Proceedings of the Workshop on Social Threats in Online Conversations: Understanding and Management (STOC-2020), pages 21–28
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A Privacy Preserving Data Publishing Middleware for Unstructured, Textual Social Media Data

Prasadi Abeywardana, Uthayasanker Thayasivam
Department of Computer Science and Engineering
University of Moratuwa, Sri Lanka
prasadiapsara.18@cse.mrt.ac.lk, rtuthaya@cse.mrt.ac.lk

Abstract

Privacy is going to be an integral part of data science and analytics in the coming years. The next hype of data experimentation is going to be heavily dependent on privacy preserving techniques mainly as it's going to be a legal responsibility rather than a mere social responsibility. Privacy preservation becomes more challenging specially in the context of unstructured data. Social networks have become predominantly popular over the past couple of decades and they are creating a huge data lake at a high velocity. Social media profiles contain a wealth of personal and sensitive information, creating enormous opportunities for third parties to analyze them with different algorithms, draw conclusions and use in disinformation campaigns and micro targeting based dark advertising. This study provides a mitigation mechanism for disinformation campaigns that are done based on the insights extracted from personal/sensitive data analysis. Specifically, this research is aimed at building a privacy preserving data publishing middleware for unstructured social media data without compromising the true analytical value of those data. A novel way is proposed to apply traditional structured privacy preserving techniques on unstructured data. Creating a comprehensive twitter corpus annotated with privacy attributes is another objective of this research, especially because the research community is lacking one.

Keywords: Privacy Preserving Data Mining, Privacy Preserving Data Publishing, Disinformation, Micro-targeting, Anonymization, Data Utility, Social Networks, Data Science

1. Introduction

Big data being a buzz word which has created an immense hype in the society, many analytical models are employed in order to repurpose those data and derive insights. With the advancements of distributed systems and theoretically cheap storage, there are less constraints to capture data as much as possible and store them. Collection of data related to individuals in a global scale has become mainstream because of this.

Data are collected in big scales and published to be used by different parties for different purposes. At this point of publishing, there should be a proper insurance for personal data, as the publishing party cannot guarantee for which purposes this personal information will be used by the utilizing party.

Micro-targeting based on the third-party analysis done on personal data is used as a means of disinformation campaigns. A famous example for this is dark advertisements targeting specific users in a very personalized manner for sharing misinformation in political campaigns. This is achieved by identifying target users by analyzing their political preferences and showing them personalized dark ads with content they are highly likely to believe. Analyzing sensitive personal information and using them for various intentions without user consents makes it a combination of an ethical and legal concern (Alaphilippe et al., 2019).

1.1 Data Protection Regulations

Until recently, privacy was just a social responsibility, but it's no more like that, because many legal systems have begun to enforce laws on protecting individuals' privacy. Specially incidents like what happened between Facebook and Cambridge Analytica have forced the governments and policy makers to look at personal information protection as an emerging concern. Following are some of such novel legal requirements which arouse recently.

1.1.1 General Data Protection Regulation (GDPR)

This is a regulation imposed by European Union (EU) on data protection and privacy for all individuals within the EU and the European Economic Area (EEA) (Wikipedia, 2016). This is applicable to exporting and processing personal data in a region outside EU as well. The intention of this regulation is to make it easy for non-European companies to work with European bodies without any data breaches.

1.1.2 Russian Federal Law on Personal Data

This is a regulation which emphasizes on systemizing the data processing of individuals in Russia. This emphasizes on localizing personal data of Russian citizens to Russia (KPMG, 2018).

1.1.3 German Bundesdatenschutzgesetz (BDSG)

This governs the exposure of personal data, which are manually processed or stored in IT systems. This was being modified with certain amendments for a long period of time and has become stricter in the recent past.

1.2 Social Threats of Personal Data Analysis

Personal data are coming into analytical systems through various domains. Mobile data, health care data, social media data and web usage data are a few such domains which can pump a huge amount of personal data into analytical systems without the knowledge or consent of individuals. There's one prominent area, which has reformed the sharing of personal information, that is none other than social media. People choose to share many information about themselves as well as their close ones, compromising the privacy of both parties (Mehta and Rao, 2015).

