IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 6, 2021, accepted October 24, 2021, date of publication October 28, 2021, date of current version November 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3124163

Review: Privacy-Preservation in the Context of Natural Language Processing

DARSHINI MAHENDRAN, CHANGQING LUO, (Member, IEEE), AND BRIDGET T. MCINNES
Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

Corresponding author: Darshini Mahendran (mahendrand@vcu.edu)

ABSTRACT Data privacy is one of the highly discussed issues in recent years as we encounter data breaches and privacy scandals often. This raises a lot of concerns about the ways the data is acquired and the potential information leaks. Especially in the field of Artificial Intelligence (AI), the widely using of AI models aggravates the vulnerability of user privacy because a considerable portion of user data that AI models used is represented in natural language. In the past few years, many researchers have proposed NLP-based methods to address these data privacy challenges. To the best of our knowledge, this is the first interdisciplinary review discussing privacy preservation in the context of NLP. In this paper, we present a comprehensive review of previous research conducted to gather techniques and challenges of building and testing privacy-preserving systems in the context of Natural Language Processing (NLP). We group the different works under four categories: 1) Data privacy in the medical domain, 2) Privacy preservation in the technology domain, 3) Analysis of privacy policies, and 4) Privacy leaks detection in the text representation. This review compares the contributions and pitfalls of the various privacy violation detection and prevention works done using NLP techniques to help guide a path ahead.

INDEX TERMS Data privacy, natural language processing, privacy preservation, privacy policy.

I. INTRODUCTION

Data privacy is a highly discussed issue, and we encounter data breaches and privacy scandals in our day-to-day life. This is mainly due to the collection of exponentially increasing data and the use of the data on various applications and research. This raises many concerns about the ways data is acquired and potential information leaks. We find potential risks of private/sensitive information leaks in different instances. The introduction of Machine Learning (ML) models has spiked the use of vast amounts of data for the training of Artificial Intelligence (AI) models [1]. There are many opportunities where privacy of the data could be violated when used in AI models, for example, an adversary could listen to the latent representation of the input in the ML models and obtain sensitive information. Therefore, there is an increased interest in privacy-preserving data mining techniques and privacy-preserving data analysis in recent years, protecting individual information. Preserving the privacy of training data for ML models is essential to guarantee data security and maintain user trust for continuous access to unlimited data that improve the performance of the models [1].

The sensitivity of the data can be categorized as 1) implicit information and 2) explicit information. When the information is directly derived from a user's query (e.g., web search), it is called implicit information (e.g., age, gender). In contrast, when the information is derived using pattern matching, it is called explicit information (e.g., Personal Identification Number (PIN), Social Security Number (SSN)) [1]. The traditional privacy protection methods are unable handle this growing need to protect data. They are very time and resource consuming unlike the AI models. Therefore, it is necessary to build systems that can not only provide such privacy assurances but also with increased automation and reliability [2]. The medical field has a high risk of exposing privacy details, where the records hold each patient's entire history and details. There is a potential risk of exposure to medical records while stored in the databases online or shared between institutions. Another field that is highly susceptible to privacy leakage is social media networks, applications, and software. In the past decade, we have seen enormous growth in people's interest in using social media networks, and often they do not realize the threat social media pose. Mostly the privacy policies used by software and apps are

The associate editor coordinating the review of this manuscript and approving it for publication was Gautam Srivastava.

long, verbose and some exploit this situation to collect and misuse the personal information of the users [3].

Natural Language Processing (NLP) is a field that combines linguistics and computer science to analyze and understand meaning from human language. NLP is used in many applications we see in our day-to-day life, such as chatbots, voice assistants, and search engines. A considerable portion of user-contributed data comes from natural language (e.g., text and voice recordings), including user-privacy data. In the past few years, researchers proposed many techniques for solving privacy-related issues and preserving privacy, including quantum cryptography, adversarial ML, and access control techniques, and recently they started to apply NLP-based methods to address the data privacy challenges, which results in an intersection of NLP and Privacy [1]. This makes privacy a well-motivated application domain for NLP researchers. However, to the best of our knowledge, there is currently no interdisciplinary review discussing the intersection of privacy preservation and NLP.

This paper provides an overview of past works where NLP was used to identify privacy leaks, help build a system for privacy preservation, and identify techniques and challenges of building and testing privacy-preserving systems. The motivation for our review is to gain an understanding of the utilization of NLP in the privacy field. We divide the different applications into four categories: 1) Data privacy in the medical domain, 2) Privacy preservation in the technology domain, 3) Analysis of privacy policies, and 4) Privacy leak detection in the text representation. The remainder of this review is structured as follows. First, we discuss the different approaches under the four categories mentioned above. Then we present a table summarizing all the works related to privacy in NLP and the future directions we propose. Finally, we conclude this review with a conclusion that summarizes the review.

II. DATA PRIVACY IN MEDICAL DOMAIN

Protected Health Information (PHI) is the information in medical records or information systems that can be used to identify patients. Some examples of PHI are patient name, phone number, physician name, and medication history. Due to the medical field's advancement, there is a growing need to share medical records between institutions. Sharing data can improve clinical decision support systems, big data medical research, and treatment quality assurance [4]. However, one of the biggest challenges is the sharing and dissemination of medical records while maintaining a commitment to patient confidentiality [5]. There is an ethical and legal responsibility towards respecting the individuals' privacy which led to the introduction of specific laws that address this issue, such as the European Union's General Data Protection Regulation (GDPR) directive or the United States' Health Insurance Portability and Accountability Act (HIPAA) [6].

To secure patients' privacy, the PHI is required to be anonymized prior to sending it to another institute. Many efforts have been devoted to this endeavor, including manual and the automatic approaches [7]. Due to the recent exponential growth in the literature, the cost of manually anonymizing large data is exceptionally high. Therefore there is an increased interest in automating the anonymization procedure through the use of NLP techniques. Anonymization is considered one of the complex tasks due to the unstructured nature of clinical notes.

Here we divide the proposed systems into three categories: Rule-based, ML-based, and Deep Learning (DL)-based systems. Each has both advantages and disadvantages. Rule-based systems utilize rules and patterns to represent knowledge. They include regular expressions and pattern matching and are easy to build, maintain. However, these technologies require tedious manual labor to generate and update the rules [8] by domain-specific experts. ML-based systems use machine learning algorithms and statistical analysis for knowledge representation. Machine learning approaches including Hidden Markov Models (HMM), Conditional Random Field (CRF), Maximum Entropy Models (MaxEnt), Support Vector Machines (SVMs), Naïve Bayes (NB), and Random Forests (RFs) [9]. These have an advantage over rule-based systems as they do not require manual rule or expert knowledge, but they require labeled data for training and typically require manual feature engineering. Recently DL-based systems have obtained very high performance across many NLP tasks and do not require manual feature engineering. Two common techniques used are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs can capture continuous local features of sequences through the convolution operation, whereas RNNs obtain long-term dependencies through the recursive process. Long short-term memory (LSTM) is an RNN that has brought more flexibility in controlling the outputs. Bi-directional Long Short-Term Memory (Bi-LSTM) is an extension of LSTMs and consists of two LSTMs and controls the flow from both directions. In this section, we describe previous works within each of these categories.

A. RULE-BASED SYSTEMS

Earlier systems used rule-based or template-based approaches to match patterns and detect PHI from clinical notes. For example, Sweeney, *et al.* [5], Berman, *et al.* [10], and Beckwith, *et al.* [11] proposed the concept of scrub system or tool for anonymization. Sweeney *et al.* [5] proposed a Scrub system for anonymization which uses two approaches to identify a PHI: a computer-based approach, which used detection algorithms competing in parallel to label the identifiers, and a human-based approach where five individuals with no medical experience or experience with the information contained in the database used a template and a set of rules to identify a PHI. Berman, *et al.* [10] used a concept based scrubs algorithm for a similar problem, and the algorithm works as follows: when the algorithm encounters a nomenclature term, it replaces the term by the nomenclature code and a synonym of the original term, but when it encounters another type of words it replaces them with asterisks. This method was considered safe as the output of this method contains only medical terms. Beckwith, *et al.* [11] designed an open-source software tool to de-identify patient information from electronic medical records, including pathology reports using a three-step process: look for identifiers associated with the patient, predict patterns likely to represent identifying data, and compare with a database of proper names and geographic locations. Recently, Iwendi *et al.* [12] proposed a semantic privacy framework named *N-Sanitization* that effectively sanitizes the sensitive and semantically related terms in healthcare documents. First, they used dictionaries, regular expressions, and Stanford NER Tagger to detect maximum PHIs and sensitive terms. Then they used a medical ontology (knowledgebase) named SNOMED-CT to sanitize the previously detected sensitive terms by substituting them with their generalized terms. They removed the negative sentences (assertions) from documents before the sanitization process.

