Personal Information Privacy: What's Next?

Khodor Hammoud, Salima Benbernou, Mourad Ouziri, Yücel Saygin, Rafiqul Haque, Yehia Taher

Personal Information Privacy: What's Next?

1st Khodor Hammoud
Université de Paris
Paris, France
ik19544@etu.parisdescartes.fr

2nd Salima Benbernou
Université de Paris,
Paris, France
salima.benbernou@parisdescartes.fr

3rd Mourad Ouziri
Université de Paris Paris,
Paris, France
mourad.ouziri@parisdescartes.fr

4th Yucel Saygin
Sabanci University
Istanbul, Turkey
ysaygin@sabanciuniv.edu

5th Rafiqul Haque
Intelligencia R&D
Paris, France
Rafiqul.haque@intelligenciaia.fr

6th Yehia Taher
Laboratoire DAVID
Université de Versailles – Paris-Saclay
yehia.taher@uvsq.fr

***Abstract*—In recent events, user-privacy has been a main focus for all technological and data-holding companies, due to the global interest in protecting personal information. Regulations like the General Data Protection Regulation (GDPR) set firm laws and penalties around the handling and misuse of user data. These privacy rules apply regardless of the data structure, whether it being structured or unstructured.**

In this work, we perform a summary of the available algorithms for providing privacy in structured data, and analyze the popular tools that handle privacy in textual data; namely medical data. We found that although these tools provide adequate results in terms of de-identifying medical records by removing personal identifyers (HIPAA PHI), they fall short in terms of being generalizable to satisfy nonmedical fields. In addition, the metrics used to measure the performance of these privacy algorithms don't take into account the differences in significance that every identifier has.

Finally, we propose the concept of a domain-independent adaptable system that learns the significance of terms in a given text, in terms of person identifiability and text utility, and is then able to provide metrics to help find a balance between user privacy and data usability.

***Index Terms*—privacy, k-anonymity, l-diversity, t-closeness, NLP, textual data, privacy in text**

I. Introduction And Motivation

The legal right to privacy is a fundamental human right recognized in the UN (United Nation) declaration of human rights [52]. The unprecedented growth of highly advanced technologies – in the last two decades – has imperiled privacy significantly. Today, different aspects of human life has been digitalized including communication medium, socialization, entertainment, purchasing and many others. People adopted the digital systems due to increasing efficiency in day-to-day tasks. In some cases the adoption is forced by social practices such as the use of social media. Nevertheless, the digital transformation created an ample of opportunities for various organizations and adversaries to abuse privacy since the digital systems enable them to hold information of people forever. Organizations such as Google can profile anyone without the users being aware of it. The concrete evidence of personal information abuse are the two recent incidents: Facebook trail [46] and google testimony [47]. Its not only the software vendors, social medias, hardware companies violates the privacy as well. As an example, Samsung's smart TVs recording audio [51], and the recent events that happened unavailing the potential of the giant Chinese tech company Huawei using its mobile phone to spy on users, blocking its phones from using google services and banning them from the US [50]. There are regulations that govern user data handling, latest of which is the European's General Data Protection Regulation (GDPR) [49] which needs transparency, and user-anonymity when performing statistical analysis, and places heavy fines on violating parties.

The task of making the web a safe place for users is a largely difficult problem due to the inherently open, nondeterministic nature of the Web, and the complex, leakage-prone information flow of many Web-based transactions that involve the transfer of sensitive, personal information. Despite considerable attention, Web privacy continues to pose significant threats and challenges. One major step is securing the way companies store, share and publish user information, as data regulations impose data publication, which if not secured, can be used to re-identify the individual owners. Securing stored/published data depends on the way data is stored. In the past, information was almost strictly stored in the form of structured relational databases [44]. Consequently, shared data was in the form of structured datasets. Ensuing privacy to these datasets was first in the form of deleting the unique identifiers, but then L. Sweeney [28] published a research result that proved that users can still be identified from their quazi-identifiers, and proposed a new methodology known as k-anonymity [1]. Following K-anonymity, several solutions were proposed including *l*-diversity [3] and *t*-closeness [4] that address shortcomings discovered in k-anonymity. However, in 2006, Dwork and Aaron introduced differential privacy [6] as a solution for privacy-preserving data analysis which can be used to provide security for both data storage and analysis.