Social platforms offer their data to third parties and advertisers to use in their analysis and campaigns. But sometimes these data are used in micro targeted disinformation campaigns to share dark ads. These highly personalized adverts are heavily used in political contexts to influence voters by sharing misinformation. In order to host micro targeted ad campaigns, a lot of information related to individuals, their preferences and personality are required, and social media undoubtedly contain a fortune of such data. In the recent incident that involved Facebook, Cambridge Analytica and Global Science Research (GSR), millions of US Facebook users' data were analyzed without their consent and used in voter targeting, which is unethical as it sounds (Alaphilippe et al., 2019). A solution to these concerns might be a law enforced privacy preserving middleware that has to be adopted by any social media platform, before publishing their data to a third party.

The purpose of this research is to come up with a framework to sanitize data and preserve privacy, which can be utilized before publishing textual social media data to any analytical 3rd party. This will ensure that any sensitive personal data will not be used in a way where a person's identity is revealed, and the individuals will not be subjected to disinformation campaigns. Specifically, this research addresses the problem of sanitizing social media data, which becomes more challenging due to their unstructured nature. Twitter is used as the selected social media platform to train and evaluate the capabilities of this framework. A corpus of 3000 tweets is built and annotated to be used in the model training process.

The rest of the paper is organized as follows. Some theoretical concepts related to privacy preserving data publishing particularly in the context of unstructured data will be discussed in the background section. Then the methodology adapted will be described followed by a section dedicated towards the dataset. Next section is about the experimental design and the results and after that a section is contributed for discussion and future work. Finally, the paper is concluded with a conclusion section.

2. Background

Publishing sensitive data related to individuals in a way that protects their privacy was a topic of interest for some time and many techniques are implemented with the contribution from various fields such as computer science, statistics and social science. A few theorical concepts from the PPDP domain are described under this section.

2.1 Different Types of Attributes Related to Personal Data

Attributes related to personal data can be classified as follows based on how they can identify an individual. These attributes are extracted and used in PPDP techniques (Mehta and Rao, 2015).

2.1.1 Personal Information Identifiers

These are the attributes such as ID, name or email address that can be directly used to identify an individual. These attributes uniquely recognize individuals from others.

2.1.2 Quasi Identifiers

These are the attributes that can be combined with other external data and used to identify an individual. For instance, age, gender, profession, race, religion can be considered as quasi identifiers. These are not unique identifiers by themselves but can be combined with another set of quasi identifiers to uniquely recognize a person.

2.1.3 Sensitive Attributes

These are the attributes that individuals do not want to reveal about themselves. Examples can be salary, relationship statuses and diseases.

2.1.4 Non-Sensitive Attributes

These are the attributes other than the above mentioned 3 types. They may not have a direct or indirect relationship to identify individuals.

Any PPDP process should include a mechanism to identify these attributes related to personal data before applying any sanitization technique. Based on the nature of the attribute, different sanitization techniques must be applied.

2.2 Existing Data Sanitization Techniques

Many research works have been carried out to come up with various sanitization techniques to protect personal data (Mehta and Rao, 2015; Fung et al., 2010)

2.2.1 Suppression

This mechanism replaces some attribute values by a symbol like '*' to indicate those attributes are repressed. For instance, a credit card number can be suppressed as 34** **** ****.

2.2.2 Generalization

This implies replacing an attribute with a generalized value of its class, for instance male and female values of the gender attribute or a nationality attribute can be replaced with 'Any' which is a more general value. Generalization makes sure that a combined set of quasi identifiers cannot be used to uniquely identify a person after generalizing.

2.2.3 Swapping

As the name implies this includes swapping some attribute values. For example, swapping the gender values of two records.

2.2.4 Anatomization

This involves separating quasi identifiers and sensitive attributes into different tables so that the relationship among them will be broken.

2.2.5 Permutation

This is about creating groups or buckets based on quasi identifiers and then shuffle the values of their respective sensitive attributes in each group to break the relationship between quasi identifier and the sensitive attributes.

2.2.6 Perturbation

This is about replacing the original values of some sensitive attributes using some fake values.

Table 1 shows some health records which contain different types of attributes mentioned above. Name can be considered as a direct identifier where age, gender, zip code and nationality can be considered as quasi identifiers. These direct identifiers and quasi identifiers can be used to recognize diseases different individuals have without their consent and diseases can be something these individuals don't want to reveal.

