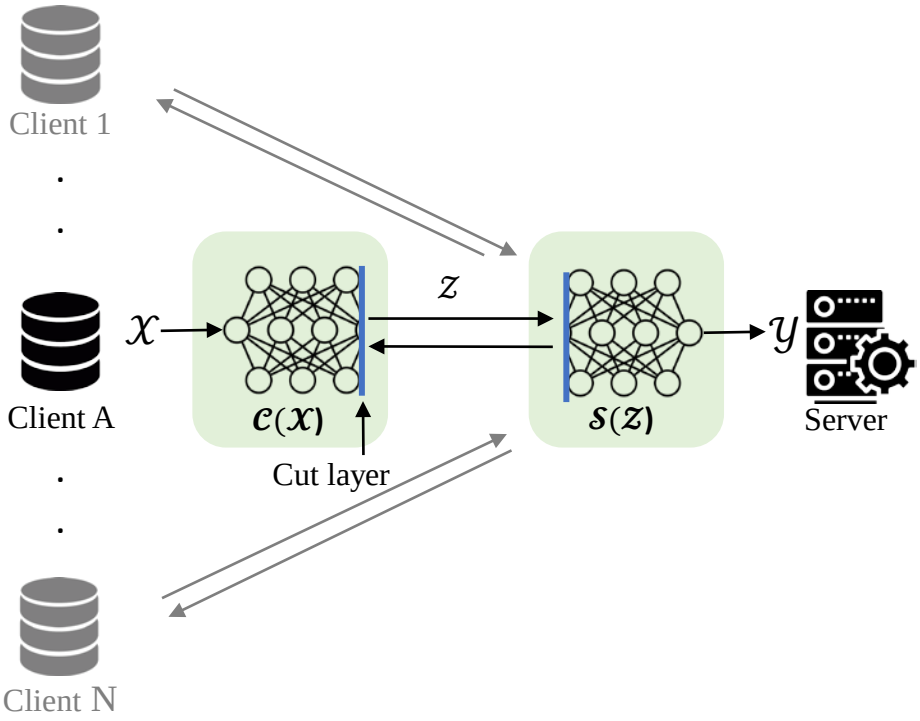

Figure 4.1: Split Learning Overview

leakage of sentiment classifier on split learning setting?

To answer this question, we explore the privacy leakage of the classifier on split learning by training discriminators to predict private information using the exchangeable variables between the entities.

Q(3): What are the impacts of increasing the number of hidden layers or training time on the performance and the privacy of split learning?

Then, we investigate these two strategies and study their impacts on the performance and the privacy of split learning.

The main contributions of this work are:

- We explore the feasibility of split learning to deal with text data.
- We propose a novel privacy-preserving text classification framework using split

learning that protects the private attributes.

- We perform an experiment on a real-world text dataset to show the efficiency of our proposed framework. We show the trade-off of privacy with the utility of the data.

The rest of the chapter is organized as follows: Section 4.2 presents the related work. Section 4.3 proposes the problem statement. 4.4 proposes the PPSL framework. Section 4.5 details our experiments. Section 4.6 discusses the experimental results, followed by the conclusion in Section 4.7.

4.2 Related Work

4.2.1 Distributed Collaborative Machine Learning

Distributed collaborative machine learning approaches have been adopted in different applications instead of the centralized approach for several reasons, such as managing the computational load and improving the privacy of the training data. Federated learning approach, which was introduced by (McMahan *et al.*, 2017), allows the training process without sharing the raw data by sending the whole model to the parties (data owners). The design principle of federated learning supposes that the neural networks' weights are exchangeable during the training process. Albeit there are core challenges related to federated learning in some applications, expensive communication, model privacy compromised, and privacy issues (Li *et al.*, 2020). As a solution to the previous challenges, split learning has emerged as a new distributed machine learning approach (Gupta and Raskar, 2018). The basic idea is to split up the model into multiple portions and execute each portion at different parties. Each party has different privileges regarding accessing the raw data. In vanilla split learn-

ing, where the model is divided between a client and a server, the client train a part of the neural network that will access the raw data while the other part will be held in the server. The server will only receive the intermediate output from the split layer (Gupta and Raskar, 2018) (Poirot *et al.*, 2019) (Vepakomma *et al.*, 2018). Split learning has been applied on sequential data (Abuadbba *et al.*, 2020) such as ECG signals, image data (Vepakomma *et al.*, 2020) such as UTKFace and CIFAR10, and users' click records (Li *et al.*, 2021) such as Avazu and Criteo.

Federated learning and split learning are both distributed collaborative machine learning, and there are two main differences related to (1) model architecture, and (2) computational resources needed. In federated learning, the whole model is transferred to the client-side to do the training process on the local data and communicate with the server with the updated model parameter. On the server-side, the aggregated updated parameters will be combined together using different approaches. In split learning, the neural network is divided into two or more parts to do the training process. The simplest split is a vanilla split where the neural network is divided into two parts, some of the beginning layers will be on the client-side, and the remaining layers will be on the server-side. The training process is started from the client where the data is stored locally, and the communication between the client and the server will be through the output of the last layer on the client side, which is called the cut layer. The second difference is the computation resources needed in federated learning and split learning. The former needs considerable computational resources on the clients' side since the whole model is transferred to them. On the other hand, the latter needs fewer computational resources on the clients' side since part of the neural network is transferred (Vepakomma *et al.*, 2018).

4.2.2 Privacy in Split Learning

Even though split learning improves the privacy by training part of the model in the data locally on the client, there is a leakage coming from the exchange of the intermediate output between the parties in the split setting. Various works have been proposed to improve privacy while maintaining accuracy by combining several protection techniques with split learning. One work (Abuadbba *et al.*, 2020) adapted two privacy mitigation techniques to address the leakage in the split learning. The first technique adds more hidden layers to the part of the neural network on the client-side, and the second technique applies differential privacy. Other work proposed a method against the reconstruction attack by minimizing the distance correlation between raw data and the intermediate output of the split layer (Vepakomma *et al.*, 2020). Their method adds an additional loss function called distance correlation with categorical cross-entropy commonly used loss function.

On the other side, numerous works propose attacks in the split learning setting, which can leak privacy. First, Norm attack, which is a method that uses the norm of the exchangeable gradients in split learning setting (Li *et al.*, 2021). Using this method, the attacker will uncover the data labels that might be sensitive and supposed to be private. Another attack is called reconstruction attack (Vepakomma *et al.*, 2020) which is a neural network taking the intermediate activation as an input, and the output is the generated data. The attacker is located at any party that receives the intermediate activation to reconstruct the raw data. Other work analyzed the security vulnerability in split learning and introduced the feature-space hijacking attack (FSHA), which showed how the server could obtain the training data (Pasquini *et al.*, 2021). This attack happens when the server exploits its control on the learning process, leading to changing specific properties in the intermediate

activation generated by the clients, inference, or reconstruction attacks.

4.2.3 Text Data and Private Attributes

Text data may cause a privacy leakage since it includes private attribute information about users, which can be defined as the information that individuals do not want to explicitly disclose, such as gender, occupation, location, and age. Several methods are reported in the literature to address how to preserve the text information privacy (Beigi *et al.*, 2019)(Alnasser *et al.*, 2021) (Mosallanezhad *et al.*, 2019). Recent works review different aspects of user privacy and compare traditional privacy models for protecting user private attributes (Alnasser *et al.*, 2020a) (Beigi and Liu, 2020).

4.3 Problem Statement

PROBLEM 1. *Let N be the number of parties (devices) and each party denoted as $Client_i$ where $i \in [1, N]$ and a supercomputing $Server$. Each party has a set of K documents as $X = \{x_1, x_2, ..., x_k\}$ and each document x_i composed of a sequence of words. Let P be a set of private attributes associated with each document and T is a downstream task (i.e. classification). We would like to use x_i in the given task T. However, we want to preserve users' privacy by preventing a potential adversary from inferring the users' private attribute information. We define the problem as building a distributed privacy-preserving text classifier among the parties and the server so that 1) the first part of the model is trained on the $Client$ and the second part on the $Server$, 2) the potential adversary cannot infer the targeted user's private attributes P from any intermediate variables Z that are exchanged between the client and the server in the training process, and 3) the utility of the given task T is preserved.*

Note that in this work, we assume that the potential adversary can only have access to the exchangeable data Z and not any other information. Moreover, the goal

is to achieve a protection against possible private-attribute inference attacks, but not against other types of attacks such as reconstruction attacks.

4.4 Framework Architecture

Overview: our goal is to design a privacy-preserved text classification framework on split learning setting where the raw text data on the client-side has never been shared with any other entity in the environment. In this context of split learning, the training process will be divided between the client, and the server, and the intermediate activations are communicated between these parties. The simplest form of split learning will be used where we have a single party and supercomputing resource, server (Gupta and Raskar, 2018). In order to protect private attribute information, we follow the idea of adversarial learning. In our proposed model, the model on the client-side includes two major components, 1) text classifier, and 2) private attribute discriminator. The goal of the text classifier component is to perform the given downstream task. One part of this component is trained on the client-side and the other part is trained on the server-side. Intermediate variable Z is also exchanged accordingly between the client and the server. The private attribute discriminator component ensures that the intermediate variable Z does not contain private attribute information. In particular, since the system does not know the malicious attacker's model, this component has been added to mimic the behavior of a potential malicious attacker. The private attribute discriminator component seeks to accurately infer users' private attribute information from the intermediate variable Z. This component could be leveraged in the adversarial learning process to regularize the way Z is learned by incorporating necessary constraints in order to fool the adversary component and further avoid the leakage of private attributes from Z. These two components are discussed in details in the next sections. Figure 4.2 illustrates PPSL

framework.