Recently, the changes in applications, user and infrastructure characteristics, mostly of the Web 2.0 domain [53] and cloud

This work was made possible thanks to the funding provided by Cognitus.

platform, led to an exponential growth of the internet and the explosion of data sources such as sensors, social media, etc. and massive workloads. This kind of data is typically referred to as Big Data [7]. This foster the requirement of a new format of data storage, known as unstructured data [48] which is essentially the central focus of this research. To be specific, the textual form of this unstructured data is the key focus of this research. Privacy in unstructured text is critically important for several reasons yet the most substantial reason is the amount of unstructured data generated by companies. More than 80% of the data generated in the last ten years are unstructured (mostly in textual form). This implies the fact that a massive volume of data is recorded in textual form yet the privacy in unstructured text, to the best of our knowledge, lack robust solutions. We studied several use cases that belong to different industrial domains including finance, healthcare, and insurance etc. According to our study, banking and healthcare sectors generate a huge volume of unstructured text; both of these industrial domains are facing several challenge concerning privacy of user information - which is the key motivation of this research. The need of a privacy mechanism for unstructured data confidentiality exceeds expectations, especially for textual data in the healthcare sector [8] [9]. A large number of researches in this field thrived aiming at providing anonymity in text. Many works rose that provide different privacy solutions for text, mostly focusing on medical data, governed by the regulations placed by the Health Insurance Portability and Accountability Act (HIPAA) [36]. Older work used rule-based approaches, but more recent work is more centered around the use of neural networks and deep learning.

In this paper, we provide a review of the different potential methodologies used for user data privacy in structured data and unstructured textual data. One of our key objectives in this research is to discover the most promising methodologies that have been proposed in literature. Therefore, we've reviewed the key existing solutions and conducted a deep and wide comparative study. Also, we reviewed the most prominent tools available on the web for natural language processing, that can be/are being used for providing privacy in text. In our comparative study, we look into the privacy methodologies used for structured data, before the governance of the use of unstructured data. We also identify the major weaknesses of existing approaches for privacy in natural texts. Based on our finding we proposed a novel methodology that would address these weaknesses. In this paper, we merely presented the architecture of our work-in-progress that is aimed at providing user anonymity in text. In addition, our proposed solution is capable of providing metrics concerning the risk of privacy leakage, sensitivity, and usability of a given text document containing personal information.

The remainder of this paper is organized follows. We start by discussing privacy methodologies used in structured data in Section II. In Section III we discuss privacy methodologies and tools used for textual data, and talk about their advantages and shortcomings. Then we introduce our approach for privacy in text in Section IV, and finally conclude and debate future work in Section V.

II. Privacy in Structured Data

There are two natural models for privacy mechanisms: interactive and noninteractive. In the noninteractive setting the data collector, a trusted entity, publishes a sanitized version of the collected data; the literature uses terms such as anonymization and de-identification. Traditionally, sanitization employs techniques such as data perturbation and sub-sampling, as well as removing well-known identifiers such as names, birth dates, and social security numbers. It may also include releasing various types of synopses and statistics. In the interactive setting, the data collector, again trusted, provides an interface through which users may pose queries about the data, and get (possibly noisy) answers.

Originally, data were published in tabular format, and made anonymous by simply removing all the explicit identifiers like names and phone numbers. However, in most of these cases, the remaining data can be used to re-identify individuals by linking it to other purposely collected data or by looking at unique characteristics in the released data [28] [29] [30]. Combinations of few characteristics often combine in populations to uniquely or nearly uniquely identify some individuals. Most known study on this is one done by Archie et al. [29] in the university of Texas where they applied their own de-anonymization methodology to a dataset published by Netflix (Netflix Prize dataset) [25], which contained anonymous movie ratings of 500,000 subscribers of Netflix, and demonstrated that an adversary who knows only little information about an individual subscriber can easily identify this subscriber's record in the dataset. A more recent work by Narayanan et al. [30] shows a similar context, only this time de-anonymizing the Netflix Prize dataset users using publicly available Amazon review data [26] [27]. Here, [30] were able to uncover more user information like a user's full name and shopping habits.

A. Noninteractive Approach

1) K-Anonymity: k-anonymity [1] is a property of a dataset that describes its level of anonymity. Developed in 1998 as a means to address the problem of releasing person-specific data while preserving the anonymity of the individuals to whom the data refers using generalization and suppression techniques. A dataset is k-anonymous if every combination of identity-revealing characteristics (quazi-identifiers) occurs in at least k different rows of the dataset. Table I shows a dataset that has been 2-anonymized; note how the attributes ”Age” and ”Gender” are identical in the top 2 and bottom2 rows.

2) /!-Diversity: /!-diversity [3] was developed in 2006 to solve 2 privacy problems found in k-anonymity. First one is that an attacker can discover the values of sensitive attributes in a k-anonymous dataset when there is little diversity in those sensitive attributes. Second is background knowledge attacks. To give an example, if there are 100 different men with ages above 70 years living in area *A* who all have allergies to peanuts, then I know that Bob, who is 72 years of age, living

TABLE I
2-ANONYMOUS DATASET

Age	Gender	Score
[10-12]	Male	98
[10-12]	Male	77
[11-12]	Female	97
[11-12]	Female	80

TABLE II
3-DIVERSE DATASET

Nonsensitive			Sensitive
Zip Code	Age	Nationality	Condition
1305*	≤ 40		Heart Disease
1305*	≤ 40		Viral Infection
1305*	≤ 40		Cancer Cancer
1305*	≤ 40		
1485*	> 40		Cancer
1485*	> 40		Heart Disease
1485*	> 40		Viral Infection
1485*	> 40		Viral Infection

in area *A*, also has an allergy to peanuts. /!-diversity aims to solve these problems by applying the following principle: a generalized quasi-identifier q*-block (equivalence class) is /!-diverse if it contains a minimum of '/!' properly depicted values under the sensitive attribute present in these blocks. If every q*-block in a dataset is /!-diverse, then the dataset meets the /!-diversity concept. Table II shows an example of an /!-diverse (3-diverse) dataset.