	Age	Gender	Zip Code	Nationality	Disease
John	28	M	13053	Russian	Heart Disease
Jack	29	M	13055	Chinese	Heart Disease
Bruce	22	M	13061	Japanese	Heart Disease
Ann	24	F	14332	Russian	Heart Disease
Lewis	41	M	14556	American	Cancer
Richard	45	M	13227	American	Cancer
Anders	50	M	13226	American	Cancer
Paul	37	M	13221	American	Flu
Janet	34	F	13229	American	Flu
Cary	56	M	13225	American	Flu

Table 1: Health records

Table 2 shows the application of different sanitization techniques to identifiers so that it is difficult to distinguish individuals from each other. For age, gender and nationality columns, generalization is applied whereas for the zip code column, suppression is applied.

	Age	Gender	Zip Code	Nationality	Disease
******	20-29	Any	130**	Any	Heart Disease
******	20-29	Any	130**	Any	Heart Disease
******	20-29	Any	130**	Any	Heart Disease
******	20-29	Any	14***	Any	Heart Disease
******	40-59	Any	14***	American	Cancer
******	40-59	Any	1322*	American	Cancer
******	40-59	Any	1322*	American	Cancer
******	30-39	Any	1322*	American	Flu
******	30-39	Any	1322*	American	Flu
******	40-59	Any	1322*	American	Flu

Table 2: Sanitized health records

2.3 Existing Privacy Models

As privacy is a very subjective concept there should be some baseline models to measure it against. Research community has come up with such benchmarks over time.

2.3.1 K-anonymity

A set of data is said to have k-anonymity property if the information for each individual cannot be eminently differentiated from at least k - 1 other individuals who are in the same dataset (Samarati and Sweeny, 1998).

2.3.2 L-diversity

This is an extension to the k-anonymity model, which diminishes the granularity of data using mechanisms including generalization and suppression. This tries to overcome a couple of weak points of the k-anonymity model (Machanavajjhala et al., 2006). If the variability of sensitive attributes is little, then it is possible to recognize individuals with some background knowledge, even though the data is k anonymized. L-diversity tries to solve this by setting a rule on distinct number of sensitive values an equivalence class (the set of records with similar quasi identifier values after anonymizing) can have.

2.3.3 T-closeness

This is an enhancement to l-diversity model to overcome its flows. Further reduction using this causes some loss of usefulness of the data as it tries to distort data (Li et al., 2007). This tries to find solutions for the problems of semantic closeness and skewness of data, that are not addressed by l-diversity model.

Any system which is intended to adopt a privacy preserving process should adhere to a couple of steps.

- Extract personal privacy related attributes from the data
- Sanitize those extracted attributes using a sanitization mechanism that suits the nature of the attribute
- Evaluate the level of privacy using privacy measures
- Evaluate the level of utility or usefulness using utility measures

But this process becomes very challenging if the data is unstructured, due to a couple of reasons.

2.4 Challenges with Textual, Unstructured Social Media Data

Social media has become an essential part of people’s life. There are many prevailing social media platforms that tend to connect individuals forming complex networks. And the number of users who actively participate in these platforms are drastically increasing over time pumping a huge amount of data in a high velocity. This obviously creates challenges for data scientists.

People are not reluctant anymore to share their personal information on the world wide web. Even though they don’t consider the privacy aspects a lot at the point of sharing, no one will prefer any sensitive information about their privacy being compromised.

There is various analysis that can be done on top of social media data to derive many interesting patterns. Facebook status analysis and Twitter’s tweet analysis are two such analysis that involve unstructured data. Obviously, these data involve so many sensitive facts about individuals. Unstructured nature of these data makes the privacy preservation more difficult. For example, think about the following sentence.

“My teacher who lived in Corktown died of cancer yesterday at age 65”

Even though this sentence does not contain any names or direct identifiers of an individual, the details provided there such as occupation, city and age can be used to disclose the individual. So, the things shared on social media can reveal many personal information indirectly. Ensuring this kind of data does not reveal any personal information has some inherent challenges.

- Extracting personal information related attributes from unstructured data is not straight forward
- Sanitization techniques cannot be directly applied on unstructured data
- As social media is a huge platform of information for analysis, any privacy preservation technique should not corrupt its original value, so that data will be useless
- As social media falls into big data category, any PPDP framework should cater to the challenges like variety, volume and velocity

3. Related Work

Privacy preserving data publishing is being a topic of interest in the research community for a long time now. But the advancements in digitization and computing introduces new challenges in the area of privacy preserving data publishing too. This section describes a couple of related work in the context of PPDP and unstructured data.