Note that in our work, we assume the client is a trusted environment, and any other entities are not trusted. Also, we assume the attacker will be an external attacker who will have access to the exchangeable parameters only.

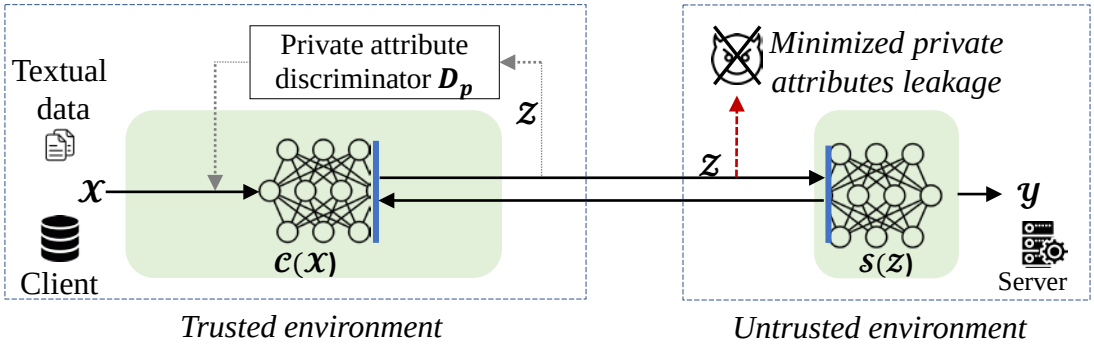

Figure 4.2: The Framework of PPSL Architecture

4.4.1 Text Classifier

First, BERT Base Uncased model (Devlin *et al.*, 2019) is used to extract the embedding of the text data. Then, we utilize the gated recurrent unit (GRU) as a cell type of Recurrent Neural Network. Then, we concatenate the two hidden state outputs of the GRU. We define the loss function of the classifier as follows:

$$\mathcal{L}_{\mathcal{D}_S} = \sum_{i=1}^{G} -y(i) \log \hat{y}(i) \tag{4.1}$$

where G is the number of classes, y is the ground truth label for the classification task, and $\hat{y}(i)$ represents the inferred label of the $i-th$ element.

The classifier will be built on a split learning setting where the client has the model's first layers and the server controls the remaining layers. In the training process, the client will start the training on the first layers of the neural network and

send the intermediate variable Z to the server as follows:

$$Z = C(X) \tag{4.2}$$

Then, the server will propagate the Z through the remaining layers as follows:

$$Y = S(Z) \tag{4.3}$$

4.4.2 Private Attribute Discriminators

We follow the idea of adversarial learning by training private attributes discriminators D_p for age, location, and gender. We want the intermediate variable Z to predict Y with high accuracy and poor accuracy to predict the private attributes. The adversary learning can be formally written as:

$$\min_{\left\{\theta_{D_P^t}\right\}_{t=1}^{T}} \max_{\alpha} \mathcal{L}_{D_P} = \min_{\left\{\theta_{D_P^t}\right\}_{t=1}^{T}} \max_{\alpha} \frac{1}{K.T} \sum_{t=1}^{T} \sum_{k=1}^{K} \mathcal{L}_{D_P^t} \left(\hat{p}_t^k, p_t\right) \tag{4.4}$$

where θ_{D_P} demonstrate the parameters of the private attributes discriminator, and T demonstrates three different private attributes (gender, age, location) and K instances. α is scalar weight to control the privacy. p_t and $\hat{p}$ are the true and the predicted private attribute. The goal of the outer minimization is to find the strongest private attribute inference attack and the goal of the inner maximization is to fool the discriminator by obscuring private information. The predicted private attribute $\hat{p}_t^k$ is defined as:

$$\hat{p} = softmax(z, \theta_{D_P}) \tag{4.5}$$

Regarding the split setting, the total loss function will be calculated for the whole neural network as follows:

$$\alpha_1 \, \mathrm{CCE}(p, \hat{p}) + \alpha_2 \, \mathrm{CCE}\,(y, \hat{y}) \tag{4.6}$$

Where: α_1 and α_2 are scalar weights to control privacy and utility. CCE is categorical cross-entropy loss function that will be used.

The objective function that we want to achieve is:

$$\min_{\theta_{D_S}} \max_{\{\theta_{D_P}\}_{t=1}^{T}} \frac{1}{K.T} \sum_{t=1}^{T} \sum_{k=1}^{K} \mathcal{L}_{D_S}(\hat{y}, y) - \alpha \mathcal{L}_{D_P^t}\left(\hat{p}_t^k, p_t\right) \tag{4.7}$$

where θ_{D_S} and θ_{D_P} demonstrate the parameters of the sentiment analysis and the discriminators, respectively, $\mathcal{L}_{D_S}$ and $\mathcal{L}_{D_P}$ represent the cross-entropy loss functions. y and $\hat{y}$ are sentiment true and the predicted label. α is the the scalar weights to control the contribution of the private attribute discriminator in the learning process. This objective function seeks to minimize the privacy leakage by maximizing the loss in the private attribute discriminator and minimizing the loss in the sentiment classification task.

4.5 Experiments

This section aims to answer the following research questions:

1. Is the proposed PPSL framework achieving high accuracy in the downstream task, i.e., classification?

2. How is the trade-off between privacy with utility maintained?

3. How does using adversarial learning improve the privacy of the proposed framework, PPSL?

4.5.1 Dataset

We use a dataset from TrustPilot from Hovy *et al.* (2015). In this collected dataset, there are reviews on different products and ratings from one to five stars. Each review is associated with three private attributes: gender (male, female), age, and location (Denmark, France, United Kingdom, and the United States). We follow the setting of (Hovy and Søgaard, 2015) by categorizing age attributes into three groups, above 45 years, under 35 years, and between 35 years and 45 years. For the sentiment ground truth, we consider the review's sentiment positive if its rating score is 4 or 5 and consider it as negative if the rating is 1, 2, or 3.

4.5.2 Experimental Design

In this work, we follow the simple vanilla split learning configuration (Vepakomma *et al.*, 2018). In addition we do the experiment with one client and the server. The parameters α and λ are determined as $\alpha = 0.5$ and $\lambda = 0.01$. The optimization algorithm used in the sentiment analysis classifier and the private attribute discriminators is SGD. The batch size we use in the experiments is $b = 32$. We report the accuracy score to evaluate the utility of the proposed framework, PPSL, for downstream task sentiment analysis in split learning after improving the privacy using adversary learning. Specifically, we use the rating score of each review instance to be the label. In addition, we report the accuracy of the three private attributes discriminators. It is worth to mentioning that the lower accuracy for the discriminators shows high privacy since it shows, it becomes hard for the attacker to infer the private information. The higher accuracy for the sentiment classification means high utility.

4.5.3 Experimental Results

The proposed framework, PPSL, is compared with:

- **ADV-ALL:** This method utilizes a generator and a discriminator to create a text representation that has high utility for a given task while preserving privacy by protecting the private attributes (Li *et al.*, 2018).
- **Original:** This is a sentiment classification on split learning setting with out using the adversarial learning as a privacy protection technique.

The experimental results are demonstrated in table 4.1.

Table 4.1: Experimental Results. Higher Sentiment Accuracy Values Show Higher Utility, While Lower Private Attribute Accuracy Indicated Higher Privacy.

Model	**Sentiment Acc**	**Private Attributes Acc**		
		Age	***Loc.***	***Gen***
ADV-ALL	89.12%	31.67%	45.90%	50.23%
Original	90.69%	65.97%	76.81%	63%
PPSL	**89.46 %**	**29.50%**	**44.76%**	**49.43%**

4.6 Discussion

The privacy preservation capabilities of our proposed model, PPSL, are due to two reasons. First, it conducts split learning which is one of the distributed machine learning where the training data is kept locally on the client, and the training process is divided between the client and the server. In split learning, not only the training data is protected, the model architecture is protected compared with federated learning, where the model needs to be sent to all the clients. Second, it utilizes the idea of adversarial learning to prevent private attribute inference attacks. Defining private

attribute discriminators helps mimic the attacker's behavior when he utilizes the exchangeable parameters between the privacy and utility to infer individuals' private information.

We have conducted experiments to answer three questions:

- **Q1-** *Utility:* Does the proposed framework preserve the utility?
- **Q2-** *Privacy:* Does the proposed framework improve privacy by minimizing the private attributes leakage?
- **Q3-** *Utility-Privacy trade-off:* Does the trade-off between the privacy and the utility maintain well without sacrificing any of them?