3) t-Closeness: t-closeness [4] comes as a betterment of /!-diversity by decreasing the granularity of the interpreted data. Introduced in 2007, where Li et al. [4] showed that /!-diversity is neither necessary nor sufficient to prevent attribute disclosure, and instead provided *t*-closeness which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of a sensitive attribute in the overall table. The distance between distributions is measured using Earth Movers Distance (EMD). For a categorical attribute, EMD is used to measure the distance between the values in it according to the minimum level of generalization of these values in the domain hierarchy. Table III shows an example of a dataset that has 0.167-closeness with respect to Salary and 0.278-closeness with respect to Disease.

These methods are not applicable for providing privacy for textual data. They were made at a time were structured data was the governing method for data storage.

B. Differential Privacy (Interactive Approach)

Differential privacy was introduced in 2006 by Dwork and Aaron [6]. It offers a robust mathematical definition of privacy, and was developed as a solution for privacy-preserving data analysis. It ensures that the result of an algorithm is not overly dependent on any instance, and states that there should be a strong probability of producing the same output even if an instance was added to or removed from the dataset. Differential privacy leapt from research papers to tech news headlines when, in the 2016 WWDC keynote, Apple VP of Engineering

TABLE III
DATASET WITH 0.167-CLOSENESS WITH RESPECT TO SALARY AND 0.278-CLOSENESS WITH RESPECT TO DISEASE

ZIP Code	Age	Salary	Disease
4767*	*leq* 40	3K	gastric ulcer
4767*	*leq* 40	5K	stomach cancer
4767*	*leq* 40	9K	pneumonia
4790*	*leq* 40	6K	gastritis
4790*	*leq* 40	11K	flu
4790*	*leq* 40	8K	bronchitis

Craig Federighi announced Apple's use of the concept to protect user privacy in iOS [39]. According to linknovate.com, tech corporations are researching heavily into differential privacy with Microsoft, Google and Apple being the top entities worldwide leading the innovations and advancements as of the date of publishing this work [41]. Google developed new algorithmic techniques for deep learning and a refined analysis of privacy costs within the framework of differential privacy to solve the problem of models exposing private information [40]. Google also announced in September 5, 2019 that it is open-sourcing an internal tool the company uses to securely draw insights from datasets that contain the private and sensitive personal information of its users, called differential privacy library [42].

Although differential privacy is praised for being an interactive solution that can be adapted to different scenarios (data collection, data analysis, machine learning...), it is not without its flaws. Kifer and Machanavajjhala [43] provide a no-free-lunch theorem to show that it is necessary to make assumptions about how the data is generated, to provide privacy, which is unlike what differential privacy claims. There is also the open problem of setting the optimum value of the algorithm's parameters based on the scenario at hand, like the parameter "Epsilon" ($\mathcal{E}$). In addition, the main criticism against differential privacy is the fact that it produces noisy results, decreasing the accuracy of the output. This means that in order to get decent results from a query, one needs to have a reasonably large dataset so that the added noise doesn't interfere much with the accuracy of the results.

III. PRIVACY IN TEXTUAL DATA

Unstructured data have internal structure but is not structured via pre-defined data models or schema. It may be textual or nontextual, and human or machine-generated. It doesnt fit neatly into the traditional row and column structure of relational databases. Examples of unstructured data include: emails, videos, audio files, web pages, and social media messages. According to mongoDB, in today's world of Big Data [7], most of the data that is created is unstructured with some estimates of it being more than 95% of all data generated [21].

Our work focuses on privacy in textual data. There has been lots of work on applying privacy to text, mostly in the form of de-identification. Challenges like the n2c2 2006: De-identification and Smoking Challenge [8] and the n2c2 2014:

De-identification and Heart Disease Risk Factors Challenge [9] (previously housed at i2b2) motivated research in textual data de-identification, namely in the field of healthcare. This influenced most work being done on textual data privacy to primarily target medical documents, due to the relative ease of access to pre-labeled training data; these challenges provided pre-labeled data from the domain to facilitate any training/testing required by the algorithms in development.