Fung et al. have done a comprehensive survey on the topical developments of privacy preserving data publishing techniques. They have discussed about the current status of privacy preservation and highlighted the fact that it's getting more and more attention over time. They have thoroughly discussed about anonymization techniques such as generalization and suppression, anatomization and permutation, and perturbation etc. Additionally, they have highlighted mechanisms to preserve privacy in a way that the data will remain practically useful. They talk about various information metrics that can be used to measure data usefulness such as special purpose metrics, general purpose metrics and trade-off metrics. A couple of existing anonymization algorithms are brought forward in this research, and they are classified into a set of subsets, based on the underlying methodology – whether it is based on record linkage, table linkage or attribute linkage (Fung et al., 2010).

Ramya et al. present an attempt to do privacy preserving data publishing on unstructured data with a somewhat different approach. They too have understood the fact that it is challenging to apply traditional PPDP techniques when the dataset it semi/unstructured. They have followed a document classification approach to categorize documents to indicate whether a document contains sensitive information or not. Before doing the actual classification, documents are preprocessed to remove any stop words and do the stemming. They have used a boolean label called Sensitivity Disclosure Label (SDL) to indicate a document contains sensitive information or not. Two different classifiers are employed to do the document classification. They are Multinomial Naive Bayes and K-Nearest Neighbor classifiers. As the dataset for model building and verification, they have used i2b2 (Informatics for Integrating Biology & the Bedside) medical dataset. In this approach they are only concerned about the domain level document classification, but not about a detailed tagging where the content inside the documents can be sanitized (Ramya et al., 2019).

Gardner et al. have described in their paper, an approach to de-identify unstructured medical data. They try to fill in some gaps in the privacy preservation techniques of current medical data domain. The scholars argue that current methodologies mainly consider simple anonymization techniques without taking the full advantage of the already done research work. So, they come up with an integrated framework, which embeds many powerful privacy preservation mechanisms. They employ a Bayesian classifier with a sampling-based technique and a conditional random field-based classifier to extract sensitive information from medical data. And, a k-anonymity based model is used for de-identifying information at the same time maintaining maximum data usefulness. As further work, they mention that we can explore into a mechanism where we can prioritize attributes based on their relatedness to the privacy. And extracting indirect identifiers like quasi identifiers are not focused under this research (Gardner and Xiong, 2009).

Thavavel et al. come up with another framework which talks about privacy preservation in a distributed environment with unstructured data. The proposed approach is about converting unstructured data to structured data before applying any privacy preservation mechanisms. They have converted the unstructured data to XML and then mapped that XML to node representation and the outcome is structured data. A distributed mechanism which vertically partitions the heterogeneous data are proposed under this mechanism. Data volume becomes a constraint here again, as it's not practical to convert a large amount of unstructured data to structured data (Thavavel and Sivakumar, 2012).

Liu et al. propose a privacy preserving middleware called LinkMirage which controls privacy preservation of social relationships. They claim that their novel algorithm de-identify the social relationship graph and at the same time it does not distort graph utility or usefulness. They have done an analysis using a huge real-world Google+ dataset which contained 940 million links. And they claim that the proposed algorithm guarantee 10x privacy preservation compared to the existing research work. This algorithm mainly depends on perturbation mechanisms (Liu and Mittal, 2016).

4. Methodology

Figure 1 summarizes the overall process adopted in the proposed methodology. The suggested approach mainly consists of a Twitter data publisher, a privacy preserving middleware and privacy and utility evaluator. The purpose of the implementation was to come up with an end to end system which can realize the concept of privacy preserving data publishing for unstructured and textual social media data. Each of these modules will be discussed in this section.

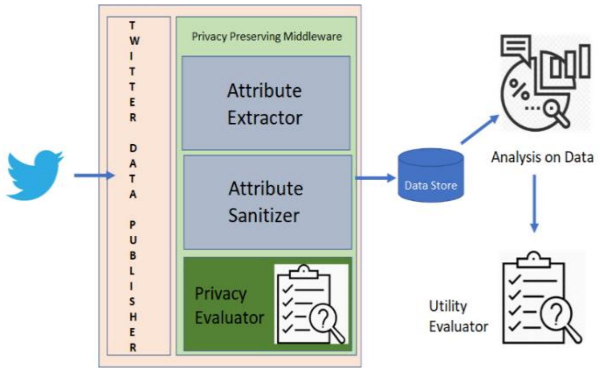


Figure 1: Overall system architecture.