B. MACHINE LEARNING-BASED SYSTEMS

Named Entity Recognition (NER), also known as entity extraction, automatically identifies and classifies terms from unstructured text into pre-defined categories or classes. For example, categories in the privacy domain include names, addresses, gender, age, country, profession, or any other personal details [6]. Many past works mapped the text de-identification problem to a Named Entity Recognition (NER) problem. The entities in the text that contain the patients' personal information (entities to be de-identified) are treated as the entities that need to be extracted. The anonymization task is similar to the NER, but it is more complex as it deletes personal information and attempts to classify the personal information in the text to one of the HIPAA-defined categories [13].

Over the years, many researchers proposed ML-based approaches to achieve anonymization such as Medlock *et al.* [4], Szarvas *et al.* [14], Lopez *et al.* [6]. Medlock *et al.* [4] proposed an NLP-based text anonymization technique to preserve patients' privacy. They utilized three different strategies to achieve anonymization: a) removing the sensitive reference with a blank placeholder, b) replacing the reference with the name of its category, and c) replacing the reference with the same category pseudo reference. Following features were used to train an ML model and classify whether the cluster contains sensitive information: Part-of-Speech (POS), inner left constituent label, 2*nd* inner left constituent label, outer left constituent label, outer left constituent token, and orthography. Szarvas *et al.* [14] used a decision tree ML-based, iterative NER approach to deanonymize semi-structured documents such as discharge summary records. Here, the iterative learning method utilizes the information given in the structured parts of the texts to improve PHI recognition accuracy in flow text. Recently, Lopez *et al.* [6] proposed HITZALMED,[1] a web-framed tool that assists with the anonymization of clinical free text in Spanish. Similar to Medlock *et al.* [4], this supports identification, classification, masking, and replacement of sensitive information. Also, once sensitive information is detected, different anonymization techniques are implemented, configurable by the user. They utilized a hybrid approach that combines ML techniques to detect PHI and a rule-based system for anonymization.

In 2014, i2b2/UTHealth NLP shared task featured a de-identification track that focused on identifying PHIs in clinical narratives [15]. They introduced a newly de-identified corpus of longitudinal medical records drawn from the Research Patient Data Repository of Partners Healthcare. Popular submissions of the shared task included CRF-based systems. He *et al.* [16] trained a CRF system with the following features: lexical, orthographic, and syntactic. They pre-processed their data with OpenNLP's tokenizer. Grouin *et al.* [17], Liu *et al.* [18], and Yang *et al.* [19] utilized both CRF and rule-based approaches in their systems. The CRF-based approach of Grouin *et al.* [17] included linguistic features such as surface features such as token itself, token length, typographic case, presence of punctuation or digits, and morpho-syntactic features such as POS, distributional analysis features, such as the frequency in the corpus, document section, and cluster ID based on context. They also utilized regular expressions in their rule-based approach to correct CRF outputs. The CRF-based approach of Yang *et al.* [19] utilized word-token (lemma, POS, chunk), context (lemma, POS, chunk of nearby tokens), orthographic (capitalization, punctuation, regex patterns for dates, usernames), sentence-level features (position of the token in a sentence, section headers). They used dictionaries and regular expressions to identify PHI with few sample instances. The CRF-based approach of Liu *et al.* [18] included bag-of-words, POS, orthography features, section information, and word representation features, and the rule-based approach used regular expressions to identify standardized PHI.

C. DEEP LEARNING-BASED SYSTEMS

DL-based NLP approaches have improved data extraction performance and require no handcrafted features or rules. Recent works have utilized DL techniques for detecting PHIs. Dernoncourt *et al.* [20], Jiang *et al.* [21], and Catelli *et al.* [22] developed two systems based on CRFs and Bi-LSTMs for patient de-identification. Jiang *et al.* [21] developed a CRF and a Bi-LSTM network-based system that focus on de-identifying psychiatric evaluation records. They manually extracted rich features to train the model for CRFs, and applied a character-level Bi-LSTM network to represent tokens and classify tags. Dernoncourt, *et al.* [20] used a combination of n-gram, morphological, orthographic, and gazetteer features for the CRF model. They also map each token using a character-enhanced embedding into a vector representation for the Bi-LSTM model. Dernoncourt, *et al.* [23] presented *NeuroNER*[2],

[1] https://snlt.vicomtech.org/hitzalmed

[2] http://neuroner.com

an easy-to-use NER tool based on Artificial neural networks (ANNs). They utilize the NER tool for patient de-identification entities and utilize LSTM-based RNN for non-overlapping label prediction. Furthermore, Dobbins *et al.* [24] utilized the same tool used by Dernoncourt *et al.* [23] to compare the performance differences across two datasets for patient de-identification. They also created a dataset specifically for this study SIRM[3] COVID-19 de-identification corpus from medical records provided by *NeuroNER [23]*

Recently, Catelli *et al.* [22], [25] focused on how different word embeddings affect the input representation. Catelli *et al.* [22] built a network combining Bi-LSTM and CRF network to predict the target PHI entities. Here, they utilized the Flair contextualized and character-level language model [26], a contextualized language model, working at the character level, to capture the polysemy of words and manage the morpho-syntactic variations typical of handwritten notes. They argued that the stacked word representations capture latent syntactic and semantic similarities better. Catelli *et al.* [25] further investigated the effectiveness of cross-lingual transfer learning to de-identify medical records written in a low resource language such as Italian, using one with high resources such as English while maintaining the necessary features to perform the NER task for de-identification correctly. Here, they utilized with stacked embedding consisting of MultiBPEmb [27] and Flair embeddings [26] and Multilingual Bidirectional Encoder Representations from Transformers (mBERT)-cased[4] model. The mBERT provides sentence representations for 104 languages, which are useful for many multi-lingual tasks.

Most of the proposed Bi-LSTM based models utilized only the global context to detect clinical entities and PHIs, not the local context. Therefore, Moqurrab *et al.* [28] proposed a combination of CNN, Bi-LSTM, and CRF with non-complex embeddings to utilize both local and global context. Here, CNN was used to capture local context, while Bi-LSTM was used to capture global context. First, six independent CNN models are applied to extract the local context with various window sizes, then the combined local context is concatenated with the input representation and passed to the three-layered sequential Bi-LSTM architecture. Finally, the combined local and global context is passed to the CRF layer.

Li *et al.* [29], Sadat *et al.* [7] tried an alternative approach named frequency-filtering, to remove text that might contain sensitive terms related to personal information. Li *et al.* [29] investigated the use of a frequency-filtering approach where they filter out rare sentences (frequency < 3) and sentences containing bigrams under a certain frequency threshold (frequency < 256). Their approach is based on the assumption that sentences that appear frequently tend to contain no PHI, which originates from the observation collected over many records. This approach is applicable for data anonymization from a single source. Improving the work of Li *et al.* [29], Sadat *et al.* [7] extended the model to be applicable for distributed sources. Sadat *et al.* [7] used frequency-based filtering to improve privacy protection on distributed sources of medical data. This framework first identified uncommon and low-frequency bigrams used to remove sentences from clinical notes containing PHI. This work also demonstrated the usefulness of homomorphic encryption for secure multi-party data analysis on medical records.

Table 1 shows an overview of the works done related to data privacy in the medical domain. For each work, it shows the year the work is published, the dataset used, and the type of approach used.

III. PRIVACY PRESERVATION IN TECHNOLOGY DOMAIN

We have seen enormous growth in people's interest in using social media networks, apps, and software in the past decade. Although these social media platforms allow people to freely interact and simplify their day-to-day activities, we often do not realize how much private and sensitive information is leaked [40]. This is primarily due to the user's lack of knowledge about the risks of privacy. Previous studies demonstrated that privacy preservation is conditioned by the following reasons [41]:

1) Individuals believe that they are less exposed to risks than others.
2) Individuals consider themselves with higher skills than those they exhibit.
3) Individuals cannot evaluate the relevant risk factors as they are unaware of the most privacy risks.