A. Medical Field

In many countries including the United States, medical professionals are strongly encouraged to adopt electronic health records (EHRs) and may face financial penalties if they fail to do so [34] [35]. One of the key components of EHRs is patient notes. However, before patient notes can be shared with medical investigators, some types of information, referred to as protected health information (PHI), must be removed in order to preserve patient confidentiality. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) [36] defines 18 different types of PHI:

1) Names
2) Dates, except year
3) Telephone numbers
4) Geographic data
5) FAX numbers
6) Social Security numbers
7) Email addresses
8) Medical record numbers
9) Account numbers
10) Health plan beneficiary numbers
11) Certificate/license numbers
12) Vehicle identifiers and serial numbers including license plates
13) Web URLs
14) Device identifiers and serial numbers
15) Internet protocol addresses
16) Full face photos and comparable images
17) Biometric identifiers (i.e. retinal scan, fingerprints)
18) Any unique identifying number or code

Performing such a task manually proved to be time-consuming and quite expensive. Douglass et al. [37] [38] reported that annotators were paid 50 US dollars per hour and read 20 000 words per hour at best. This motivated research in this domain to automate this process.

Earlier research in the field were oriented towards **rule-based** or **pattern-matching** solutions, using either complex regular expressions, dictionaries or a combination of both [10] [11] [12] [13] [14]. The advantages of rule-based and pattern matching de-identification methods is that they require little or no annotated training data, and can be easily and quickly modified to improve performance by adding rules, dictionary terms, or regular expressions. Disadvantages of pattern matching de-identification methods are that developers have to craft many complex algorithms in order to account for different categories of PHI, and the required customization to fit a particular dataset. As such, PHI pattern recognition performance may not be generalizable to different datasets (i.e. data from a different institution or a different type of medical report). Another disadvantage is the need for developers to be aware of all possible PHI patterns that can occur, such as location patterns that use nonstandard abbreviations (e.g., 'Cal' for California).

Later work tended to be mostly based on machine learning methods to classify words as PHI or not PHI, and in different classes of PHI in the former case. The methods used a range of techniques from Support Vector Machines, to Conditional Random Fields, Decision Trees, and Maximum Entropy [15] [16] [17] [18]. More recent work is more focused on utilizing neural networks and deep learning in its approach to de-identify patient data. Ji Young Lee et al. [19] incorporate human-engineered features as well as features derived from electronic health records to a neural-network-based de-identification system composed of a Long Short Term Memory neural network [20].

B. Differential Privacy with Textual Data

Benjamin Weggenmann et al. provide an automated text anonymization approach that applies differential privacy to the vector space model [54]. They obscure term frequencies in textual documents' TF-IDF vectors in a differentially private manner. Their aim is to prevent a document's author attribution through the evaluation of the document's TF-IDF vectors using different data-mining techniques. They also demonstrate that this approach has a low impact on accuracy when mining these document vectors. Our goal is different from that of Weggenmann in that we aim to provide privacy methods to the actual text documents, and not their vector representations.

C. Review of Available De-Identification Tools

1) MITdeid: MITdeid [23] is an automated de-identification software package that is generally usable on most medical records aimed at removing HIPAA PHI, and an extended PHI set that includes doctors' names and years of dates. The software achieves that by utilizing lexical look-up tables, regular expressions, and simple heuristics. This tool has a precision of 93.2%, recall of 99.8%, and an F1-score of 96.4%.

2) De-identification of Patient Notes with Recurrent Neural Networks (2016) - DeID [31]: The solution presented in this paper uses RNNs (Recurrent Neural Networks) with LSTM (Long Short-Term Memory) to de-identify medical text documents. The system is composed of 3 layers:

1) Character-enhanced token embedding layer.
2) Label prediction layer
3) Label sequence optimization layer

The solution was evaluated on two datasets: n2b2 2014 [9] and MIMIC de-identification datasets (assembled by the writers of the system and is twice as large as the n2b2 2014 dataset). It has a precision of 97.9%, a recall of 97.8%, and an F1-Score of 97.8%.

Algorithm	Target Data Structure	Approach	Scores %		
			Precision	Recall	F1-Score
K-anonymity	Structured	generalization/suppression of quazi-identifiers	-	-	-
R-diversity	Structured	adds on k-anonymity by making the sensitive attributes in every equivalence class of the dataset records contain a minimum of R properly depicted values	-	-	-
t-closeness	Structured	adds on R-diversity by decreasing the granularity of the interpreted data	-	-	-
Differential Privacy	Structured	adding noise to ensure that the result of an algorithm is not overly dependent on any single instance	-	-	-
MITdeid	Textual	utilizes lexical look-up tables, regular expressions and simple heuristics to remove HIPAA PHI frommedical records	93.2	99.8	96.4
DeID	Textual	uses RNNs with LSTM to remove HIPAA PHI from medical records	97.9	97.8	97.8
NLM-Scrubber	Textual	uses a combination of regular expressions and pattern-matching to remove 12 personal identifier categories in patient medical notes	93.2	99.8	96.4
CliniDeID	Textual	uses twelve de-identification models that use deep learning, shallow learning,or rule-based approaches to remove HIPAA PHI	97	94.4	95.7

TABLE IV
COMPARISION OF PRIVACY ALGORITHMS

3) NLM-Scrubber: NLM-Scrubber [32] is advertised as a HIPAA compliant, clinical text de-identification tool designed and developed at the National Library of Medicine. It looks for 12 personal identifier categories in patient medical notes. It uses a combination of regular expressions and pattern-matching to locate and remove identifiers from documents.