4.1 Twitter Data Publisher

A data publisher was implemented as the main source of test data generating for the solution. This is a Python program which can perform a keyword search via the Twitter API to extract some tweets, or through which the users can push a precompiled set of tweets into the system. This will be the entry point in the developed prototype.

4.2 Privacy Attribute Extractor

This is the most critical module in the prototype. A decision tree-based tagging model was developed to tag tweets with attributes related to personal information. Figure 2 shows the methodology adopted in training the tagger. The manually tagged corpus was transformed before it was fed to the decision tree classifier.

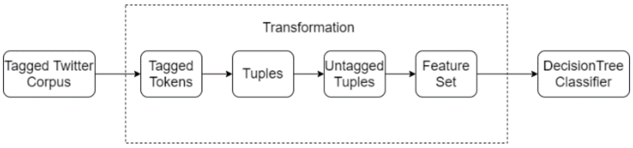


Figure 2: Data transformation.

A data set of 3000 tweets were manually annotated by 3 annotators for different identifiers related to privacy to build a corpus. The annotation scheme is described in table 3.

Attribute Type	Tag	Attribute
Direct Identifiers	DI	Name, TwitterId
Quasi Identifiers	QIAGE	Age
Quasi Identifiers	QIRACE	Race
Quasi Identifiers	QIREGION	Region
Quasi Identifiers	QIGENDER	Gender
Quasi Identifiers	QILANG	Language
Quasi Identifiers	QIJOB	Occupation
Sensitive Attribute	SA	Health Conditions, Relationship Status, Salary, Political Preferences
Non-Sensitive Attribute	NONE	Anything that does not belong to above

Table 3: Annotation scheme

This corpus was preprocessed and transformed before being fed into the decision tree classifier for training. The transformation utilities contained methods for tokenizing, untagging and extracting features from the words.

A set of syntactic, orthographic, gazetteer and affix features were used in the transformation process. Table 4 shows all the features extracted in the process.

Feature	Feature Type
is_first	Orthographic
is_last	Orthographic
is_capitalized	Orthographic
is_all_caps	Orthographic
is_all_lower	Orthographic
prefix-1	Affix
prefix-2	Affix
prefix-3	Affix
suffix-1	Affix
suffix-2	Affix
suffix-3	Affix
prev_word	Orthographic
next_word	Orthographic
has_hyphen	Orthographic
is_numeric	Orthographic
pos_tag	Syntactic
named_entity	Gazetteer

Table 4: Features selected

Insights for features were obtained by a named entity extractor that was built using AdaBoost (Carreras and Marques, 2003). As gazetteer features, values suggested by spaCy's named entity recognition are used (spaCy.io, 2016).

Then this transformed dataset was input into the classifier and trained. Train and test data were split based on the 70:30 rule and the confusion matrix was computed to score the model. This brings a macro average of 0.74 for the F1 score. Accuracy stays at 0.92 as the result is highly impacted by the 'none' label proportion.

	Precision	Recall	F1-Score
DI	0.83	0.7	0.76
QIAGE	0.69	0.77	0.73
QIGENDER	0.57	0.65	0.61
QIRACE	0.74	0.49	0.6
QIREGION	0.78	0.82	0.8
QIJOB	0.59	0.7	0.64
SA	0.73	0.81	0.77
NONE	0.98	0.98	0.98

Table 5: Classifier confusion metrics

Following is a sample tweet automatically tagged through the tagging model.

('My', 'None')('teacher', 'QIJOB')('who', 'None')('lived', 'None')('in', 'None')('USA', 'QIREGION')('died', 'None')('of', 'None')('cancer', 'SA')('at', 'None')('age', 'None')('65', 'QIAGE')

All the new tweets published through the data publisher will go through this module and get automatically tagged with appropriate tags.