Due to the above reasons educating individuals about potential privacy risks and building privacy preservation systems is essential. Many works such as Cappellari *et al.* [42], Canfora *et al.* [41] utilized NLP-based solutions along with ML models to detect and prevent privacy violations. Cappellari *et al.* [42] proposed a method to detect messages that carry sensitive information, and they built a privacy protection framework where a client-side privacy awareness mechanism can alert users of the potential private information leakages in their communications. They employ ML methods to build a privacy decision-making tool. They utilized NLP techniques during pre-processing, such as remove stop words, replace each word with a common synonym via the *WordNet* lexical database [43]; and each word is stemmed to reduce the dictionary of terms to words in their root form. Canfora *et al.* [41] proposed a method, and an accompanying tool, to automatically intercept the sensitive information delivered in a social network post. They recognized specific recurrent patterns used in natural language by the user to express specific privacy leakage classes using the syntactic structures and classified the classes automatically. Following are the features they used: tokenization, lowercase conversion, stop-word removal, and stemming. They ensure sentence classification performance does not change with the features' selection or training set and outperforms the state-of-the-art ML techniques. They also developed a browser

[3] https://sirm.org/category/senza-categoria/covid-19/

[4] https://github.com/google-research/bert/blob/master/multilingual.md

TABLE 1. Overview of the works done related data privacy in medical domain.

Year	Paper	Dataset	Model
1996	Sweeney, et al. [5]	Scrubbed subset of a pediatric medical record system	Detection algorithms using templates and knowledge base
2003	Berman, et al. [10]	Pathology free text	Pattern matching
2006	Beckwith, et al. [11]	Pathology reports [30]	Pattern matching
2006	Medlock, et al. [4]	Informal Text Anonymization Corpus (ITAC)[5]	HMM
2007	Szarvas et al. [14]	i2b2-NLP shared task dataset [31]	Boosting, C4.5 for pattern matching
2014	He, et al. [16]	i2b2-2014 dataset [32]	CRF
2014	Liu, et al. [18]	i2b2-2014 dataset [32]	CRF, Pattern matching
2014	Yang, et al. [19]	i2b2-2014 dataset [32]	CRF, Dictionaries, Pattern matching
2014	Grouin, et al. [17]	i2b2-2014 dataset [32]	CRF, Pattern matching
2015	Li, et al. [29]	Enterprise Data Trust (EDT) [33] notes	Frequency-filtering
2017	Jiang, et al. [21]	Psychiatric evaluation records [34]	CRF, Bi-LSTM
2017	Dernoncourt, et al. [20]	i2b2-2014 dataset [32], MIMIC de-identification dataset [20]	CRF, Bi-LSTM
2017	Dernoncourt, et al. [23]	i2b2-2014 dataset [32], CoNLL-2003 [35]	CRF, Bi-LSTM
2017	Dobbins, et al. [24]	UW-Dataset [24], i2b2-2014 dataset [32]	CRF, Bi-LSTM
2019	Sadat, et al. [7]	Medical Information Mart for Intensive Care (MIMIC) [36] notes	Frequency-based filtering
2020	Lopez, et al. [6]	MEDDOCAN[6]	Pattern matching, Dictionaries, and ML algorithms
2020	Iwendi, et al [12]	i2b2-2010 NLP dataset[7]	dictionaries, regular expressions and Stanford NER Tagger
2021	Catelli, et al. [22]	i2b2-2014 dataset [32]	CRF, Bi-LSTM
2021	Catelli, et al. [25]	i2b2-2014 dataset [32], Italian SIRM COVID-19 [37]	CRF, Bi-LSTM, mBERT-cased
2021	Moqurrab, et al. [28]	i2b2-2010 [38], i2b2-2012 [39]	CNN, Bi-LSTM, CRF

extension for privacy-preserving when users write posts on Twitter.

In Europe, organizations are legally bound to release contractual information containing specific personal information of individuals. Therefore, for privacy assurance, several systems are built to auto-monitor *Personally Identifiable Information (PII)*. PII indicates any representation of information that can expose the identity of an individual same as PHI. Therefore, from here on, we use both terms interchangeably. Silva *et al.* [2] proposed a system where they used NER to identify, monitor, and validate the PII. The experiments used three of the most well-known NLP tools to analyze their characteristics and capabilities: Natural Language Toolkit (NLTK[8]), Stanford CoreNLP [44], and spaCy.[9] NLTK is an open-source Python software that allows manipulating different corpora, analyzes the linguistic structure, and categorizes text. Stanford CoreNLP is an open-source Java software containing higher-level NLP components, including sentiment analysis, dependency parsing, or NER. Finally, spaCy is an open-source software library for NLP written in Python and Cython and is considered one of the fastest NLP libraries. First, they assessed the tools' effectiveness with a generic dataset, then applied to datasets that contained any publicly available PII like names, addresses, contact numbers, or other related types. Further, they established that their method could act as a Privacy Enhancing Technology (PET) and the potential risks and associated impacts.

Nan *et al.* [45] addressed the challenge in analyzing information leaks within mobile apps for automatically detecting code operating on user-sensitive data. Mobile apps usually contain semantic documentation of meaningful programs. Leveraging this documentation, the authors designed an NLP-driven solution that locates the program elements (variables, methods) and performs an ML-based program structure analysis to detect the program element of apps carrying sensitive content. Following NLP techniques were used in their approach: (i) stemming, (ii) POS tagging, and (iii) dependency relation parsing.

Other means of privacy leaks in the technical domain are malicious hyperlinks pointing to various types of viruses, phishing texts to lure individuals into providing sensitive data such as personal information, banking, and credit card details, and passwords [46]. Fattahi *et al.* [46] put forward a new tool, called SpaML, for spam detection using a set of supervised and unsupervised classifiers, and two techniques imbued with NLP: Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). SpaML operates in two modes (BoW, TF-IDF) and utilizes seven supervised and unsupervised detectors: Multinomial Naive Bayes (MNB), Logistic Regression (LR), Support Vector Machine (SVM), Nearest Centroid (NCC), Extreme gradient boosting (Xgboost), K-Nearest Neighbors (KNN) and perceptron. In addition, it utilizes the majority of vote strategy to make the final decision founded on the prediction of its base learners.

Graph convolutional networks (GCNs) are a robust architecture for graph-based data representation such as citations, social networks. Nevertheless, they are prone to privacy leaks due to their training specifics. Igamberdiev *et al.* [47] proposed a method to apply differentially private stochastic gradient descent and its variants to GCNs, allowing to maintain strict privacy guarantees and performance. Also, they proposed a differentially private version of the Adam optimizer. They conducted experiments on five datasets in two languages (English and Slovak), covering a variety of NLP tasks, such as research article classification in citation networks, Reddit post-classification, and user interest classification in social networks.

Table 2 shows an overview of the works done related to privacy preservation in technological domain. For each work, it shows the year the work is published, dataset

[8] https://www.nltk.org/

[9] https://spacy.io/

used, the domain of the dataset, and models used in work.

IV. ANALYSIS OF PRIVACY POLICIES

A privacy policy is a statement that explains how an organization of an app or software collects, uses, retains, and discloses personal information. This is often called "privacy notice," "privacy statement," or "privacy terms." The privacy policies mainly contain the data-use practices of an app or software. Information privacy is built on the basic principle of *notice and choice*, meaning users should be able to make informed decisions about what information is collected and how it should be used [54]. In other words, the policies allow the users to read and decide to use a product or service only if they find the conditions acceptable. However, most of the privacy policies are lengthy, verbose, and challenging to understand. This imposes reading fatigue on the users, which plays an active role for the user in deciding on what app/software to use [3]. Furthermore, studies show that even if users do read the policies and understand, they would often still not be able to answer basic questions about what these policies say [55]. Recently, the growing number of online services and apps with privacy policies makes the situation more complicated. In addition, some app developers/owners exploit this situation to collect and misuse the personal information of the users [3]. There have been many techniques and proposals designed to make the policies user-friendly and increase user awareness, but the semantic complexity of the privacy terms, the length of the text, and the application-dependent variables still make this challenging. However, the above techniques are still insufficient to shape a coherent idea about app's/software's data gathering practice.