To answer the first question (Q1), we report experimental results for our proposed framework, PPSL, using the downstream task sentiment analysis. We predict the sentiment of the reviews and measure the performance using metric accuracy. The sentiment analysis result for PPSL is comparable with the *Original* baseline, which is a sentiment classifier in a split learning setting without an adversarial learning technique. *Original* performs better in terms of sentiment accuracy than PPSL. The reason is that *Original* is a sentiment classifier in a split learning setting without any protection techniques that affect the utility.

To answer the second question (Q2), three different private attributes are considered in our experiments, i.e., age, location, and gender. We measure the efficiency of the framework in privacy protection using its accuracy in predicting the values of these attributes. We measure the performance of the private attributes discriminators using the accuracy metric. Our proposed framework, PPSL, has a remarkably lower accuracy than the *Original* baseline, which indicates the privacy improvement by minimizing the private attribute leakage. In addition, PPSL exceeds

$ADV - ALL$ in terms of hiding the private attributes. To elaborate, the intermediate variable Z has been chosen to give a poor accuracy in predicting the private attributes. The accuracy of the private attribute discriminator to predict the age in $ADV - ALL$ is 31.67%, and it reduced to be 29.50% in our proposed model PPSL. While in *Original* model, the accuracy of predicting the age is 65.97%. This improvement in privacy indicates that even though *Original* is a classification model in a split learning setting, it causes a privacy leakage that needs to be addressed. Utilizing adversarial learning has successfully improved the privacy in our proposed framework, PPSL.

To answer the third question (Q3), we evaluate both the utility and the privacy to make sure that the trade-off is maintained well in the proposed model, PPSL. It is clearly shown that PPSL has achieved better trade-off results which are indicated in high privacy (low accuracy for the private attribute discriminators) and low utility loss (high accuracy for the sentiment analysis). Improving privacy comes with the cost of distorting the utility, so maintaining the privacy and utility trade-off is one of the objectives of the proposed model, PPSL. Utilizing the adversarial learning helps us to choose the exchangeable parameter, Z, to give a poor accuracy in predicting the private attributes and good accuracy for the classification task.

The experimental results have shown that our proposed framework, PPSL, achieves high accuracy in sentiment classification in a split learning setting where the neural network is divided between the client and the server while preserving the private attributes information.

We do further investigation and study the impact of the training time on the privacy, particularly in split learning. We measure the accuracy of the proposed framework, PPSL, after each epoch. As illustrated in Fig 4.3, the sentiment accuracy converged to 99.06% around 200 epochs. In addition, we show the accuracy of the

three discriminators, age, location, and gender in Fig 4.4. It is shown that gender and location discriminators have lower accuracy when the number of epochs increases, which means high privacy. It is shown in the literature that some datasets and models are achieving better performance at higher epochs number of training (Thapa *et al.*, 2020). However, some scenarios require limiting the number of training to satisfy the cost limitations. On the other hand, age discriminator accuracy has increased when the training time has increased. As possible explanation of this, neural networks memorize the information in the training data (Feldman and Zhang, 2020). There is a privacy concern in the memorization from a privacy perspective, especially when the data is sensitive. Membership attacks, for example, take advantage of the memorization by looking at the prediction of the model and inferring the existence of the data sample in the training data.

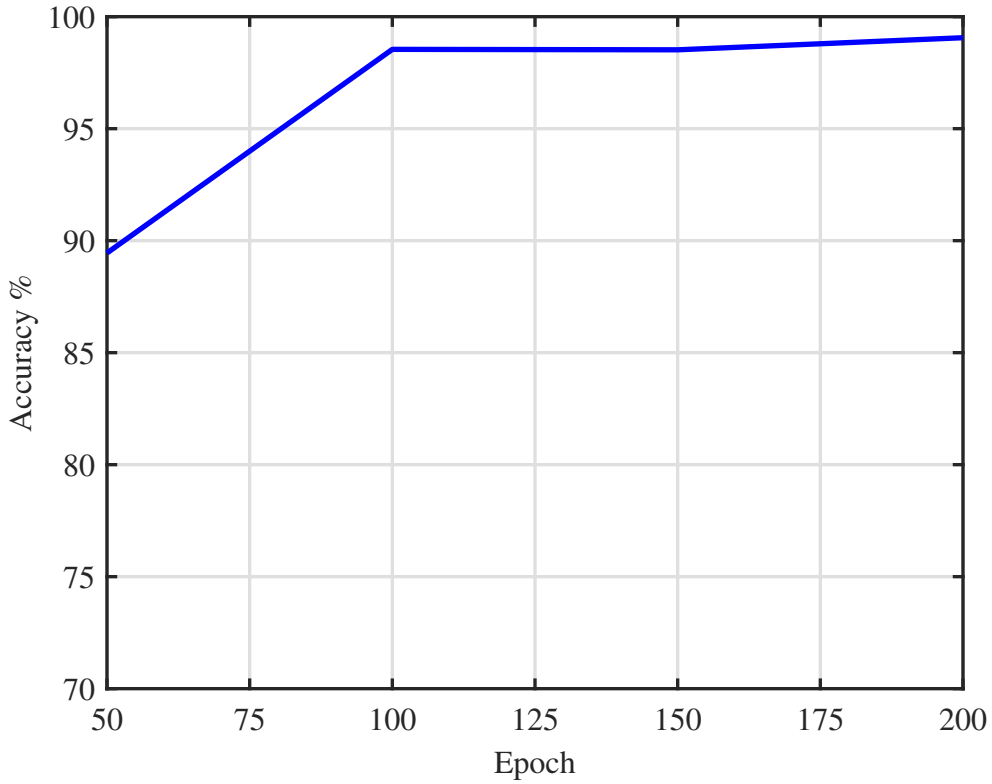

Figure 4.3: Sentiment accuracy over the training time

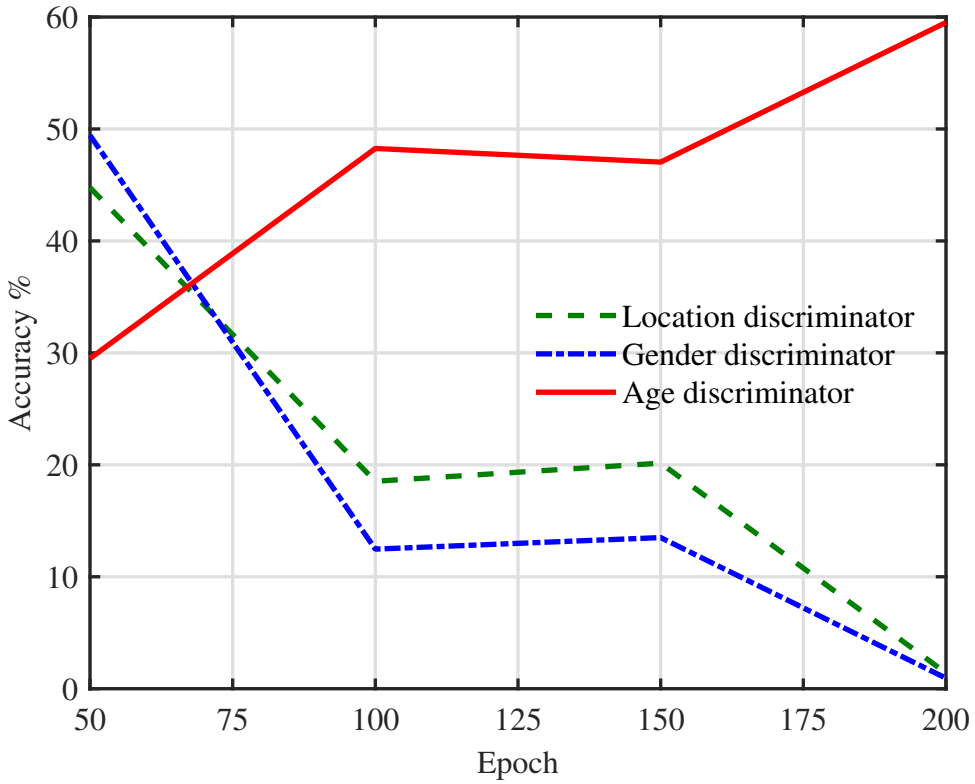

Figure 4.4: Private Attributes Discriminators Accuracy Over The Training Time

4.6.1 Parameter Analysis

Our proposed framework, PPSL, has a parameter α that controls the contribution of the privacy discriminator component. We do the experiments and check the effects of different values of α by investigating these values: 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. Figure 4.5 shows the accuracy of the sentiment prediction and gender discriminator for different values of α. We can see clearly that the increase of α will increase the accuracy of sentiment, which means high utility. At the same time, it decreases the gender discriminator's accuracy, which means minimizing the gender information leakage. Similar performances are shown in Figure 4.6 and Figure 4.7 for location and age, respectively. We observe that setting $\alpha = 0.5$ will improve the sentiment

classification accuracy and preserve the private attributes' privacy by minimizing the leakage.