4.3 Privacy Attribute Sanitizer

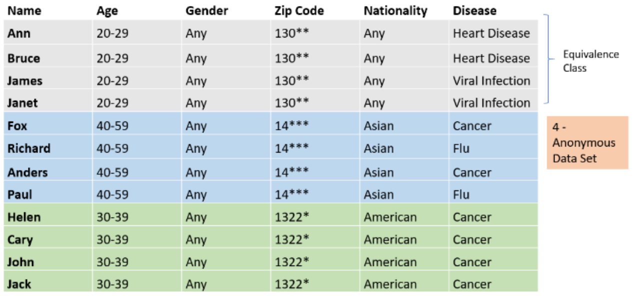
The next module in the privacy preserving pipeline is attribute sanitizer. This incorporates some sanitization techniques from the literature that suits the nature of each identifier. Sanitization Techniques applied to each attribute is presented in table 6.

Attribute	Sanitization Technique	Original Value	Sanitized Value
Name, TwitterId	Complete Anonymization	John	******
Age	Generalization (to a number range)	65	60-70
Race	Union	Indian	American, Indian, African
Region	Union	Sri Lanka	India, Sri Lanka, America, Germany, Canada
Gender	Generalization	Female	Any
Language	Generalization	English	Any
Occupation	Swapping	Teacher	Doctor

Table 6: Sanitization scheme

After applying the sanitization techniques on the tagged sentence, the original tweet is rebuilt with the anonymized values. Following shows how the above tagged tweet looks after applying anonymization techniques.

“My doctor who lived in India, Sri Lanka, USA, Canada died of Cancer at the age of 60-70.”



Name	Age	Gender	Zip Code	Nationality	Disease
Ann	20-29	Any	130**	Any	Heart Disease
Bruce	20-29	Any	130**	Any	Heart Disease
James	20-29	Any	130**	Any	Viral Infection
Janet	20-29	Any	130**	Any	Viral Infection
Fox	40-59	Any	14***	Asian	Cancer
Richard	40-59	Any	14***	Asian	Flu
Anders	40-59	Any	14***	Asian	Cancer
Paul	40-59	Any	14***	Asian	Flu
Helen	30-39	Any	1322*	American	Cancer
Cary	30-39	Any	1322*	American	Cancer
John	30-39	Any	1322*	American	Cancer
Jack	30-39	Any	1322*	American	Cancer

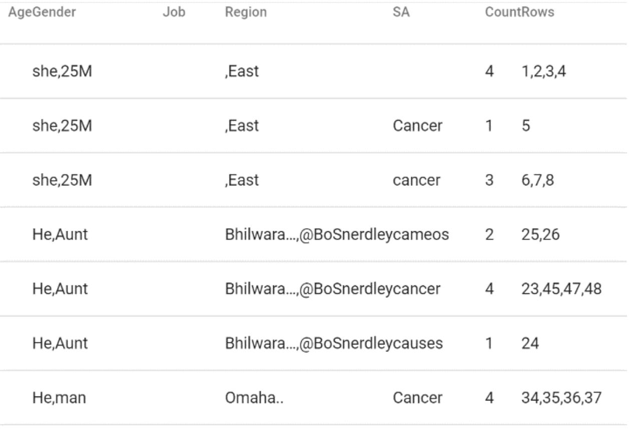
Figure 3: A 4-anonymized data set.

The developed prototype provides the ability to do the anonymization in two ways.

1. Simple anonymizing: Under this category all the quasi identifiers and direct identifiers will be anonymized without considering the fact to which extent they contribute to revealing the privacy.
2. K-anonymizing: Under this category, anonymization will be performed according to the k-anonymity model where the user can specify the k value. According to k-anonymity model, dataset is divided into equivalence classes based on similar quasi identifiers and the objective is to anonymize data in a way, a record can’t be distinguished from other records in its equivalence class. Figure 3 shows an example of a 4-anonymized data set where each record is not distinguishable from 4-1 other records. Increasing the value of k strengthens the privacy. But it can be challenging to find the correct k value which can preserve the privacy at the same time protects the utility.

The above anonymization mechanisms can be applied on either a single tweet or a set of tweets. If it is a single tweet, simple anonymization will be applied and if it is a set of tweets, user can select between simple anonymization and k-anonymization.

Textual dataset is converted into the form of structured data to perform k-anonymization and after doing the anonymization, the textual dataset is rebuilt using the structured dataset. Figure 4 shows how tweets look when they are converted to the structured format.