To address this, Alohaly *et al.* [3] proposed an approach to quantify the amount of data collected by an app by analyzing its privacy policy text using NLP techniques, and their proposed design not only allows the users to understand the policy easily but also allow them to compare with other applications in the market based on their data gathering practices. They used NLP techniques to analyze the privacy policy, extract potentially collected "information types" or "data items," which are noun phrases associated with collection practice, and then compare them against all possible information types. Then they normalized the resulted subset and initiated a four-step quantification process:

1) locate the text segments that are relevant to collection practices
2) extract noun phrases that are potentially collected items
3) compare the extracted noun phrases with the information types in the lexicon, using similarity measures
4) count the number of collected items

Studies on user preference modeling suggest that a few essential features in privacy policies largely determine the user's comfort level [56]. Researchers focused on using NLP to identify and extract essential fragments of a privacy policy to increase the ease of understanding for the user, such as Ammar *et al.* [57], Sadeh *et al.* [56], and Sathyendra *et al.* [54]. Ammar *et al.* [57] conducted a pilot experiment to estimate the extractability of salient features from website privacy policies. They combined NLP techniques and ML algorithms to extract the salient features. They utilized logistic regression, a classic high-performance probabilistic model, to map privacy policy documents to categorical labels. Both works of Sadeh *et al.* [55], [56] focus on developing an NLP framework to automate the extraction of vital information from the privacy policies to enable users more control of their privacy. They combine privacy preference modeling, crowd-sourcing, formal methods, and privacy interface design. Their objectives are to extract key privacy policy features semi-automatically and present them to users in an easy-to-digest format that enables them to make more informed privacy decisions. They used NLP techniques in pre-processing when crowd-sourcing reduces manual labor, filters out unnecessary text fragments and focuses on the relevant segments in a privacy policy. They also proposed augmenting crowd-sourcing results with ML algorithms and NLP techniques to develop the tools needed to extract answers to privacy terms questions automatically. Xiao *et al.* [58] adapted NLP techniques designed around a model to extract instances from software documents and produce formal specifications automatically. The linguistic-analysis component of their approach adapts the following NLP techniques that parse the software documents and annotate the words and phrases in the document sentences with semantic meaning: shallow parsing, utilizing domain dictionary, anaphora resolution, negative-expression identification, syntactic and semantic-pattern matching. Sathyendra *et al.* [54] focused on identifying and extracting *choice* instances automatically, which allow users to choose statements in a policy that give them discretion over aspects of their privacy. They focused on a two-stage ML procedure and treated the identification of choice instances as a binary classification problem, where they label each sentence in the text whether it contains a choice instance. They further annotated another dataset[11] and developed a hybrid model architecture to identify and label different types automatically. They used the following NLP techniques for feature selection: stemmed unigrams, stemmed bigrams, relative location in the document, topic model features, modal verbs, opt-out specific phrases, and syntactic parse tree features. They then used a two-stage architecture of ML models for classification.

Few researchers developed a corpus or lexicon (vocabulary of a language or a branch of knowledge) to support and improve the analysis of privacy policies. For example, Bhatia *et al.* [59] conducted a study and developed an information type lexicon based on privacy policy annotations obtained from crowd-sourcing entity extractor based on POS tagging. Using the lexicon, they suggested performing a richer analysis of policies or measure the degree of ambiguity. Lexicon construction consists of three parts: 1) obtain manual annotations, 2) entity extraction, and 3) lexicon construction.

[11] https://www.usableprivacy.org/data

TABLE 2. Overview of the works done related to privacy preservation in technology domain.

Year	Paper	Dataset	Domain	Model
2017	Cappellari, et al. [42]	Twitter sample stream	Social media platforms	Nearest neighbor, Rule induction, Random forest, Naive bayes, SVM
2018	Canfora, et al. [41]	Facebook statuses	Social networks	Logistic regression, Simple logistic, J48, FT, Random forest, Naive bayes
2018	Nan, et al. [45]	popular android apps	Mobile apps	Pattern matching
2020	Silva, et al. [2]	Kaggle[10]	Online websites	NER using NLTK, Stanford CoreNLP, and spaCy
2021	Fattahi, et al. [46]	SMS dataset [48]	SMS	MNB, LR, NCC, Xgboost, KNN, SVM, perceptron
2021	Igamberdiev, et al. [47]	Cora, Citeseer, PubMed [49]–[51], Reddit [52], Pokec [53]	Social networks	GCN

Automatically aligning segments of the policies makes it more comprehensible for the users. Liu *et al.* [60] contributed to an improved annotated dataset for pairwise evaluation of automatic methods and an exploration of clustering and HMM-based alignment methods. They employed a first-order Hidden Markov Model (HMM) and POS tagging. Recently, Ravichander *et al.* [61] presented a corpus named *PRIVACYQA* consisting of 1750 questions about the privacy policies of mobile applications and over 3500 expert annotations of relevant answers to aid the development of QA methods in the privacy domain. They further evaluated ML methods' ability to identify relevant evidence for questions in the privacy domain by establishing three baselines: ML - SVM, DL - CNN, and BERT.

Furthermore, some app developers collect data about their users and share it with advertising companies to raise revenue, which serves as targeted ads to end-users [62]. Given the size of the app market places verifying the third-party data recipients in each policy is a tedious task. Therefore, Hosseini *et al.* [62] developed an automated approach to extract and categorize third-party data recipients (i.e., entities) declared in privacy policies. They characterized the detection and classification of third-party entities as a NER problem, utilized Stanford CoreNLP for tokenization. Further, they used POS tags to identify each token, utilized Bag-of-Words (BoW) and Word2Vec [63] for vectorization, then passed into a Bi-LSTM-CRF model for classification. Word2Vec is a technique used to deliver distributed representation of words by studying the word associations.

In Europe, privacy policies are subject to compliance with GDPR. Since manual completeness checking is both time-consuming and error-prone, Torre *et al.* [64] proposed an AI-based automation system for the completeness checking of privacy policies recently. First, they built two artifacts to characterize the privacy-related provisions of GDPR then, they developed an automated solution on top of these artifacts with a combination of NLP and supervised ML. Their NLP pipeline combines six consecutive NLP modules divided into three categories:

1) Parsing the policy text - tokenization, sentence splitting
2) Extracting information from the text - NER, regular expressions
3) Normalizing text - lemmatization, stop words removal,

Finally, they utilized SVM for multi-class and multi-label classification.

Table 3 shows an overview of the works described here that are related to the analysis of privacy policies. For each work, the table includes the year the work is published, the dataset used, and the domain of the policies the dataset came from.

V. PRIVACY LEAKS DETECTION IN TEXT REPRESENTATION

Writing styles vary from person to person. This variation is mainly due to the authors' background and personal attributes such as gender, age, education, and nationality [40]. Therefore, a written text often leaves enough clues that can lead to the identification of the author. This situation can lead to problems when these texts are used to train NLP models [40]:

1) Variations in the text eventually lead to significant variation in inferences across different types of corpora. Moreover, models that fit these datasets would be biased.
2) The texts in the data compromise the authors' privacy, especially data collected from emails, SMS messages, social media posts, and medical records.
3) The latent representations generated from these data can still have sensitive information, which can fall into the hands of an adversary who can reverse engineer and gain the information.

Figure 1 illustrates a possible attack where an adversary could listen to the latent representation in the middle and obtain the sensitive information. For example, the classifier predicts class y from text x, and an adversary tries to recover the private information z in x through the classifier's latent representation. The naive solution for these attacks is removing protected attributes which is insufficient as other features may be highly correlated with the protected attributes [67]. Several works have been done in the past that deal with adversarial attacks NLP-based systems to prevent sensitive information leaks through representations.

Alawad *et al.* [68] used a DL-based approach to automatically extract cancer characteristics from the high volume of unstructured pathology text reports of cancer registries. They used a multitask CNN method, and the privacy-preserving model outperformed the single registry model in preserving privacy. Li *et al.* [40] proposed an approach for privacy-preserving learning of unbiased representations to explicitly obscure individuals' private information. They employed adversarial learning models inspired by Ganin *et al.* [69] for domain adaptation. This suggests that, for effective domain transfer to be achieved, predictions must be made based on features that cannot discriminate between the training (source) and test (target) domains. They jointly learn a discriminator model along with a supervised model

TABLE 3. Overview of the works done related to analysis of privacy policies.