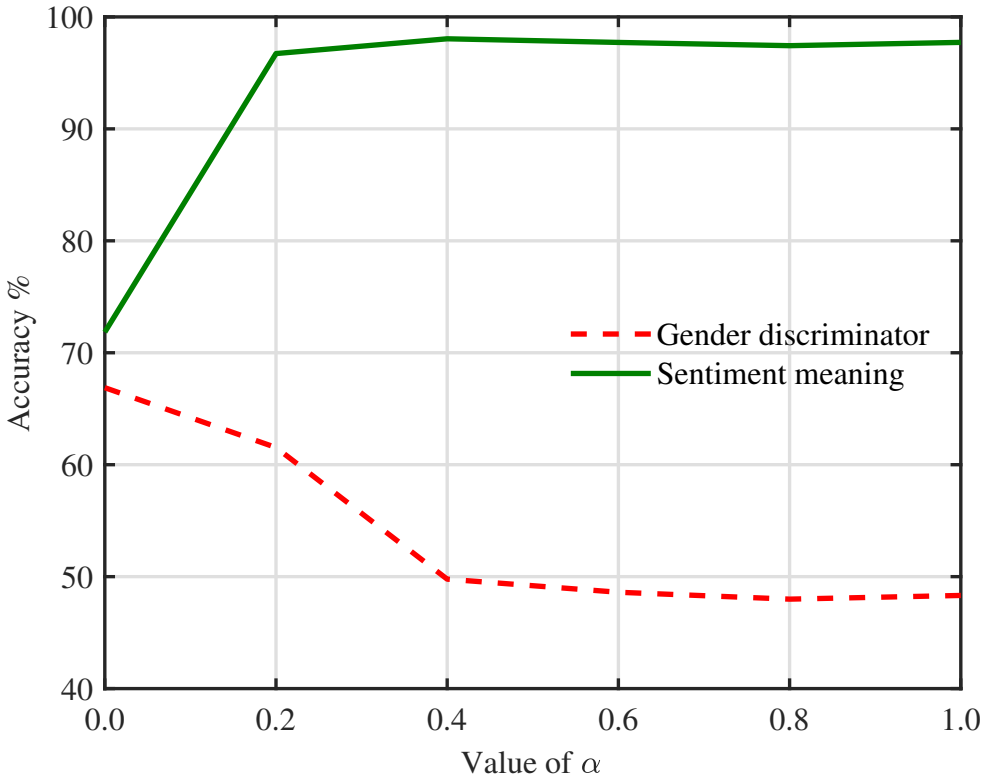

Figure 4.5: Gender Discriminator and Sentiment Meaning Accuracy For Different Values of α

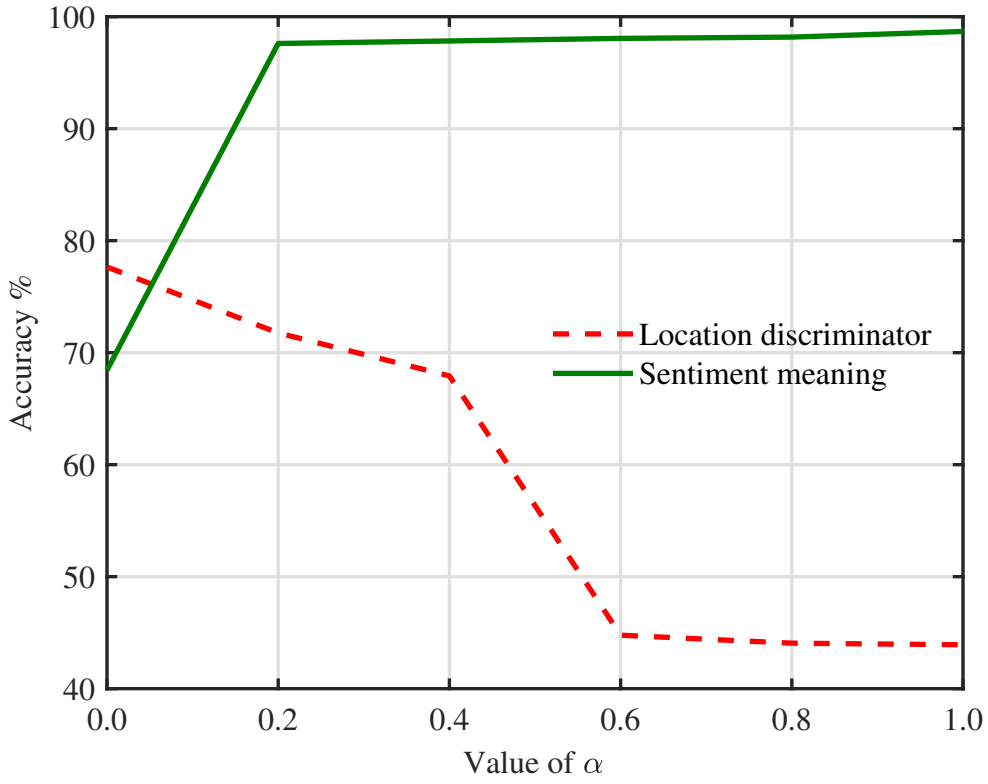

Figure 4.6: Location Discriminator and Sentiment Meaning Accuracy For Different Values of α

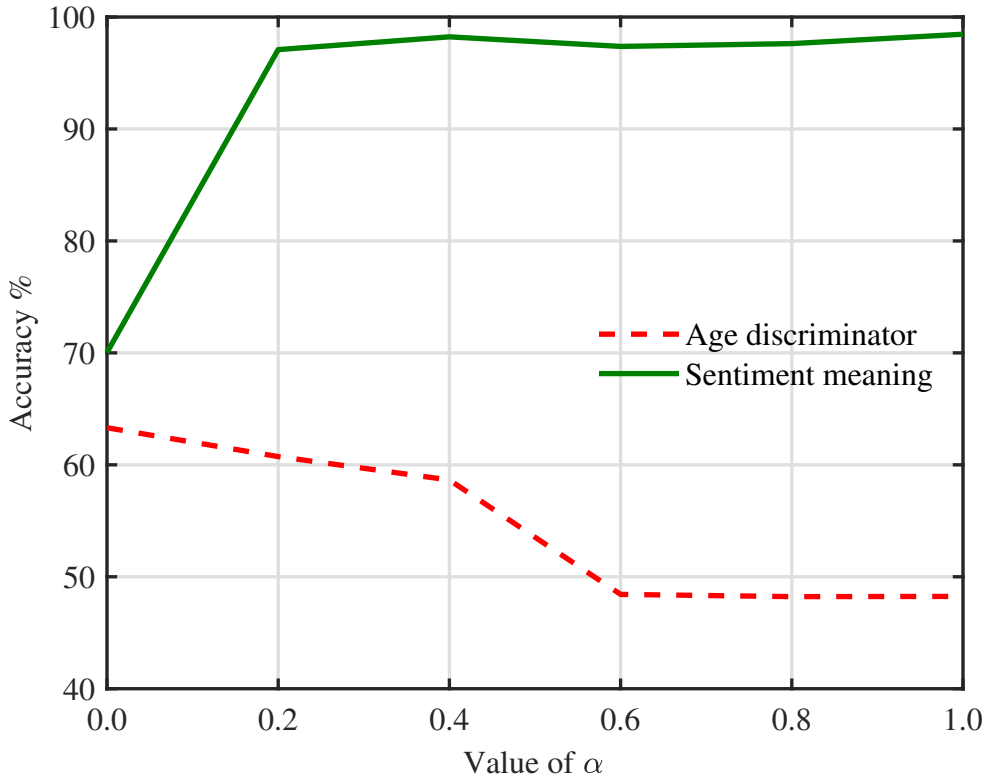

Figure 4.7: Age Discriminator and Sentiment Meaning Accuracy For Different Values of α

4.6.2 Adding More Hidden Layers

Improving the network complexity by adding more hidden layers impacts the performance and computational resources. We investigate how increasing the number of the hidden layers on the client-side before the split layer will affect the performance of our proposed framework, PPSL. Assuming that the number of layers in the server is constant during the experiment, we show the impact of increasing the number of hidden layers starting from 2 to 10 layers. Figure 4.8 illustrates how the gender discriminator and sentiment meaning accuracy change with the change of the model architecture, specifically with the increasing of the hidden layers. Sentiment

classification accuracy decreases when the number of hidden layers is increased. On the other hand, the privacy leakage is decreased which is measured by the private attribute discriminators accuracy. As the accuracy of predicting the private attribute decreased, the privacy improved. The accuracy of the gender discriminator reduced from 58% to 48% when we increased the number of hidden layers from 2 to 10 layers. This finding demonstrates that in some applications when the data is highly sensitive, increasing the number of the hidden layers will minimize the privacy leakage (Abuadbba *et al.*, 2020).

For the location private attribute, Figure 4.9 shows the accuracy of the location discriminator and sentiment meaning. The sentiment meaning accuracy is dramatically decreased when the number of the hidden layers is increased. On the other hand, privacy is slightly improved by the increasing number of the hidden layers.

Figure 4.10 shows the accuracy of age discriminator and sentiment meaning. The sentiment meaning accuracy is dramatically decreased when the number of the hidden layers is increased. On the other hand, privacy is slightly improved by the increasing number of the hidden layers.

Overall, the accuracy decline is acceptable until having three hidden layers that retain the accuracy above 90%. However, an increasing number of layers on the client-side will require more computational resources, which need to be addressed in some domains. On another side, privacy is improved with the increasing number of hidden layers. The reason behind this improvement in the privacy is that as we add more layers, the dependency between the raw data and the exchangeable intermediate output of the split layer is reduced.