AgeGender	Job	Region	SA	CountRows	
she,25M		,East		4	1,2,3,4
she,25M		,East	Cancer	1	5
she,25M		,East	cancer	3	6,7,8
He,Aunt		Bhilwara...,@BoSnerdleycameos		2	25,26
He,Aunt		Bhilwara...,@BoSnerdleycancer		4	23,45,47,48
He,Aunt		Bhilwara...,@BoSnerdleycauses		1	24
He,man		Omaha..	Cancer	4	34,35,36,37

Figure 4: Textual data that are converted to structured format and k-anonymized

4.4 Utility Evaluator

A couple of metrics are provided to evaluate the quality or the utility of the privacy preserved dataset. These measures specifically target the quality of the quasi identifier groups.

4.4.1 Discernibility Metric (DM)

This assigns a penalty to each tuple based on how many other tuples in the database are indistinguishable from it (Fung et al, 2010). If a record belongs to qid group of size n, then the penalty for the record will be n and the penalty for the group will be $n^{2.}$ Whenever an anonymization task is performed, user is given the ability to calculate the discernibility metrics for each quasi identifier. The specialty with discernibility metric is it can compare the cost of generalizing for each qid value. Higher the discernibility value, higher the cost of generalization is.

4.4.2 Loss Metric (LM)

This calculates the normalized loss of each attribute of every tuple. This, in particular targets the information loss caused by the generalization. LM is defined as the number of nodes a record's value has been made indistinguishable from (via generalization) compared to the total number of original leaf nodes in the taxonomy tree (Fung et al, 2010). Loss metric is created as n-1/m where n is the number of descendants of a parent value in a generalization tree and m is the total number of domain values of an attribute.

4.4.3 Generalization Counting

This counts how many generalization/suppression operations were performed.

5. Dataset

The dataset developed for tagging the tweets is one of the biggest achievements of this research. The research community was lacking a dataset which has annotated textual data for privacy related attributes. One of the greatest intentions of this research was to come up with an annotated corpus including tweets, which can be used for future privacy preserving tasks and that goal was successfully achieved.

Tweets to build the corpus was selectively picked from a public Kaggle dataset based on a keyword search (Kaggle, 2018). This dataset contains 3000 tweets which are annotated adhering to the scheme shown in Table 3 using 3 annotators. These attributes are subjective; therefore, the tweets were cross annotated by each annotator and an agreement study was performed. The average kappa values lie between 0.6-0.7 proving our dataset is reliable.

6. Experimental Results

A keyword search was performed on Twitter using Twitter's public API to create an experimental data set. A couple of sensitive attribute values like 'cancer', 'lesbian' and 'gay' were used as keywords and a dataset of 1000 tweets were created. Then both simple anonymization and k-anonymization were performed on this tweet set and utility metrics were computed. The objective of this experiment was to simulate a real-life data anonymization operation. K value used was 4.

First the no. of sanitizations was counted, and a percentage of sanitized terms were measured.

Total number of terms sanitized: 334 (simple anonymization)
Percentage of terms sanitized: 7.7% (simple anonymization)
Total number of terms sanitized: 303 (k-anonymization)
Percentage of terms sanitized: 7.1% (k-anonymization)

Table 7: Sanitization counts

Table 8 summarizes the discernibility metric values for each quasi identifier type.

Quasi Identifier Group	Anonymized Value	DM	Privacy Type
QIGENDER	Any	9025	Simple Anonymization
QIREGION	Any	6084	Simple Anonymization
QIGENDER	She, Men, Women	4900	k-Anonymization
	Girl, Woman	64	
	Girlgriend, Girl	49	
	Girl	64	
QIREGION	Hollywood	9	k-Anonymization

Table 8: DM values for different attributes

First two records of table 8 show the DM values for gender and region under simple anonymization. An interpretation for those two results will be the cost of generalization of gender is higher than cost of generalization of region. At the same time, we can say that more originally distinguishable values have become indistinguishable under region generalization, but at a lesser cost.

Figure 5 shows the DM variation within a quasi-identifier and how each generalization has costed. Through that we can get an idea about what are the costliest generalizations.

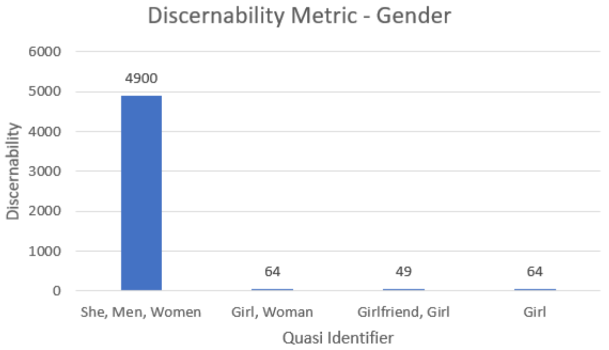


Figure 5: Discernibility metrics – gender.