Year	Paper	Dataset	Domain
2012	Ammar, et al. [57]	Crowd-sourced websites [57]	Websites
2012	Xiao, et al. [58]	iTrust, IBM enterprise application, published papers and public websites	Software documents
2013	Sadeh, et al. [56], [55]		Mobile apps
2014	Liu, et al. [60]	Websites ranked by Alexa[12]	Websites
2015	Bhatia, et al. [59]	Information type lexicon	Websites
2016	Alohaly, et al. [3]	Crowd-sourced websites [57]	Websites
2017	Sathyendra, et al. [54]	OPP-115 Corpus [65]	Websites
2019	Ravichander, et al. [61]	PRIVACYQA [61]	Mobile apps
2020	Hosseini, et al. [62]	App policies from Google Play Store[13]	Mobile apps
2021	Amaral, el al. [64]	Web service or App policies	Web service

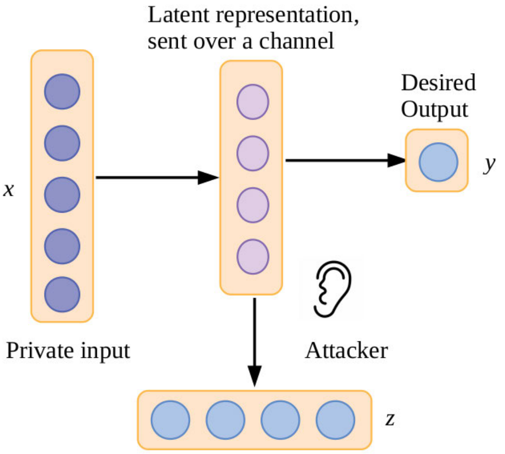

FIGURE 1. An illustration of a possible attack situation [66].

and aim for a good prediction of the target and a poor representation of the sensitive information.

Coavoux *et al.* [66] proposed a metric to measure the privacy of the neural representation of input for many NLP tasks such as sentiment analysis and topic classification. The metric they used is based on an attacker's ability (performance of the attacker's classifier) to recover information about the input from the latent representation. They presented three defense mechanisms designed against this type of attack by minimizing some measure of information and making it hard for the adversary to predict three training methods: multidetasking, adversarial generation, and de-clustering.

Both the above works provide only empirical improvements in privacy without any formal guarantees. Therefore, researchers moved into building systems in the context of Differential Privacy (DP) that provides formal privacy guarantee of the extracted representation from the user-authored text [70]. Lately, DP has become a de facto standard for privacy analysis, where researchers introduce noise into the data to make data related to specific people more difficult to trace. DP algorithms guarantee that the algorithm's behavior hardly changes when a single individual joins or leaves the dataset. Lyu *et al.* [70] proposed a novel approach called Differentially Private Neural Representation (DPNR), which utilizes DP to provide a formal privacy guarantee. They introduced a DP noise layer to preserve the extracted test representation's privacy without degrading the main classification task. They controlled how much noise to add for the robust algorithm through this layer. Fernandes *et al.* [71] combined ideas from "generalized DP" and ML techniques to model privacy for text processing. They demonstrated how to use ideas from differential privacy to provide strong a priori privacy guarantees in document disclosures. Here, they used BoW for text document representation as they contain sufficient information for the representation, and they used $d_X - privacy$ [72] a metric-based extension of differential privacy, to implement an automated privacy mechanism. The mechanism takes the BoW as input and produces *noisy* BoW outputs.

Pre-trained contextualized language models have been shown to increase the performance of several NLP tasks, but existing text sanitization mechanisms still provide low utility, as cursed by the high-dimensional text representation [73]. Yue *et al.* [73] built a privacy-preserving NLP (PPNLP) pipeline to address privacy from the root to produce sanitized text documents directly. Here they sanitize the public data before feeding them to training because they believe it prepares the model to work with sanitized queries, increasing accuracy. They proposed two token-wise sanitization methods: *SANTEXT* and $SANTEXT^+$, which were built atop a variant of the exponential mechanism (EM) [74] to avoid going to the "cursed dimensions" of token embeddings. Finally, they passed the output token into BERT for classification.

Recently in NLP, building general-purpose language models such as ELMo [75], BERT [76], and Generative Pre-trained Transformer-2 (GPT-2) [77] to convert text to vectors has become successful. Nevertheless, these embeddings from general-purpose language models would also capture much sensitive information from the plain text and be a potential risk. Pan *et al.* [78] is the first to present a systematic study on the privacy risks of eight state-of-the-art language models by constructing two novel attack scenarios such as pattern reconstruction attacks and keyword inference attacks. Pattern reconstruction attack aims to recover a specific segment of the plain text with a fixed format like date of birth or gender, and keyword inference attack aims to infer the sensitive information using the existing words in the text. Through the study, they confirm the existence of privacy risks. Also, they proposed four different defense mechanisms to obscure the unprotected embeddings for alleviation purposes as follows:

1) *Rounding* - Apply floating-point Rounding on each coordinate of the sentence embeddings for obfuscation.
2) *Laplace Mechanism* - Perturb the embedding coordi nate-wise with samples from a Laplace distribution whose parameters are determined by the sensitivity of the language model.
3) *Privacy-Preserving Mapping* - Apply adversarial training as mentioned by Li *et al.* [40].
4) *Subspace Projection* - Remove the unwanted subspace that encodes the keyword's occurrence from the universal sentence embedding space.

Table 4 provides a summary of the works done to prevent privacy violations in the learned text representations. The table shows the year, authors, state-of-the-art datasets used for experiments, and the NLP tasks the datasets were evaluated for each paper.

VI. DISCUSSION

In this section, we provide a summary of all works we discussed in the above sections and our novel insights in the future directions we can take to tackle privacy issues with NLP-based techniques.

A. SUMMARY

Table 5 provides an overview of the works done in the privacy domain using NLP techniques in chronological order. For each paper reference, the table shows the year, authors, and the paper's main objectives. For ease of understanding, we grouped the papers into four categories as discussed in the above sections as follows:

- A - Data privacy in the medical domain
- B - Privacy preservation in the technology domain
- C - Analysis of privacy policies
- D - Privacy leaks detection in the text representation

B. FUTURE DIRECTIONS

So far, we have discussed previous works that used NLP-based techniques to address the data privacy challenges. Here, we discuss some privacy-related issues and the future directions we propose to utilize NLP techniques in privacy preservation.

This review focuses on the de-identification or anonymization of personal identifiers in the medical and technological domains. However, there are other domains where documents or artifacts are shared between institutions that contain personal identification details such as financial documents, Curriculum Vitae (CV), resumes. Therefore, the similar techniques that we discuss here can be enhanced to be applied for data from other domains. Also, here we focus on the personal identifiers only, but researchers could apply these techniques to identify quasi-identifiers. These quasi-identifiers are not unique identifiers themselves, but they create a unique identifier that correlates with specific entities.

During this review, few studies explored utilizing different word embeddings to capture different aspects of the input representation, such as Flair embeddings and MultiBPEmb embeddings. We should further explore utilizing different word embeddings, especially the deep contextualized word embeddings such as ELMo [75], BERT [76]. Since most of the datasets belong to the clinical or healthcare domain, we can specifically use BioBERT [91] or Clinical BERT [92]. Pre-trained word embeddings trained on these large-scale data help to represent the token more efficiently.

In the future, we can explore the possibility of utilizing transfer learning when studying data where we do not have much data. For example, most of the clinical or healthcare datasets we use to study ways to secure patients' privacy are smaller than other domains. Catelli *et al.* [25] investigated the effectiveness of transfer learning across languages. It would be interesting to explore transfer learning to learn on datasets with more instances and test on the dataset with fewer instances. Also, we can investigate new optimizations that can reduce the resource requirements and training data to analyze domains where we do not have much data [90].

In the course of our review, we noticed there had not been any research using NLP-based approaches for privacy preservation in Twitter data. Twitter is the second most popular social networking site, and Twitter data is used for research purposes in multiple domains such as political campaigns, movie reviews, industry-related reviews. These data can carry sensitive information about the users that can be exploited. NLP-based techniques can be used to remove or anonymize the personalized information from the tweets.

Another area we would like to focus on user data privacy is the location privacy of the users. Many apps and social media networks track the location details of the users. An adversary can use this data to link records of the same individual, study and predict the movement patterns of an individual, identify points of interest that can endanger a targeted individual [93]. In the future, more research should focus on preserving the privacy information from these data, and many NLP techniques can be applied to identify and extract user's location privacy information and normalize so that the information does not fall into the wrong hands.

Furthermore, we discussed developing user-friendly privacy policies. In the future, we can focus on improving the usability of privacy policies by extracting relevant data practices and making them more accessible to users. We can use information extraction techniques utilized in NLP-based research.

In the recent past, there was an urgency to manage and find cures for the COVID-19 pandemic. It was necessary to share large volumes of data between national and international organizations to share information for the studies [94]. We should look into efficient organizational and technical measures to remove or replace PIs in the Big Data

TABLE 4. Overview of the works done related to detect privacy violations in text representations.