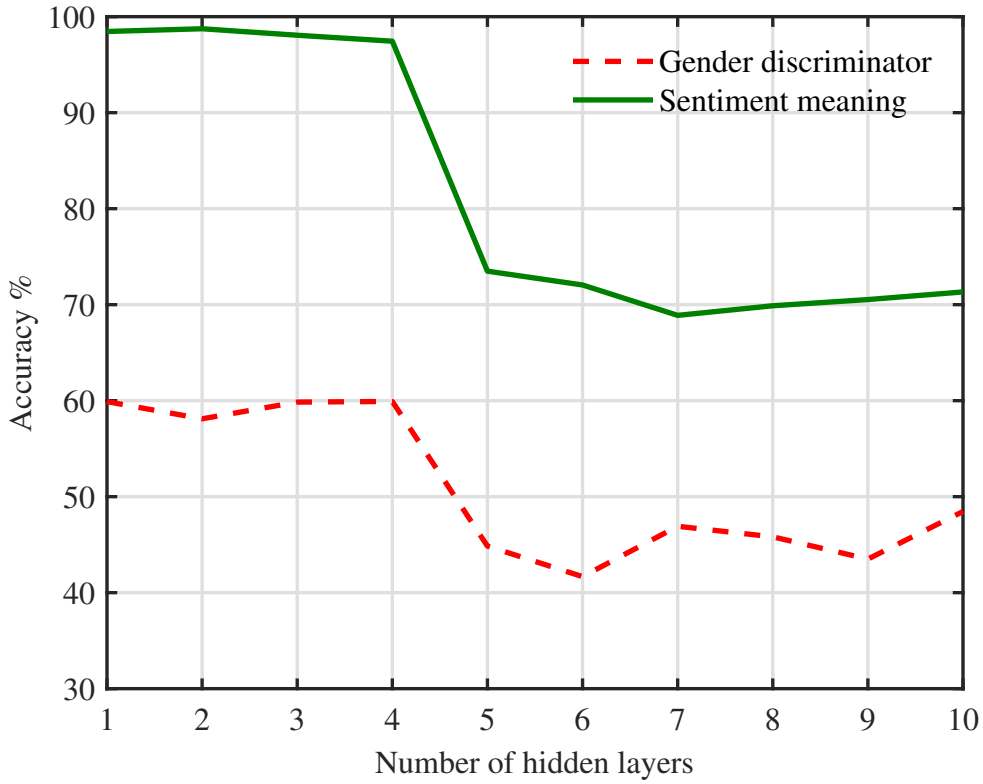

Figure 4.8: Gender Discriminator and Sentiment Meaning Accuracy With The Increase Number of The Hidden Layers

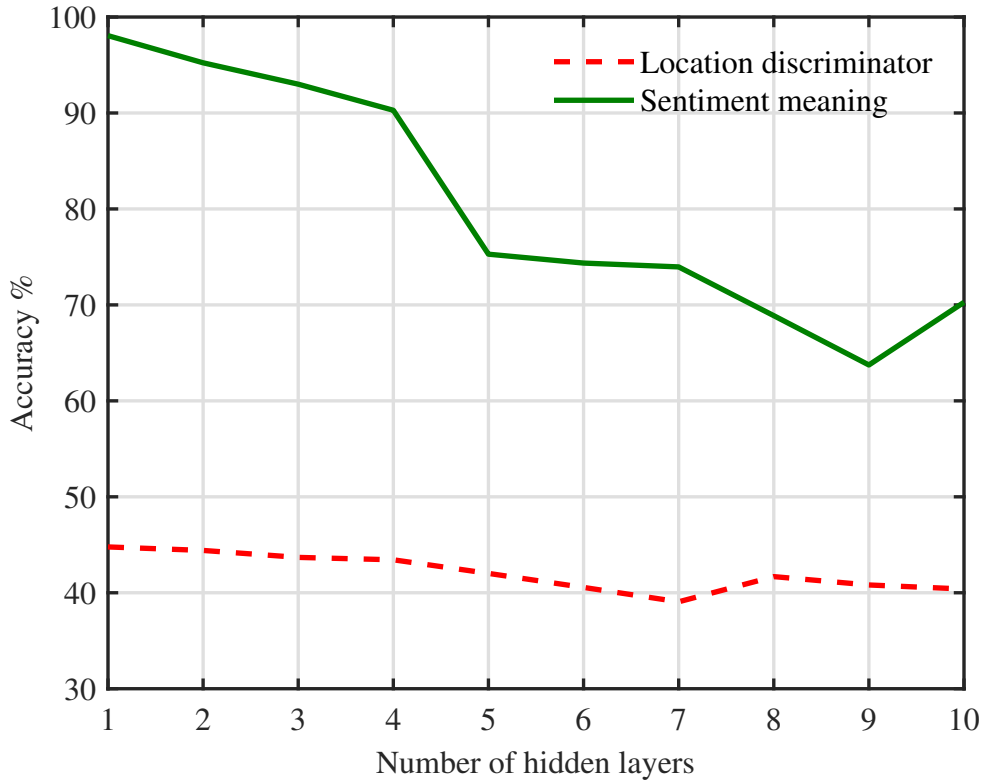

Figure 4.9: Location Discriminator and Sentiment Meaning Accuracy With The Increase Number of The Hidden Layers

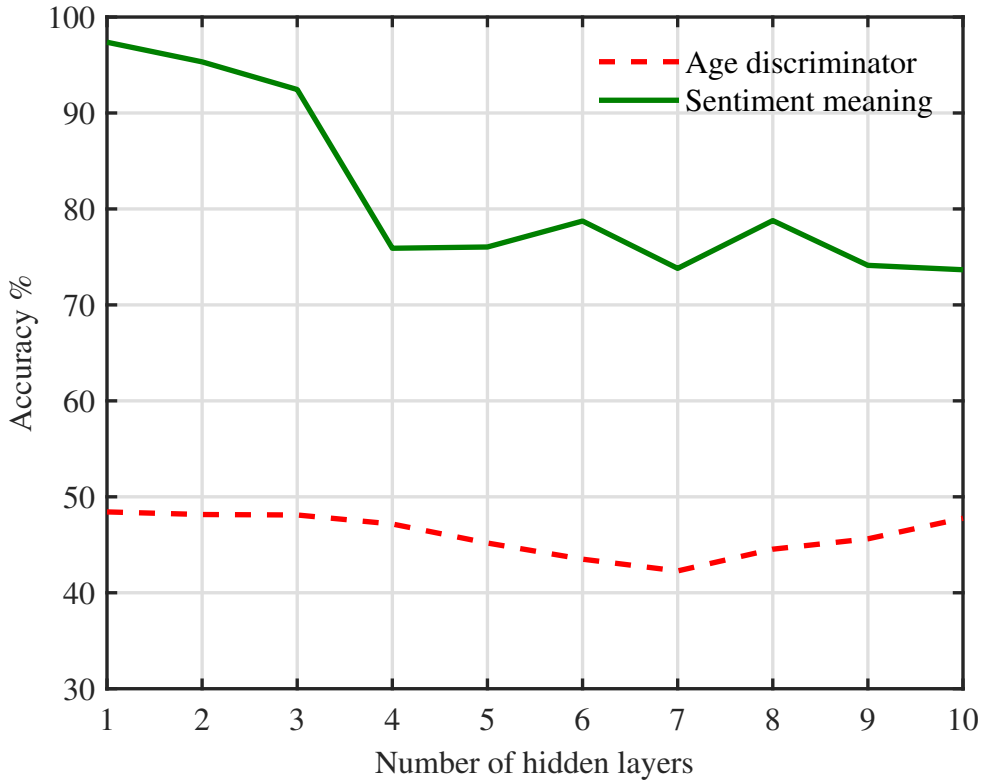

Figure 4.10: Age Discriminator and Sentiment Meaning Accuracy With The Increase Number of The Hidden Layers

4.7 Conclusion

In this chapter, we propose a privacy-preserving text classification for split learning, PPSL, which is a sentiment classification that protects the individual's private attributes by using adversarial learning. We explore the feasibility of split learning to deal with the text dataset. Our results show the effectiveness of the proposed framework, PPSL, in preserving the utility of a sentiment analysis task and improving privacy by minimizing the private attributes leakage. We evaluate the performance of our proposed framework, PPSL, in terms of model accuracy in sentiment classification as a measure of the utility and the accuracy of the private attributes discriminators

as a measure of privacy leakage. Future research can be directed towards applying different privacy techniques in split learning settings and investigating each protection technique's impact. Another future direction is to explore the impact of the number of clients on the performance of the proposed model. Last but not least, another extension to this work could be considering the implicit dependency between the private attributes and protecting against that.