Testing on some sample text reveals that it is not difficult to manipulate with this tool. For example, a name that isn't capitalized is not detected as a name. Also, the age (65) in a sentence like: *Dave is a **65 year old** man* is not detected as an age. This tool offers a precision of 93.2%, recall of 99.8%, and an F1-score of 96.4%.

4) CliniDeID: Is a tool for de-identifying clinical notes according to the HIPAA Safe Harbor method. Owned by the company ClinAcuity and based on the work done by Youngjun Kim et al. [33], it finds identifiers and tags or replaces them with surrogates for anonymity. The tool includes twelve de-identification models that use deep learning, shallow learning, or rule-based approaches. This tool has a precision of 97%, recall of 94.4% and an F1-score of 95.7%.

Table III-A compares all the tools and algorithms mentioned in this section so far.

5) Named Entity Recognition Tools: Due to the strong correlation between Text de-identification and Named Entity Recognition (NER), here we shall discuss the tools we found that handle text analysis and (NER).

1) Stanford CoreNLP Tool (2014) [55]: Is an NLP tool created by Stanford university, initially developed in 2006, further work led to the system being released as free open source software in 2010. The tool supports, to varying degrees, the languages Arabic, Chinese, English, French and German. In our experimentation, the tool was quite successful at part-of-speech tagging, but had varying inaccuracies when it came to NER, especially with identifying person names.
2) NeuroNER [56]: is a program that performs named-entity recognition (NER), used by the Stanford CoreNLP tool. It's composed of a 3-layer recurrent neural network with LSTM. The tool was good at detecting names, unless they were lower-cased. It was also good at detecting locations, but it doesnt detect any date or age.
3) Gate [57]: Gate offers text analysis services (part-of-speech, NER,), but it uses a static approach. Its not so good at detecting named entities, and its quite easy to trick it by changing the structure of the sentence.
4) IBMs Watson, Natural Language Understanding [58]: Is a collection of APIs that offer text analysis tools using Natural Language Processing. Watson's performance in NER was inconclusive for us; on one side, it offers the most complex analysis where it doesn't only name the entities with decent granularity (an address is divided into ”location” and ”facility”), but can also detect the tonality of the speech. On the other hand, it was still not so difficult to mislead. It doesn't always detect dates, and still doesn't detect person names if they are not capitalized.

D. Discussion

The application of **differential privacy** to textual data is mostly possible by using the word vectorization of key terms chosen from the text, then applying differential privacy to these vectors. This is useful for running privacy-preserving statistical analysis on these terms, or to prevent a document's author attribution [54]. However, it falls short when it comes to preserving the structure of the text, since it picks out only specific terms discarding the rest of the text. Our aim

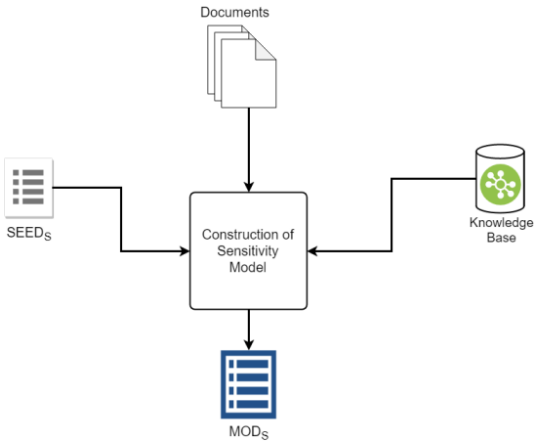

Fig. 1. MOD_S Model Construction Process for a Given Domain D.

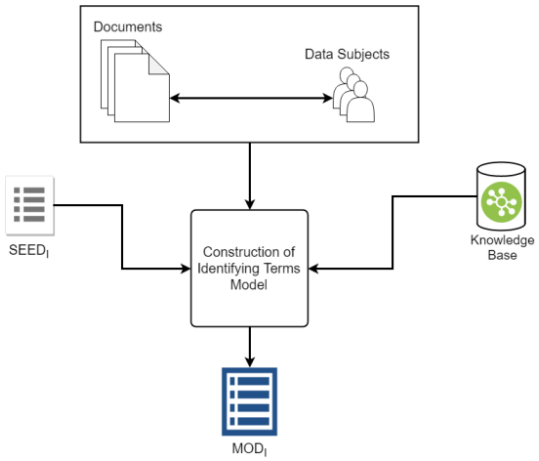

Fig. 2. MOD_I Model Construction Process for a Given Domain D.

is not to choose specific keywords from a given text to run a specific analysis, rather we want to conserve most of the text, removing/obscuring only what's necessary to preserve the privacy of any individuals present in it while preserving most of the text's utility.