Year	Paper	Dataset	NLP tasks
2018	Li, et al. [40]	TrustPilot English POS tagged dataset [79], Google Universal POS tagset [80], African-American Vernacular English (AAVE) [81]	POS-tagging, Sentiment analysis
2018	Coavoux, et al. [66]	TrustPilot English POS tagged dataset [79], AG news corpus [82], Deutsche Welle (DW) news corpus [83]	Sentiment analysis, Topic classification
2019	Fernandes, et al. [71]	*Fan fiction* dataset [84]	Authorship attribution, Topic identification
2020	Lyu, et al. [70]	TrustPilot English POS tagged dataset [79], AG news corpus [82], Blog posts dataset (BLOG) [85]	Sentiment analysis, Topic classification
2020	Pan, et al. [78]	Homo Sapiens Splice Sites Dataset (HS3D) [86], Airline review dataset from Skytrax[14], CMS public healthcare records[15]	Word representation
2021	Yue, et al. [73]	MedSTS [87], QNLI [88], SST-2 [89]	Sentiment Classification, Medical Semantic Textual Similarity, Question Natural Language Inference

TABLE 5. Overview of the state-of-the-art works done in the privacy domain related to NLP. Categories represent as follows: A - Data privacy in the medical domain, B - Privacy preservation in the technology domain, C - Analysis of privacy policies, D - Privacy leaks detection in the text representation.

Category	Year	Authors	Objective
A	1996	Sweeney, et al. [5]	utilized detection algorithms for PHI anonymization
A	2003	Berman, et al. [10]	proposed a concept-based scrubs pattern matching for PHI anonymization
A	2006	Beckwith,et al. [11]	designed a pattern matching tool for PHI anonymization
A	2006	Medlock, et al. [4]	proposed a feature extraction technique for PHI anonymization
A	2007	Szarvas, et al. [14]	proposed decision tree-based pattern matching approach for PHI anonymization
C	2012	Anmar, et al.	conducted experiment to estimate the extractability of salient features from privacy policies
C	2012	Xiao, et al. [58]	proposed approach which adapts NLP techniques to auto-extract instances from software documents
C	2013	Sadeh et al. [55]	proposed algorithm to answer privacy questions of users semi-automatically
C	2014	Sadeh et al. [56]	developed NLP framework to auto-extract vital information from privacy policies
C	2014	Breaux, et al. [90]	mapped privacy requirements to a formal language description
C	2014	Liu, et al. [60]	contributed to an improved annotated dataset for pairwise evaluation of automatic methods
A	2014	He, et al. [16]	proposed a CRF-based system for patient anonymization in clinical narratives
A	2014	Liu, et al. [18]	proposed CRF-based pattern matching system for patient anonymization in clinical narratives
A	2014	Yang, et al. [19]	proposed CRF-based pattern matching for patient anonymization in clinical narratives
A	2014	Grouin, et al. [17]	proposed CRF-based pattern matching system for patient anonymization in clinical narratives
A	2015	Li,et al. [29]	proposed frequency-filtering approach for patient anonymization
C	2015	Bhatia,et al. [59]	developed information type lexicon based on privacy policy annotations
C	2016	Alohaly, et al. [3]	proposed algorithm to quantify the amount of data collection of an application
C	2017	Sathyendra, et al. [54]	built a 2-stage classifier using feature selection
A	2017	Jiang, et al. [21]	proposed a CRF and LSTM-based system for patient anonymization in psychiatric evaluation records
A	2017	Dernoncourt, et al. [20]	proposed CRF and LSTM-based systems for patient anonymization in psychiatric evaluation records
A	2017	Dernoncourt, et al. [23]	designed a CRF and LSTM-based for patient de-identification
A	2017	Dobbins, et al. [23]	utilized a CRF and LSTM-based tool [23] to compare performance of datasets
A	2017	Dernoncourt, et al. [20]	proposed a CRF and LSTM-based system for PHI anonymization
B	2017	Cappellari, et al. [42]	built privacy protection framework with ML algorithms
B	2018	Canfora, et al. [41]	designed tool with ML algorithms to intercept private information in social media post
B	2018	Nan, et al. [45]	proposed pattern matching -based solution to auto-detect the code operating on private data in mobile apps
D	2018	Li, et al. [40]	proposed approach to train model for adversarial training in parallel
D	2018	Coavoux, et al. [66]	proposed metric to measure the privacy of the neural representation of input text
C	2019	Ravichander, et al. [61]	built a corpus for QA methods in privacy domain
A	2019	Sadat, et al. [7]	proposed homomorphic encryption for secure multi-party data analysis
D	2019	Fernandes, et al. [71]	proposed a combined approach of generalised DP and ML to model privacy for text documents
D	2020	Lyu, et al. [70]	proposed representation to formally quantify DP
D	2020	Pan, et al. [78]	presented systematic study on the privacy risks
C	2020	Hosseni, et al. [62]	developed an automated approach to extract and categorize third-party data recipients
B	2020	Silva, et al. [2]	proposed NER using NLP-based tools to identify, monitor and validate PII
D	2020	Alawad, et al. [68]	designed privacy-preserving model using CNN
A	2020	Lopez, et al. [6]	designed pattern matching, dictionaries, and ML-powered web tool for auto-detection of PHI
A	2020	Iwendi, et al [12]	proposed semantic privacy framework to effectively sanitize sensitive terms in healthcare documents
D	2021	Yue, et al. [73]	proposed two token-wise sanitization methods for text sanitization
B	2021	Fattahi, et al. [46]	proposed a tool for spam detection
B	2021	Igamberdiev, et al. [47]	applied differentially private stochastic gradient descent to GCNs to maintain strict privacy guarantees
C	2021	Amaral, el al. [64]	proposed an AI-based automation system for the completeness checking of privacy policies
A	2021	Catelli, et al. [22]	combined contextualized word representation and sub-document level analysis for clinical de-identification
A	2021	Catelli, et al. [25]	cross-lingual transfer learning to de-identify medical records
A	2021	Moqurrab, et al. [28]	proposed model uses local and global context to extract clinical entities

applications in the era of COVID-19. We can conduct an inter-domain study to investigate ways to combine with NLP to increase efficiency.

VII. CONCLUSION

This inter-disciplinary review categorized state-of-the-art research in the privacy domain that utilized NLP-based techniques into four categories. We investigated methods to protect patients' health information in the medical domain through PHI anonymization and de-identification techniques. We analyzed techniques to educate individuals about potential privacy risks and building systems for privacy preservation in social media networks, software, and apps. We further looked into designs to make the policies user-friendly, increase user awareness, and quantify the sensitive information in the policies. Next, we studied methods that prevent an adversary from listening to the latent representation in the middle and obtaining sensitive information Finally, we provide a tabular summary of related work and discuss future directions to help guide a path ahead.

REFERENCES

[1] O. Feyisetan, S. Ghanavati, and P. Thaine, "Workshop on privacy in NLP (PrivateNLP 2020)," in *Proc. 13rd Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 903–904.

[2] P. Silva, C. Goncalves, C. Godinho, N. Antunes, and M. Curado, "Using NLP and machine learning to detect data privacy violations," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Jul. 2020, pp. 972–977.

[3] A. Manar and T. Hassan, "Better privacy indicators: A new approach to quantification of privacy policies," in *Proc. Symp. Usable Privacy Secur.*, 2016, pp. 1–7.

[4] M. Ben, "An introduction to NLP-based textual anonymisation," in *Proc. 5th Int. Conf. Lang. Resour. Eval.*, Genoa, Italy, May 2006, pp. 1051–1056.

[5] S. Latanya, "Replacing personally-identifying information in medical records, the scrub system," in *Proc. AMIA Annu. Fall Symp. Amer. Med. Inform. Assoc.*, 1996, p. 333.

[6] L. SalvadorLima, P. Naiara, G.-S. Laura, and C. Montse, "Hitzalmed: Anonymisation of clinical text in Spanish," in *Proc. 12nd Lang. Resour. Eval. Conf.*, 2020, pp. 7038–7043.

[7] M. N. Sadat, M. M. A. Aziz, N. Mohammed, S. Pakhomov, H. Liu, and X. Jiang, "A privacy-preserving distributed filtering framework for NLP artifacts," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–10, Dec. 2019.

[8] C. Laura, L. Yunyao, and R. Frederick, "Rule-based information extraction is dead long live rule-based information extraction systems," in *Proc. Conf. Empirical Methods Natural Lang. Process.* 2013, pp. 827–832.

[9] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[10] B. Jules, "Concept-match medical data scrubbing: How pathology text can be used in research," *Arch. Pathol. Lab. Med.*, vol. 127, no. 6, pp. 680–686, 2003.

[11] B. A. Beckwith, R. Mahaadevan, U. J. Balis, and F. Kuo, "Development and evaluation of an open source software tool for deidentification of pathology reports," *BMC Med. Informat. Decis. Making*, vol. 6, no. 1, p. 12, Dec. 2006.