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Summary

In the first two chapters of this dissertation, we introduced and proposed the problem of improving users' privacy in machine learning and reviewed the literature from two points of view. First, existing private attribute attacks are reviewed which can be classified to: (1) friend-based private attribute inference, (2) behavior-based private-attribute inference, and (3) both friend-based and behavior-based inference. Second, different privacy protection techniques have been proposed and discussed pointing out two main categories to protect individuals' private information: (1) traditional protection techniques and (2) adversarial learning. One of the earliest and powerful traditional protection techniques is differential privacy which provides a guarantee that the behavior of the dataset will hardly be affected after a single instance is added or removed from the dataset. Anonymization techniques are considered privacy protection techniques which seek to remove personally identifiable information from datasets. There are three different versions of anonymization: k-anonymity, ℓ-diversity, and t-closeness. The second category of privacy protection techniques is adversarial learning which is considered the state-of-art approach to defend against machine learning models that target and infer individuals' private information. Uncountable works have been using adversarial learning as a defense against private attribute inference attacks and they can be game-theoretic methods or computationally tractable methods. After that, we reviewed the privacy in distributed machine learning from two angles: attacks and defenses. Attacks in distributed machine learn-

ing are classified based on the source of the attacks to: (1) attacks from the server, (2) attacks from a client, and (3) external attacks. Then, we discussed the protections techniques that have been used in federated and split learning to improve individuals privacy.

In Chapter 3, we proposed the privacy-preserving text representation learning framework DB_{BERT} to improve users' privacy in centralized machine learning. The goal of DB_{BERT} is to learn a privacy preserved text representation that is differentially private to protect against identity leakage, does not leak private attributes information (age, gender, location, etc.), and preserves the sentiment meaning of the text. In order to protect from these two kinds of attacks, multiple protection techniques have been applied which are differential privacy and adversarial learning. This framework first extracts the sentence embedding for a given text using SBERT then adds noise using differential privacy mechanism. In order to decide the proper amount of added noise that improves the privacy and retains the utility, we trained a classifier to learn the amount of added noise. After that, adversarial learning idea is followed by defining private attributes discriminators to identify the private information from the text representation. Improving privacy comes at the cost of data utility, which leads to a trade-off between privacy and utility. This trade-off needs to be maintained. Our experiments demonstrated the efficiency of the proposed framework by improving individuals' privacy and preserving the utility of the data for the sentiment meaning.

In chapter 4, we extended our scope to improve the users' privacy in distributed machine learning, specifically in split learning. We proposed privacy-preserving text classification for split learning, $PPSL$, that improves the users' privacy and preserves the utility. Shifting from centralized machine learning to distributed machine learning helped to achieve the scalability that was mentioned earlier as one of the challenges in privacy-preserving models. In our model, $PPSL$, we explore the feasibility of split

learning to deal with text data by building a sentiment classification model. Adversarial learning is used in this model by training discriminators to predict private information using the exchangeable variables between the entities. Our empirical results show the effectiveness of $PPSL$ in both protecting users against private attribute inference attacks and preserving the utility of the data for a downstream task. We investigated the impact of different methods on the privacy which are model complexity, training time, and values of privacy budget. We have seen that increasing the model complexity improves the privacy and at the same time decreases the sentiment classification accuracy. In addition, increasing the model complexity required more computational cost that needs to be determined based on the application resources.

5.2 Conclusion

The availability of massive amounts of user-generated data has enabled machine learning models to provide personalized services and accurate results. On the other hand, this data contains private information about individuals which leads to privacy issues. Individuals face two types of attacks: identity attacks and private information attacks. Thus, individuals' privacy needs to be protected without affecting their data used in different applications to get the desirable results. However, protecting the privacy of the information comes at the cost of the utility, which means the quality of the data for different tasks. Consequently, this leads to a trade off between privacy and utility which needs to be addressed. In addition, no one protection technique is able to defend against all the attacks that affect users' privacy. As a result of this, multiple mechanisms should be combined to achieve better privacy. Furthermore, distributed machine learning has emerged as a new learning paradigm which improves the scalability of the models in different applications. Distributed machine learning improves the privacy by storing the data locally; however, privacy

risks have threatened individuals' information which are caused by a server, a client, or external attacker.

In my dissertation, we investigated how to measure and enhance users' privacy in centralized and distributed machine learning. In centralized machine learning, we proposed a framework to learn a privacy preserved text representation that: (1) is differentially private to protect against identity leakage, (2) protects against leakage of private attributes information, and (3) maintains the high utility for downstream tasks. This work defended against identity attacks and inference attacks while preserving the utility of the data. In distributed machine learning, we proposed a privacy-preserved text classification framework in split learning that protects individuals' private information. By utilizing adversarial learning technique, this framework defended against private attribute inference attacks. The privacy preservation capabilities of our proposed model, PPSL, come from two reasons: it conducts a split learning setting where the training data is kept locally on the client and it utilizes the idea of adversarial learning to minimize the private information leakage.

5.3 Future Work

In this dissertation, we study the research problem of improving users' privacy in centralized and distributed machine learning. Below we present some extensions that are worth investigation:

- **Applying different privacy protection techniques in split learning:** In order to defend against different attacks, multiple protection techniques need to be combined in one framework. Differential privacy, as an example, defends against membership attacks but not against private attribute inference attacks. Multiple mitigations need to be adopted to propose a privacy-preserved split learning model.

- **Exploring the impact of increasing the number of clients in split learning:** In this dissertation, we used the vanilla setting with one client and one server. It would be interesting to explore how increasing the number of clients on split learning will affect the privacy of the trained data. In a previous study, Melis *et al.* (2019) studied the impact of increasing the number of clients in federated learning. They showed how privacy improved when the number of clients increased by making the attacker's task harder. The reason behind this is the amount of aggregated updates that will not directly reveal information about one of the participants.

- **Studying the dependencies between the private information and protecting against that:** Dependencies between the individuals' information is well known. Private attributes such as, age, gender, and location have dependencies between them which affects the attacker inference. Studying the dependencies between the private information and taking them into account when proposing a privacy-preserving machine learning model will be an interesting extended work.

- **Extending the proposed privacy-preserved frameworks on different types of datasets:** Proposing privacy-preserving frameworks that applied on different kind of datasets such as images. As an example, critical problems in the health domain need to utilize a massive amount of patients' data which causes a privacy leakage. Thus, proposing a private way to share and use the data to train a model is crucial. Moreover, we plan to extend our frameworks to be applied on different datasets.

- **Extending privacy-preserved frameworks to consider fairness:** Fairness in machine learning can be divided based on the availability of the sensitive

attributes, as follows:

1. Fairness through unawareness (FTU): when the protected attributes, such as gender and race, are removed and the training is based on the other features. In other words, when the protected attributes do not explicitly contribute to the training process (Kusner *et al.*, 2017).
2. Fairness through awareness: happens when the private attributes is considered in the training process by using the distance metric. The distance metric is used as a similarity metric to treat similar individuals similarly (Dwork *et al.*, 2011).

In our proposed framework, DP_{BERT}, which learned the privacy preserved text representation that is differentially private to protect against identity leakage (if a target instance is available in the data or not), minimizes the private information leakage (age, gender, location, etc.), and preserves the utility of the text for the downstream task. In the future, we will investigate the possibility of extending our proposed framework, DP_{BERT}, to consider fairness by following fairness through unawareness (FTU).

- **Mitigate the privacy risks in large language models:** It has been denoted that recent large language models leaked some private training data. Two different attacks have been reconstructed against large language models *(1) pattern reconstruction attack* and *(2) keyword inference attack* (Pan *et al.*, 2020). *Pattern reconstruction attack* happens when the attacker has prior knowledge of the generating rule of the targeted unknown text. This is usually the case when the text format is well known, such as identity code and date of birth. While *keyword inference attack* happens when the attacker can predict if a certain keyword is contained in the unknown sentence or not. This keyword can be

sensitive information such as location or disease name. One potential research direction is to propose a defense mechanism against these attacks which affect individuals' privacy.

- **Proposing dynamic protection techniques:** Research works in the privacy of machine learning can be divided into two classes: (1) attacks and (2) defenses. In the literature, it has been shown that an uncountable number of proposed defenses have been attacked. As an example of this, the *InstaHide* which protects the privacy of images by encrypting training data and doing the deep learning directly on them (Huang *et al.*, 2020). A reconstruction attack has been proposed against *InstaHide* (Carlini *et al.*, 2020a). Ultimately, this makes it challenging to adopt specific protection techniques to protect and measure users' privacy in real-world applications. In the future, we want to study the ability to propose a defense that can dynamically defend against different attacks.

- **Empirical study on different versions of differential privacy:** Differential privacy is one of the earliest and powerful protection techniques. It has been proposed different versions such as (1) *local differential privacy* and (2) *distributed differential privacy*. Local differential privacy is a special version of the traditional differential privacy where users perturb their data locally before sending the data to an untrusted third party. Compared with traditional differential privacy, where the data is collected first from different users at the trusted party then release the perturbed data publicly. The local differential privacy is when all the users perturb their data locally before sending it to any party (Yang *et al.*, 2020). Distributed differential privacy is called the shuffled model, which takes place between the local differential privacy and center differential

privacy. The basic idea of this model is having a channel, called a shuffler, that receives the data from the users then randomly adds noises and forwards them to the data collector (or server) for the learning process. In the future, we will investigate the impact of these different versions of differential privacy on individuals' privacy and study their impact of the utility.