Dictionary based and **pattern-matching based** approaches to provide privacy in text, like MITdeid and NLM-Scrubber, aren't much complex to implement, and require little to no annotated training data. But that comes on the expense of being static, where every target term to be captured has to be manually transcribed through complex regular expressions. In addition, solutions using these approaches can't be generalized to handle different datasets; they're only made to handle a target dataset, or a category of datasets.

Neural-network based approaches like DeID and ClniDeID are the ones with the most promise in terms of accuracy, adaptability, and generalizability. Once a neural network model is created and trained to capture specific terms in a piece of text, it is then able to capture other terms that are symantically equivalent, which are learned from the context of the text. This is crucial in the field of text analysis since it is very difficult to predict every possible structure of a sentence in a given language, or every possible use of a term or word. The problem with available solutions is that they are fixated on the field of medicine and capturing PHIs, and don't take infto account texts about other domians like finance or trading. Besides, these solutions treat all identifiers equally, although one identifier (like names) can have higher priority to be removed than others (like age) since it can identify an individual more easily.

IV. Our Approach

In our work, we are concentrating on text, therefore all operations are done on textual documents, each of which contains natural language text. We assume that a document is associated with a single natural person and, without loss of generality, we assume that multiple documents may refer to a single natural person.

In order to cause a privacy problem, the document should be identifiable (containing personal identifiers), and should contain private information. If any of these conditions are not satisfied, then the document will not cause privacy leaks. Therefore, the degree of privacy risks associated with a textual document are a combination of the present identifying and private information present in it.

Due to the nature of text documents having different priorities to what is private and what is sensitive based on the *domain* that said documents belong to (finance, healthcare...), it is important to take into account the differences of each term's ”criticality” in a given document based on what domain said document belongs to. In our work, we refer to words in documents as ”terms”, but not all words, rather ones that are not stop words.

The main objective of our work is of two folds:

1) Devise a framework to measure the degree of identifiabilty and sensitivity of entities in a given text domain. Then, using these measures, construct a module to be used to assess the privacy risks of a text document belonging to said domain, if the text was to be published, based on the sensitive information and the personal identifiers present in this document. The provided measures take into consideration the semantics of the documents as the main utility guarantee.
2) Based on the devised framework, provide a set of tools and algorithms for privacy preserving textual data management including textual data publishing and mining.

A. Term Attribution

We associate three attributes to each term in a given text document.

1) identifiability: the degree to which a term can identify a real person. For example, a person's name has a high level of identifiability, but his country of residence might have a lower level of that (there are more people living in the United Kingdom than there are people called Jason).

TABLE V
SAMPLE OF THE 3 METRICS FOR GIVEN TERMS

Model			
Terms	Identifiability	Privacy Sensitivity	Semantic Value
age	0.3	0.2	0.2
sex	0.4	0.3	0.3
disease	0.2	0.7	0.6
...	...	...	...

2) Privacy Sensitivity: is this term a sensitive piece of information? Each domain has its own set of sensitive keywords, so this is domain-specific. A term such as location could be both privacy-sensitive and identifying a real person.
3) Semantic value: how much information does this term convey with respect to the general context of the document. Semantic value should be assessed with respect to the application. For example, in case of the domain of social studies, the opinion-baring terms have more semantic value, while in a health application this may be different.

For each term in a given document, each of these attributes is marked with a numerical value (a metric) reflecting the significance of the term with respect to the attribute. Each of the three metrics can be considered as a normalized weight, where a higher value indicates that the term is more significant.

Identifyability metric. For each term, this value is calculated by checking the term's uniqueness; for example, given 2 names (Bob and Jason), if the name Bob appears more often than Jason, then Bob has a lower identifyability metric than Jason,since it is more unique.

Sensitivity metric. For each term, this value is calculated based on the work done by Sánchez et al. [45], where by using the semantics of the text, we can assess the degree of sensitiveness of terms according to the amount of information they provide. Then, this assessment is represented as a normalized numerical value.

Semantic value metric. Using sentiment analysis, the sentiment of sentences in the text is studied to determine the significance of each term to the semantic structure of their respective sentence. The terms that each sentence is centered around are considered as terms of semantic significance for the text. This significance is then represented as a normalized numerical value.

A sample table is provided in Table V.

B. Adaptability

Our solution contains data-driven models adaptable to different data domains, with the ability to customize these models further through training on vendor-specific data. This concept is based on word vectorization, where dictionaries of vendor-specific terms relating to identifiability and sensitivity are first constructed, then expanded upon using domain-specific knowledge-base to form a model, which is adapted to the contexts used in vendor use cases. The model is constructed by finding terms in the vendor data that is similar in context to the terms present in the original dictionaries. Figures 1 and 2 represent the construction process of the model for sensitivity (MOD_S) from the dictionary for sensitive terms ($SEED_S$) and the model for identifiability (MOD_I) from the dictionary for identifiable terms ($SEED_I$) respectively.