[12] C. Iwendi, S. A. Moqurrab, A. Anjum, S. Khan, S. Mohan, and G. Srivastava, "N-sanitization: A semantic privacy-preserving framework for unstructured medical datasets," *Comput. Commun.*, vol. 161, pp. 160–171, Sep. 2020.

[13] V. Veronika and F. Richárd, "De-identification in natural language processing," in *Proc. 37th Int. Conv. Inf. Commun. Technol., Electron. Microelectron.*, 2014, pp. 1300–1303.

[14] G. Szarvas, R. Farkas, and R. Busa-Fekete, "State-of-the-art anonymization of medical records using an iterative machine learning framework," *J. Amer. Med. Inform. Assoc.*, vol. 14, no. 5, pp. 574–580, Jun. 2007.

[15] A. Stubbs, C. Kotfila, and Ö. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1," *J. Biomed. Informat.*, vol. 58, pp. S11–S19, Dec. 2015.

[16] B. He, Y. Guan, J. Cheng, K. Cen, and W. Hua, "CRFs based de-identification of medical records," *J. Biomed. Informat.*, vol. 58, pp. S39–S46, Dec. 2015.

[17] G. Cyril, "Clinical records de-identification using CRF and rule-based approaches," presented at the 7th i2b2 Shared Task Workshop, Natural Language Process. Clin. Data, 2014.

[18] Z. Liu, Y. Chen, B. Tang, X. Wang, Q. Chen, H. Li, J. Wang, Q. Deng, and S. Zhu, "Automatic de-identification of electronic medical records using token-level and character-level conditional random fields," *J. Biomed. Informat.*, vol. 58, pp. S47–S52, Dec. 2015.

[19] H. Yang and J. M. Garibaldi, "Automatic detection of Protected health information from clinic narratives," *J. Biomed. Informat.*, vol. 58, pp. S30–S38, Dec. 2015.

[20] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits, "De-identification of patient notes with recurrent neural networks," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 3, pp. 596–606, May 2017.

[21] Z. Jiang, C. Zhao, B. He, Y. Guan, and J. Jiang, "De-identification of medical records using conditional random fields and long short-term memory networks," *J. Biomed. Informat.*, vol. 75, pp. S43–S53, Nov. 2017.

[22] R. Catelli, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "Combining contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical de-identification," *Knowl.-Based Syst.*, vol. 213, Feb. 2021, Art. no. 106649.

[23] F. Dernoncourt, J. Y. Lee, and P. Szolovits, "NeuroNER: An easy-to-use proGram for named-entity recognition based on neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2017, pp. 97–102.

[24] D. Nicholas, W. David, L. Kahyun, U. Özlem, and Y. Meliha, "Performance of automatic de-identification across different note types," *CoRR*, vol. abs/2102.11032, pp. 1–2, Feb. 2021.

[25] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106779.

[26] A. Alan, B. Duncan, and V. Roland, "Contextual string embeddings for sequence labeling," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018 pp. 1638–1649.

[27] H. Benjamin and S. Michael, "BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, Miyazaki, Japan, May 2018, pp. 1–5.

[28] S. A. Moqurrab, U. Ayub, A. Anjum, S. Asghar, and G. Srivastava, "An accurate deep learning model for clinical entity recognition from clinical notes," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3804–3811, Oct. 2021.

[29] D. Li, M. Rastegar-Mojarad, R. K. Elayavilli, Y. Wang, S. Mehrabi, Y. Yu, S. Sohn, Y. Li, N. Afzal, and H. Liu, "A frequency-filtering strategy of obtaining PHI-free sentences from clinical data repository," in *Proc. 6th ACM Conf. Bioinf., Comput. Biol. Health Informat.*, Sep. 2015, pp. 315–324.

[30] I. S. Kohane, "Getting the data in: Three year experience with a pediatric electronic medical record system," in *Proc. Annu. Symp. Comput. Appl. Med. Care*, 1994, p. 457.

[31] Ö. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *J. Amer. Med. Informat. Assoc.*, vol. 14, no. 5, pp. 550–563, 2007.

[32] A. Stubbs, C. Kotfila, H. Xu, and Ö. Uzuner, "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task track 2," *J. Biomed. Informat.*, vol. 58, pp. S67–S77, Dec. 2015.

[33] C. G. Chute, S. A. Beck, T. B. Fisk, and D. N. Mohr, "The enterprise data trust at mayo clinic: A semantically integrated warehouse of biomedical data," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 2, pp. 131–135, Mar. 2010.

[34] A. Stubbs, M. Filannino, and Ö. Uzuner, "De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1," *J. Biomed. Informat.*, vol. 75, pp. S4–S18, Nov. 2017.

[35] E. F. Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proc. 7th Conf. Natural Lang. Learn.*, vol. 4, 2003, pp. 142–147.

[36] J. Alistair, P. Tom, S. Lu, L. Li-wei, F. Mengling, and G. Mohammad, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, May 2016, Art. no. 160035.

[37] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, ''A novel COVID-19 data set and an effective deep learning approach for the de-identification of Italian medical records,'' *IEEE Access*, vol. 9, pp. 19097–19110, 2021.

[38] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, ''2010 i2b2/VA cHallenge on concepts, assertions, and relations in clinical text,'' *J. Amer. Med. Informat. Assoc.*, vol. 18, no. 5, pp. 552–556, Jun. 2011.

[39] W. Sun, A. Rumshisky, and O. Uzuner, ''Evaluating temporal relations in clinical text: 2012 i2b2 cHallenge,'' *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 5, pp. 806–813, 2013.

[40] L. Yitong, B. Timothy, and C. Trevor, ''Towards robust and privacy-preserving text representations,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, Melbourne, VIC, Australia, 2018, pp. 25–30.

[41] G. Canfora, A. Di Sorbo, E. Emanuele, S. Forootani, and C. A. Visaggio, ''A NLP-based solution to prevent from privacy leaks in social network posts,'' in *Proc. 13rd Int. Conf. Availability, Rel. Secur.*, Aug. 2018, pp. 1–6.

[42] P. Cappellari, S. A. Chun, and M. Perelman, ''A tool for automatic assessment and awareness of privacy disclosure,'' in *Proc. 18th Annu. Int. Conf. Digit. Government Res.*, Jun. 2017, pp. 586–587.

[43] G. A. Miller, ''WordNet: A lexical database for English,'' *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[44] M. Christopher, S. Mihai, B. John, F. J. Rose, B. Steven, and M. David, ''The Stanford CoreNLP natural language processing toolkit,'' in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2014, pp. 55–60.

[45] N. Yuhong, Y. Zhemin, W. Xiaofeng, Z. Yuan, Z. Donglai, and Y. Min, ''Finding clues for your secrets: Semantics-driven, learning-based privacy discovery in mobile apps,'' in *Proc. NDSS*, Feb. 2018, pp. 1–15.

[46] J. Fattahi and M. Mejri, ''SpaML: A bimodal ensemble learning spam detector based on NLP techniques,'' in *Proc. IEEE 5th Int. Conf. Cryptogr., Secur. Privacy (CSP)*, Jan. 2021, pp. 107–112.

[47] I. Timour and H. Ivan, ''Privacy-preserving graph convolutional networks for text classification,'' *CoRR*, vol. abs/2102.09604, pp. 1–14, Feb. 2021.

[48] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, ''Contributions to the study of SMS spam filtering: New collection and results,'' in *Proc. 11st ACM Symp. Document Eng.*, 2011, pp. 259–262.

[49] Y. Zhilin, C. William, and S. Ruslan, ''Revisiting semi-supervised learning with graph embeddings,'' in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 40–48.

[50] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, ''Collective classification in network data,'' *AI Mag.*, vol. 29, no. 3, p. 93, 2008.

[51] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, ''Automating the construction of internet portals with machine learning,'' *Inf. Retr.*, vol. 3, no. 2, pp. 127–163, 2000.

[52] H. WilliamL, Y. Rex, and L. Jure, ''Inductive representation learning on large graphs,'' in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*,2017, pp. 1025–1035.

[53] T. Lubos and Z. Michal, ''Data analysis in public social networks,'' in *Proc. Conf. Int. Workshop Trends Innov.*, vol. 1, 2012, p. 6.

[54] K. Mysore Sathyendra, S. Wilson, F. Schaub, S. Zimmeck, and N. Sadeh, ''Identifying the provision of choices in privacy policy text,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2774–2779.