Abuadbba, S., K. Kim, M. Kim, C. Thapa, S. A. Camtepe, Y. Gao, H. Kim and S. Nepal, "Can we use split learning on 1d cnn models for privacy preserving training?", in "Proceedings of the 15th ACM Asia Conference on Computer and Communications Security", ASIA CCS '20, p. 305–318 (Association for Computing Machinery, New York, NY, USA, 2020), URL https://doi.org/10.1145/3320269.3384740.

Al-Rubaie, M. and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions", IEEE Security & Privacy **17**, 2, 49–58 (2019).

Ali, S., M. H. Shakeel, I. Khan, S. Faizullah and M. A. Khan, "Predicting attributes of nodes using network structure", ACM Trans. Intell. Syst. Technol. **12**, 2, URL https://doi.org/10.1145/3442390 (2021).

Alipour, B., A. Imine and M. Rusinowitch, "Gender Inference for Facebook Picture Owners", in "TrustBus 2019 - 16th International Conference on Trust, Privacy and Security in Digital Business", edited by S. Gritzalis, E. Weippl, S. Katsikas, G. Anderst-Kotsis, A. M. Tjoa and I. Khalil, vol. 11711 of *Lecture Notes in Computer Science*, pp. 145–160 (Springer, Linz, Austria, 2019), URL https://hal.univ-lorraine.fr/hal-02271825.

Alnasser, W., G. Beigi and H. Liu, "An overview on protecting user private-attribute information on social networks", in "Handbook of Research on Cyber Crime and Information Privacy", edited by M. M. Cruz-Cunha and Mateus-Coelho, chap. 6 (2020a).

Alnasser, W., G. Beigi and H. Liu, "An overview on protecting user private-attribute information on social networks", in "Handbook of Research on Cyber Crime and Information Privacy", edited by M. M. Cruz-Cunha and Mateus-Coelho, chap. 6 (2020b).

Alnasser, W., G. Beigi and H. Liu, "Privacy preserving text representation learning using bert", in "Social, Cultural, and Behavioral Modeling", edited by R. Thomson, M. N. Hussain, C. Dancy and A. Pyke, pp. 91–100 (Springer International Publishing, Cham, 2021).

Aslett, L. J., P. M. Esperança and C. C. Holmes, "A review of homomorphic encryption and software tools for encrypted statistical machine learning", arXiv preprint arXiv:1508.06574 (2015).

Beigi, G. and H. Liu, "A survey on privacy in social media: Identification, mitigation, and applications", ACM/IMS Trans. Data Sci. **1**, 1, URL https://doi.org/10.1145/3343038 (2020).

Beigi, G., A. Mosallanezhad, R. Guo, H. Alvari, A. Nou and H. Liu, "Privacy-aware recommendation with private-attribute protection using adversarial learning", in "Proceedings of the 13th International Conference on Web Search and Data Mining", pp. 34–42 (2020).

Beigi, G., K. Shu, R. Guo, S. Wang and H. Liu, "Privacy preserving text representation learning", in "Proceedings of the 30th ACM Conference on Hypertext and Social Media", HT '19, p. 275–276 (Association for Computing Machinery, New York, NY, USA, 2019), URL https://doi.org/10.1145/3342220.3344925.

Bonawitz, K., V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal and K. Seth, "Practical secure aggregation for privacy-preserving machine learning", in "Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security", CCS '17, p. 1175–1191 (Association for Computing Machinery, New York, NY, USA, 2017), URL https://doi.org/10.1145/3133956.3133982.

Cai, Z., Z. He, X. Guan and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks", IEEE Transactions on Dependable and Secure Computing **15**, 4, 577–590 (2018).

Carlini, N., S. Deng, S. Garg, S. Jha, S. Mahloujifar, M. Mahmoody, S. Song, A. Thakurta and F. Tramer, "Is private learning possible with instance encoding?", URL https://arxiv.org/abs/2011.05315 (2020a).

Carlini, N., F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea and C. Raffel, "Extracting training data from large language models", CoRR **abs/2012.07805**, URL https://arxiv.org/abs/2012.07805 (2020b).

Chaudhuri, K., C. Monteleoni and A. D. Sarwate, "Differentially private empirical risk minimization", Journal of Machine Learning Research **12**, 29, 1069–1109, URL http://jmlr.org/papers/v12/chaudhuri11a.html (2011).

Chen, J., J. He, L. Cai and J. Pan, "Profiling online social network users via relationships and network characteristics", in "2016 IEEE Global Communications Conference (GLOBECOM)", pp. 1–6 (2016).

Cheng, K., T. Fan, Y. Jin, Y. Liu, T. Chen, D. Papadopoulos and Q. Yang, "Secureboost: A lossless federated learning framework", URL https://arxiv.org/abs/1901.08755 (2019).

Choudhury, O., A. Gkoulalas-Divanis, T. Salonidis, I. Sylla, Y. Park, G. Hsu and A. Das, "Differential privacy-enabled federated learning for sensitive health data", CoRR **abs/1910.02578**, URL http://arxiv.org/abs/1910.02578 (2019).

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", (2019).

dos Santos, C. and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts", in "Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers", pp. 69–78 (Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014), URL https://www.aclweb.org/anthology/C14-1008.

Dwork, C., "Differential privacy", in "Automata, Languages and Programming", edited by M. Bugliesi, B. Preneel, V. Sassone and I. Wegener, pp. 1–12 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006).

Dwork, C., "Differential privacy: A survey of results", in "Theory and Applications of Models of Computation", edited by M. Agrawal, D. Du, Z. Duan and A. Li, pp. 1–19 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).

Dwork, C., M. Hardt, T. Pitassi, O. Reingold and R. Zemel, "Fairness through awareness", URL https://arxiv.org/abs/1104.3913 (2011).

Ethayarajh, K., "Unsupervised random walk sentence embeddings: A strong but simple baseline", in "Proceedings of The Third Workshop on Representation Learning for NLP", pp. 91–100 (Association for Computational Linguistics, Melbourne, Australia, 2018), URL https://www.aclweb.org/anthology/W18-3012.

Feldman, V. and C. Zhang, "What neural networks memorize and why: Discovering the long tail via influence estimation", **33**, 2881–2891 (2020).

Filho, J. A. B. L., R. Pasti and L. N. de Castro, "Gender classification of twitter data based on textual meta-attributes extraction", in "New Advances in Information Systems and Technologies", edited by Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis and M. Mendonça Teixeira, pp. 1025–1034 (Springer International Publishing, Cham, 2016).

Fung, B. C. M., K. Wang, R. Chen and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments", ACM Comput. Surv. **42**, 4, URL https://doi.org/10.1145/1749603.1749605 (2010).

Gao, Y., M. Kim, S. Abuadbba, Y. Kim, C. Thapa, K. Kim, S. A. Camtep, H. Kim and S. Nepal, "End-to-end evaluation of federated learning and split learning for internet of things", in "2020 International Symposium on Reliable Distributed Systems (SRDS)", pp. 91–100 (2020).

Geyer, R. C., T. Klein and M. Nabi, "Differentially private federated learning: A client level perspective", URL https://arxiv.org/abs/1712.07557 (2017).

Gong, N. Z. and B. Liu, "Attribute inference attacks in online social networks", ACM Trans. Priv. Secur. **21**, 1, URL https://doi.org/10.1145/3154793 (2018).

Gong, N. Z., A. Talwalkar, L. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. R. Shi and D. Song, "Joint link prediction and attribute inference using a social-attribute network", ACM Trans. Intell. Syst. Technol. **5**, 2, URL https://doi.org/10.1145/2594455 (2014).

Gupta, O. and R. Raskar, "Distributed learning of deep neural network over multiple agents", Journal of Network and Computer Applications **116**, 1–8, URL https://www.sciencedirect.com/science/article/pii/S1084804518301590 (2018).

Han, X., H. Huang and L. Wang, "F-pad: Private attribute disclosure risk estimation in online social networks", IEEE Transactions on Dependable and Secure Computing **16**, 6, 1054–1069 (2019).

He, Z. and Y. Huang, "Tracking histogram of attributes over private social data in data markets", in "Combinatorial Optimization and Applications", edited by Y. Li, M. Cardei and Y. Huang, pp. 277–288 (Springer International Publishing, Cham, 2019).

Hovy, D., A. Johannsen and A. Søgaard, "User review sites as a resource for large-scale sociolinguistic studies", in "Proceedings of the 24th International Conference on World Wide Web", WWW '15, p. 452–461 (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2015), URL https://doi.org/10.1145/2736277.2741141.

Hovy, D. and A. Søgaard, "Tagging performance correlates with author age", in "Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)", pp. 483–488 (Association for Computational Linguistics, Beijing, China, 2015), URL https://www.aclweb.org/anthology/P15-2079.

Huang, Y., Z. Song, K. Li and S. Arora, "InstaHide: Instance-hiding schemes for private distributed learning", in "Proceedings of the 37th International Conference on Machine Learning", edited by H. D. III and A. Singh, vol. 119 of *Proceedings of Machine Learning Research*, pp. 4507–4518 (PMLR, 2020), URL https://proceedings.mlr.press/v119/huang20i.html.

Idan, L. and J. Feigenbaum, "Show me your friends, and i will tell you whom you vote for: Predicting voting behavior in social networks", in "2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)", pp. 816–824 (2019).