C. Restriction Flexibility

For every given data domain exists a pre-calculated *criticality threshold*. This threshold is evaluated through the analysis of domain-specific datasets, and is used as an anchor point for the data holder to fine-tune the level of anonymity that is to be applied to the text in a manner that suits their privacy standards. This is important because there is a compromise that must be met between privacy and utility; the more strict the privacy rules being applied are, the less utility the document has (since we are removing data from the document). Our solution makes it easy to strike the required balance, as privacy risks are quantified, allowing data holders to alter solution parameters to fulfill their privacy requirements.

V. CONCLUSION AND FUTURE WORK

Data storage is now mostly moving towards unstructured data, and the use of textual data has become an inseparable part of our daily lives. We only continue to share more of our personal info through online services and social media. In this paper, we've introduced a new concept for a data-oriented solution that provides measures for a given text document to assess the privacy-leak potential of said document, as well as measure its semantic utility. We believe that this work can pave the way for a new data-driven orientation in the privacy-research field.

REFERENCES

[1] P. Samarati and L. Sweeney. Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression, Technical Report (SRI-CSL-98-04), 1998

[2] L. Sweeney. k-Anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10 (7), 2002

[3] A. Machanavajjhala & J. Gehrke & D. Kifer & M. Venkitasubramaniam. (2006). l-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery From Data - TKDD. 1. 24. 10.1145/1217299.1217300.

[4] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, 2007, pp. 106-115. doi: 10.1109/ICDE.2007.367856

[5] Rajendran, Keerthana & Jayabalan, Manoj & Rana, Muhammad Ehsan. (2017). A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data. 17.

[6] C. Dwork, A. Roth, The Algorithmic Foundations of Differential Privacy, Foundations and Trends in Theoretical Computer Science Vol. 9, Nos. 34 (2014) 211407, 2014 C. Dwork and A. Roth DOI: 10.1561/0400000042

[7] Chang, F., J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes and R.E. Gruber, 2006. Bigtable: A distributed storage system for structured data. Proceeding of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI 06). Berkeley, CA, USA, pp: 15-15.

[8] n2c2 2006: Deidentification and Smoking Challenge, https://portal.dbmi.hms.harvard.edu/projects/n2c2-2006/, Accessed on 2019-10-20