[55] S. Norman, A. Alessandro, B. Travis, C. L. Faith, M. Aleecia, and R. Joel, ''The usable privacy policy project,'' Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-ISR-13-119 2013.

[56] S. Norman, A. Alessandro, B. Travis, C. L. Faith, M. Aleecia, and R. Joel, ''Towards usable privacy policies: Semi-automatically extracting data practices from websites' privacy policies,'' in *Proc. SOUPS*, 2014, pp. 9–11.

[57] A. Waleed, W. Shomir, S. Norman, and S. Noah, ''Automatic categorization of privacy policies: A pilot study,'' School Comput. Sci., Lang. Technol. Inst., Pittsburgh, PA, USA, Tech. Rep. CMU-LTI-12-019, 2012.

[58] X. Xiao, A. Paradkar, S. Thummalapenta, and T. Xie, ''Automated extraction of security policies from natural-language software documents,'' in *Proc. ACM 20th Int. Symp. Found. Softw. Eng.*, 2012, pp. 1–11.

[59] B. Jaspreet and B. Travis, ''Towards an information type lexicon for privacy policies,'' in *Proc. IEEE Int. Workshop Requirement Eng. Law (RELAW)*, Oct. 2015, pp. 19–24.

[60] L. Fei, R. Rohan, S. Norman, and S. Noah, ''A step towards usable privacy policy: Automatic alignment of privacy statements,'' in *Proc. COLING*, 2014, pp. 884–894.

[61] A. Ravichander, A. W. Black, S. Wilson, T. Norton, and N. Sadeh, ''Question answering for privacy policies: Combining computational and legal perspectives,'' in *Proc. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 4947–4958.

[62] H. M. Bokaie, R. Irwin, and E. Serge, ''Identifying and classifying third-party entities in natural language privacy policies,'' in *Proc. 2nd Workshop Privacy*, 2020, pp. 18–27.

[63] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, ''Distributed representations of words and phrases and their compositionality,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[64] D. Torre, S. Abualhaija, M. Sabetzadeh, L. Briand, K. Baetens, P. Goes, and S. Forastier, ''An AI-assisted approach for checking the completeness of privacy policies against GDPR,'' in *Proc. IEEE 28th Int. Requirements Eng. Conf. (RE)*, Aug. 2020, pp. 136–146.

[65] L. Frederick, W. Shomir, S. Florian, and S. Norman, ''Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies,'' in *Proc. AAAI*, 2016, pp. 1–6.

[66] M. Coavoux, S. Narayan, and S. B. Cohen, ''Privacy-preserving neural representations of text,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1–10.

[67] D. Pedreshi, S. Ruggieri, and F. Turini, ''Discrimination-aware data mining,'' in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 560–568.

[68] M. Alawad, H.-J. Yoon, S. Gao, B. Mumphrey, X.-C. Wu, E. B. Durbin, J. C. Jeong, I. Hands, D. Rust, L. Coyle, L. Penberthy, and G. Tourassi, ''Privacy-preserving deep learning NLP models for cancer registries,'' *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 3, pp. 1219–1230, Jul. 2021.

[69] Y. Ganin, E. Ustinova, H. Ajakan, and P. Germain, ''Domain-adversarial training of neural networks,'' *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.

[70] L. Lyu, X. He, and Y. Li, ''Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness,'' in *Proc. Findings Assoc. Comput. Linguistics*, 2020, pp. 2355–2365.

[71] F. Natasha, D. Mark, and M. Annabelle, ''Generalised differential privacy for text document processing,'' in *Proc. Int. Conf. Princ. Secur. Trust*. Cham, Switzerland: Springer, 2019, pp. 123–148.

[72] C. Konstantinos, A. MiguelE, B. NicolásEmilio, and P. Catuscia, ''Broadening the scope of differential privacy using metrics,'' in *Proc. Int. Symp. Privacy Enhancing Technol. Symp.* New York, NY, USA: Springer, 2013, pp. 82–102.

[73] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S. S. M. Chow, ''Differential privacy for text analytics via natural text sanitization,'' in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 3853–3866.

[74] F. McSherry and K. Talwar, ''Mechanism design via differential privacy,'' in *Proc. 48th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2007, pp. 94–103.

[75] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, ''Deep contextualized word representations,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 2227–2237.

[76] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.

[77] R. Alec, W. Jeffrey, C. Rewon, L. David, A. Dario, and S. Ilya, ''Language models are unsupervised multitask learners,'' in *Proc. OpenAI Blog*, 2019, vol. 1, no. 8, p. 9.

[78] X. Pan, M. Zhang, S. Ji, and M. Yang, ''Privacy risks of general-purpose language models,'' in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2020, pp. 1314–1331.

[79] D. Hovy and A. Søgaard, ''Tagging performance correlates with author age,'' in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2015, pp. 483–488.

[80] S. Petrov, D. Das, and R. McDonald, ''A universal part-of-speech tagset,'' 2011, *arXiv:1104.2086*.

[81] A. Jørgensen, D. Hovy, and A. Søgaard, ''Learning a POS tagger for AAVE-like language,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1115–1120.

[82] G. M. Del Corso, A. Gullí, and F. Romani, ''Ranking a stream of news,'' in *Proc. 14th Int. Conf. World Wide Web*, 2005, pp. 97–106.

[83] P. Nikolaos and P.-B. Andrei, ''Multilingual hierarchical attention networks for document classification,'' in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 1015–1025.

[84] P. Martin, R. Francisco, T. Michael, S. Efstathios, R. Paolo, and S. Benno, "Overview of PAN'17," in *Proc. Int. Conf. Cross-Lang. Eval. Forum Eur. Lang.* New York, NY, USA: Springer, 2017, pp. 275–290.

[85] S. Jonathan, K. Moshe, A. Shlomo, and P. JamesW, "Effects of age and gender on blogging," in *Proc. AAAI Spring Symp., Comput. Approaches Analyzing Weblogs*, vol. 6, 2006, pp. 199–205.

[86] P. Pasquale and R. Salvatore, "HS^{3}D: Homo sapiens splice site data set," Tech. Rep., 2003.

[87] Y. Wang, N. Afzal, S. Fu, L. Wang, F. Shen, M. Rastegar-Mojarad, and H. Liu, "MedSTS: A resource for clinical semantic textual similarity," *Lang. Resour. Eval.*, vol. 54, no. 1, pp. 57–72, Mar. 2020.

[88] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. EMNLP Workshop BlackboxNLP, Analyzing Interpreting Neural Netw.*, 2018, pp. 353–355.

[89] S. Richard, P. Alex, W. Jean, C. Jason, M. Christopher, and D. N. Andrew, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.

[90] B. TravisD, H. Hanan, and R. Ashwini, "Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements," *Requirements Eng.*, vol. 19, no. 3, pp. 281–307, 2014.

[91] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, and C. H So, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[92] A. Emily, M. John, B. Willie, W. WeiHung, J. Di, and N. Tristan, "Publicly available clinical BERT embeddings," *CoRR*, vol. abs/1904.03323, pp. 1–7, Apr. 2019.

[93] G. Beigi and H. Liu, "A survey on privacy in social media: Identification, mitigation, and applications," *ACM/IMS Trans. Data Sci.*, vol. 1, no. 1, pp. 1–38, Mar. 2020.

[94] B. Vibhushinie, H. Chaminda, and W. Jason, "Solutions to big data privacy and security challenges associated with COVID-19 surveillance systems," *Frontiers in Big Data*, to be published.

DARSHINI MAHENDRAN received the B.Sc. degree in computer science from the University of Peradeniya, in 2015. She is currently pursuing the Ph.D. degree with the Natural Language Processing Laboratory, Virginia Commonwealth University (VCU), USA. Her research interests include relation extraction, information retrieval, and machine learning.

CHANGQING LUO (Member, IEEE) received the B.E. and M.E. degrees from the Chongqing University of Posts and Telecommunications, in 2004 and 2007, respectively, the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2011, and the Ph.D. degree from Case Western Reserve University, in 2018. He is currently an Assistant Professor with the Department of Computer Science, Virginia Commonwealth University. His research interests include cybersecurity, big data, and wireless networking. He is a member of the ACM.

BRIDGET T. MCINNES received the B.S. and M.S. degrees in computer science from the University of Minnesota Duluth, in 2002 and 2004, respectively, and the Ph.D. degree in computer science from the University of Minnesota Twin Cities, in 2009. She is currently an Associate Professor in computer science with Virginia Commonwealth University and the Director of the Natural Language Processing Laboratory.

• • •