Jia, J. and N. Z. Gong, "Attriguard: A practical defense against attribute inference attacks via adversarial machine learning", in "Proceedings of the 27th USENIX Conference on Security Symposium", SEC'18, p. 513–529 (USENIX Association, USA, 2018).

Jia, J., B. Wang, L. Zhang and N. Z. Gong, "Attriinfer: Inferring user attributes in online social networks using markov random fields", in "Proceedings of the 26th International Conference on World Wide Web", WWW '17, p. 1561–1569 (International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017), URL https://doi.org/10.1145/3038912.3052695.

Kingma, D. P. and M. Welling, "Auto-encoding variational bayes", (2014).

Kusner, M. J., J. Loftus, C. Russell and R. Silva, "Counterfactual fairness", Advances in neural information processing systems **30** (2017).

Li, N., T. Li and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity", in "2007 IEEE 23rd International Conference on Data Engineering", pp. 106–115 (2007).

Li, O., J. Sun, X. Yang, W. Gao, H. Zhang, J. Xie, V. Smith and C. Wang, "Label leakage and protection in two-party split learning", (2021).

Li, T., A. K. Sahu, A. Talwalkar and V. Smith, "Federated learning: Challenges, methods, and future directions", IEEE Signal Process. Mag. **37**, 3, 50–60, URL https://doi.org/10.1109/MSP.2020.2975749 (2020).

Li, Y., T. Baldwin and T. Cohn, "Towards robust and privacy-preserving text representations", in "Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)", pp. 25–30 (Association for Computational Linguistics, Melbourne, Australia, 2018), URL https://aclanthology.org/P18-2005.

Liu, P., Y. Xu, Q. Jiang, Y. Tang, Y. Guo, L. e Wang and X. Li, "Local differential privacy for social network publishing", Neurocomputing **391**, 273–279, URL https://www.sciencedirect.com/science/article/pii/S0925231219304229 (2020).

Machanavajjhala, A., J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity", in "22nd International Conference on Data Engineering (ICDE'06)", pp. 24–24 (2006).

Madaan, H., M. Gawali, V. Kulkarni and A. Pant, "Vulnerability due to training order in split learning", arXiv preprint arXiv:2103.14291 (2021).

Mao, J., W. Tian, Y. Yang and J. Liu, "An efficient social attribute inference scheme based on social links and attribute relevance", IEEE Access **7**, 153074–153085 (2019).

McMahan, B., E. Moore, D. Ramage, S. Hampson and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data", in "Proceedings of the 20th International Conference on Artificial Intelligence and Statistics", edited by A. Singh and J. Zhu, vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282 (PMLR, 2017), URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Melis, L., C. Song, E. D. Cristofaro and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning", 2019 IEEE Symposium on Security and Privacy (SP) pp. 691–706 (2019).

Mosallanezhad, A., G. Beigi and H. Liu, "Deep reinforcement learning-based text anonymization against private-attribute inference", in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", pp. 2360–2369 (Association for Computational Linguistics, Hong Kong, China, 2019), URL https://aclanthology.org/D19-1240.

Narayanan, A. and V. Shmatikov, "De-anonymizing social networks", in "2009 30th IEEE Symposium on Security and Privacy", pp. 173–187 (2009).

Nie, L., L. Zhang, M. Wang, R. Hong, A. Farseev and T.-S. Chua, "Learning user attributes via mobile social multimedia analytics", ACM Trans. Intell. Syst. Technol. **8**, 3, URL https://doi.org/10.1145/2963105 (2017).

Pan, X., M. Zhang, S. Ji and M. Yang, "Privacy risks of general-purpose language models", in "2020 IEEE Symposium on Security and Privacy (SP)", pp. 1314–1331 (2020).

Pasquini, D., G. Ateniese and M. Bernaschi, "Unleashing the tiger: Inference attacks on split learning", in "Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security", CCS '21, p. 2113–2129 (Association for Computing Machinery, New York, NY, USA, 2021), URL https://doi.org/10.1145/3460120.3485259.

Pereteanu, G.-L., A. Alansary and J. Passerat-Palmbach, "Split he: Fast secure inference combining split learning and homomorphic encryption", URL https://arxiv.org/abs/2202.13351 (2022).

Poirot, M. G., P. Vepakomma, K. Chang, J. Kalpathy-Cramer, R. Gupta and R. Raskar, "Split learning for collaborative deep learning in healthcare", (2019).

Rabbany, R., D. Eswaran, A. W. Dubrawski and C. Faloutsos, "Beyond assortativity: Proclivity index for attributed networks (prone)", in "PAKDD", (2017).

Raval, N., A. Machanavajjhala and J. Pan, "Olympus: Sensor privacy through utility aware obfuscation.", Proc. Priv. Enhancing Technol. **2019**, 1, 5–25 (2019).

Reimers, N. and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks", in "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", pp. 3982–3992 (Association for Computational Linguistics, Hong Kong, China, 2019), URL https://www.aclweb.org/anthology/D19-1410.

Salamatian, S., A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira and N. Taft, "Managing your private and public data: Bringing down inference attacks against your privacy", IEEE Journal of Selected Topics in Signal Processing **9**, 7, 1240–1255, URL http://dx.doi.org/10.1109/JSTSP.2015.2442227 (2015).

Shah, H., V. Kakkad, R. Patel and N. Doshi, "A survey on game theoretic approaches for privacy preservation in data mining and network security", Procedia Computer Science **155**, 686–691, URL https://www.sciencedirect.com/science/article/pii/S1877050919310142, the 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology (2019).

Shokri, R., G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux and J.-Y. Le Boudec, "Protecting location privacy: Optimal strategy against localization attacks", in "Proceedings of the 2012 ACM Conference on Computer and Communications Security", CCS '12, p. 617–627 (Association for Computing Machinery, New York, NY, USA, 2012), URL https://doi.org/10.1145/2382196.2382261.

Singh, A., P. Vepakomma, O. Gupta and R. Raskar, "Detailed comparison of communication efficiency of split learning and federated learning", (2019).

Soria-Comas, J., J. Domingo-Ferrer, D. Sánchez and S. Martínez, "t-closeness through microaggregation: Strict privacy with enhanced utility preservation", IEEE Transactions on Knowledge and Data Engineering **27**, 11, 3098–3110 (2015).

Sun, C., X. Qiu, Y. Xu and X. Huang, "How to fine-tune bert for text classification?", (2020).

Sweeney, L., "¡i¿k¡/i¿-anonymity: A model for protecting privacy", Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**, 5, 557–570, URL https://doi.org/10.1142/S0218488502001648 (2002).

Tanuwidjaja, H. C., R. Choi, S. Baek and K. Kim, "Privacy-preserving deep learning on machine learning as a service—a comprehensive survey", IEEE Access **8**, 167425–167447 (2020).

Thapa, C., M. A. P. Chamikara, S. Camtepe and L. Sun, "Splitfed: When federated learning meets split learning", arXiv preprint arXiv:2004.12088 (2020).

Vepakomma, P., O. Gupta, T. Swedish and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data", (2018).

Vepakomma, P., A. Singh, O. Gupta and R. Raskar, "Nopeek: Information leakage reduction to share activations in distributed deep learning", in "20th International Conference on Data Mining Workshops, ICDM Workshops 2020, Sorrento, Italy, November 17-20, 2020", edited by G. D. Fatta, V. S. Sheng, A. Cuzzocrea, C. Zaniolo and X. Wu, pp. 933–942 (IEEE, 2020), URL https://doi.org/10.1109/ICDMW51313.2020.00134.

Wang, B. and C. C. J. Kuo, "Sbert-wk: A sentence embedding method by dissecting bert-based word models", (2020).

Wang, Y., Y. Wang, J. Liu, Z. Huang and P. Xie, "A survey of game theoretic methods for cyber security", in "2016 IEEE First International Conference on Data Science in Cyberspace (DSC)", pp. 631–636 (2016).

Xiao, Y. and L. Xiong, "Protecting locations with differential privacy under temporal correlations", in "Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security", CCS '15, p. 1298–1309 (Association for Computing Machinery, New York, NY, USA, 2015), URL https://doi.org/10.1145/2810103.2813640.

Xu, L., C. Jiang, J. Wang, Y. Ren, J. Yuan and M. Guizani, "Game theoretic data privacy preservation: Equilibrium and pricing", in "2015 IEEE International Conference on Communications (ICC)", pp. 7071–7076 (2015).

Xu, Q., L. Qu, C. Xu and R. Cui, "Privacy-aware text rewriting", in "Proceedings of the 12th International Conference on Natural Language Generation", pp. 247–257 (Association for Computational Linguistics, Tokyo, Japan, 2019), URL https://aclanthology.org/W19-8633.

Yang, M., L. Lyu, J. Zhao, T. Zhu and K.-Y. Lam, "Local differential privacy and its applications: A comprehensive survey", arXiv preprint arXiv:2008.03686 (2020).

Zhong, Y., N. J. Yuan, W. Zhong, F. Zhang and X. Xie, "You are where you go: Inferring demographic attributes from location check-ins", in "Proceedings of the Eighth ACM International Conference on Web Search and Data Mining", WSDM '15, p. 295–304 (Association for Computing Machinery, New York, NY, USA, 2015), URL https://doi.org/10.1145/2684822.2685287.