[9] n2c2 2014: De-identification and Heart Disease Risk Factors Challenge, https://portal.dbmi.hms.harvard.edu/projects/n2c2-2014/, Accessed on 2019-10-21
[10] Beckwith BA. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak. 2006. p. 12.
[11] Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med. 2003. pp. 6806.
[12] Fielstein EM, Brown SH, Speroff T. Algorithmic De-identification of VA Medical Exam Text for HIPAA Privacy Compliance: Preliminary Findings. Medinfo. 2004. p. 1590.
[13] Friedlin FJ, McDonald CJ. A software tool for removing patient identifying information from clinical documents. J Am Med Inform Assoc. 2008;15(5):60110. doi: 10.1197/jamia.M2702.
[14] Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol. 2004. pp. 17686.
[15] Aramaki E. Automatic Deidentification by using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC. 2006.
[16] Guo Y. Identifying Personal Health Information Using Support Vector Machines. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC. 2006.
[17] Szarvas G, Farkas R, Kocsor A. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. 9th Int Conf Disc Sci (DS2006), LNAI. 2006. pp. 267278.
[18] Wellner B. Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc. 2007. pp. 56473
[19] Lee, Ji & Dernoncourt, Franck & Uzuner, Ozlem & Szolovits, Peter. (2016). Feature-Augmented Neural Networks for Patient Note De-identification.
[20] Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., & Woo, W. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. ArXiv, abs/1506.04214.
[21] mongoDB: Unstructured Data In Big Data, https://www.mongodb.com/scale/unstructured-data-in-big-data, Accessed on 2019-10-15.
[22] National Association of Health Data Organizations, A Guide to State-Level Ambulatory Care Data Collection Activities (Falls Church: National Association of Health Data Organizations, Oct. 1996).
[23] Kayaalp, M., Browne, A. C., Dodd, Z. A., Sagan, P., & McDonald, C. J. (2014). De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2014, 767776.
[24] Group Insurance Commission testimony before the Massachusetts Health Care Committee. See Session of the Joint Committee on Health Care, Massachusetts State Legislature, (March 19, 1997).
[25] Netflix Prize Dataset: https://www.kaggle.com/netflix-inc/netflix-prize-data. Downloaded on July 15, 2019.
[26] R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016.
[27] J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015.
[28] L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA: 2000. Forthcoming book entitled, The Identifiability of Data.
[29] Archie, M., Gershon, S., Katcoff, A., & Zeng, A. (2018). Who s Watching ? De-anonymization of Netflix Reviews using Amazon Reviews.
[30] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP '08). IEEE Computer Society, Washington, DC, USA, 111-125. DOI: https://doi.org/10.1109/SP.2008.33
[31] Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, Peter Szolovits, De-identification of patient notes with recurrent neural networks, Journal of the American Medical Informatics Association, Volume 24, Issue 3, May 2017, Pages 596606.
[32] Kayaalp M, Browne AC, Dodd ZA, Sagan P, McDonald CJ. De-identification of Address, Date, and Alphanumeric Identifiers in Narrative Clinical Reports. AMIA Annu Symp Proc. 2014 Nov 14;2014:767-76.
[33] Kim, Y., Heider, P., & Meystre, S. (2018). Ensemble-based Methods to Improve De-identification of Electronic Health Record Narratives. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2018, 663672.
[34] DesRoches CM, Worzala C, Bates S. Some hospitals are falling behind in meeting meaningful use criteria and could be vulnerable to penalties in 2015. Health Affairs. 2013;32:135560.
[35] Wright A, Henkin S, Feblowitz, et al. Early results of the meaningful use program for electronic health records. New Engl J Med. 2013;368:77980.
[36] Office for Civil Rights H. Standards for privacy of individually identifiable health information. Final rule. Federal Register. 2002;67:53181
[37] Douglass M, Clifford G, Reisner A, et al. De-identification algorithm for free-text nursing notes. Comput Cardiol. 2005:33134.
[38] Douglas M, Clifford G, Reisner A, et al. Computer-assisted deidentification of free text in the MIMIC II database. Comput Cardiol. 2004:34144.
[39] Apple Adopts Differential Privacy, https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf, accessed on 2019-10-20
[40] Abadi M, Chu A, Goodfellow I, McMahan B, Mironov I, Talwar K, Zhang L. Deep Learning with Differential Privacy, https://arxiv.org/abs/1607.00133, 23rd ACM Conference on Computer and Communications Security (ACM CCS), 2016, 308-318
[41] Top 10 entities worldwide leading the innovations and advancements in Differential Privacy, https://www.linknovate.com/search/?query=%22differential%20privacy%22%2C%22privacy%20by%20design%22&utm_source=blog.linknovate.com&utm_medium=referral&utm_campaign=data%20stories&utm_content=differential%20privacy, Accessed on 2019-10-20
[42] Enabling developers and organizations to use differential privacy, https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html, accessed on 2019-10-20
[43] Daniel Kifer and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD '11). ACM, New York, NY, USA, 193-204. DOI: https://doi.org/10.1145/1989323.1989345
[44] David Maier, Theory of Relational Databases, Computer Science Press; 1st edition (March 1983), 978-0914894421.
[45] David Sánchez , Montserrat Batet, Alexandre Viejo, Detecting Sensitive Information from Textual Documents: An Information-Theoretic Approach, Modeling Decisions for Artificial Intelligence, Springer Berlin Heidelberg, 2012, 173–184",
[46] The Facebook and Cambridge Analytica scandal, explained with a simple diagram, https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram, Accessed on 2019-10-20.
[47] Googles Sundar Pichai was grilled on privacy, data collection, and China during congressional hearing, https://www.cnbc.com/2018/12/11/google-ceo-sundar-pichai-testifies-before-congress-on-bias-privacy.html, Accessed on 2019-20-10.
[48] Abdullah, Ahmad & Zhuge, Qingfeng, (2015). From Relational Databases to NoSQL Databases: Performance Evaluation. Research Journal of Applied Sciences, Engineering and Technology. 11. 434-439. 10.19026/rjaset.11.1799.
[49] GDPR, https://eugdpr.org, Accessed on 2019-10-15
[50] Trump order bans US firms from dealing with Huawei, https://www.techradar.com/news/trump-order-bans-us-firms-from-dealing-with-huawei, Accessed on 2019-10-25.
[51] If you have a smart TV, take a closer look at your privacy settings, https://www.cnbc.com/2017/03/09/if-you-have-a-smart-tv-take-a-closer-look-at-your-privacy-settings.html, accessed on 2019-10-12.
[52] Privacy and Human Rights, http://gilc.org/privacy/survey/intro.html, Accessed on 2019-10-15.
[53] Hecht, R. and S. Jablonski, 2011. Nosql evaluation. Proceeding of International Conference on Cloud and Service Computing, pp: 336-41 [54] Weggenman, B., & Kerschbaum, F. (2018, July) SynTF: Synthetic and Differentially Private Term Frequency Vectors for Privacy-Preserving Text Mining *SIGIR18, July 8-12, 2018, Ann Arbor, MI, USA*
[55] Stanford Core NLP, https://stanfordnlp.github.io/CoreNLP, Accessed 2019-10-2
[56] Dernoncourt, Franck & Lee, Ji & Szolovits, Peter. (2017). NeuroNER: an easy-to-use program for named-entity recognition based on neural networks.
[57] Gate, https://gate.ac.uk/gate/doc/papers.html, Accessed on 2019-10-19
[58] Natural Language Understanding, https://www.ibm.com/watson/services/natural-language-understanding, Accessed on 2019-10-10