And also, if we closely look at the values of the qid groups, we can understand the fact that the tagger is performing very well with extracting attributes as all the attributes depicted in the graphs/tables are meaningful in their category.

Table 9 summarizes the loss metric values for each qid attribute. For the sample validation dataset used loss metric seems to be close to 0.8 for both the privacy types. Sanitization count and DM values seem slightly lower for k-anonymization than simple anonymization, but LM values are almost equal for both privacy types.

Quasi Identifier Group	LM	QID values	Privacy Type
QIGENDER	0.83	She, he, girls, men, girlfriends, shemales	Simple Anonymization
QIREGION	0.75	Odisha, China, Narsipatnam, Africa	Simple Anonymization
QIGENDER	0.86	Girl, women, girlfriend, woman, girls, men, she	k-Anonymization
QIREGION	0.8	Hollywood, Carmel, Delhi, Sindh, China	k-Anonymization

Table 9: LM values for different qids

The prototype enables the user to perform the anonymization operation and calculate these metrics through the framework itself to get a better idea.

7. Discussion and Future Work

This research has focused on building a privacy preserving data publishing middleware for textual, unstructured social media data and has successfully achieved that objective. As an additional contribution to the research community, this research has developed a reliable dataset with annotated tweets for privacy related attributes. In order to measure the usability of the newly generated data, utility metric calculation is embedded as a part of the framework. And specifically, this framework supports simple anonymization and k-anonymization which are very popular in the research community to preserve privacy of structured data. Therefore, this research can be considered as an integration of traditional privacy preserving approaches to textual and unstructured data in a novel way.

As future work, the framework can be enhanced by introducing some innovative features.

- As the tagger built in this research relies on a decision tree-based approach, some different approaches can be tried out using sequence tagging mechanisms to improve the accuracy.
- The dataset can be enhanced with introducing tweets with other different quasi identifiers than the ones used in this research
- User can be given the ability to define the attributes that are important to them, forming the foundation to personalized privacy

8. Conclusion

The core objective of this research was to come up with a novel framework, that can preserve the privacy of unstructured, textual social media data before publishing to any analytical platform. In order to achieve this task, a dataset was created and tagged with privacy related attribute tags. This dataset can be utilized by the research community to perform privacy preserving tasks on unstructured data in the future as well. This research comes up with an end to end framework for privacy preserving data publishing of unstructured data, including steps like attribute extraction, attribute sanitization and utility evaluation. The main attribute extraction module comes up with a F1 score of 0.7 for most of the quasi identifiers. Additionally, some points for improvement and promising future work too are discussed in this paper.

9. Ethics Statement

The dataset created in the research was built selectively based on a publicly available Kaggle dataset and it is not targeting any specific individual. Intermediate results containing personal data of any anonymization job will not be persisted for future use. The concept and research work are fully independent and impartial.

10. Bibliographical References

Alaphilippe A., Gizikis A., Hanot C., Automated tackling of disinformation, 2019.

Bu Y., Fu A.W.C., Wong R.C.W., Chen L., Li J., Preserving serial data publishing by role composition, in Proc. Very Large Database Endowment, 2008, pp. 845–856.

Carreras X., Marquez L., Padro L., A Simple Named Entity Extractor using AdaBoost, in Proc. Conference on Computational Natural Language Learning, 2003.

Chen B.C., Kifer D., LeFevre K., Machanavajjhala A. Privacy-preserving data publishing, in Proc. Foundations and Trends in Databases Conference, 2009, pp. 1 – 167.

Duan Y., Wang J., Kam M., and Canny J. Privacy preserving link analysis on dynamic weighted graph in Computational & Mathematical Organization Theory, 2005, pp.141–159

Fan L., Jin H., A Practical Framework for Privacy Preserving Data Analytics, in Proc. 24th International Conference on World Wide Web, 2015.

Fung B.C.M., Wang K., Philip S.Y., Introduction to Privacy-preserving Data Publishing: Concepts and Techniques. Boca Raton: CRC Press, 2010.

Fung B.C.M., Wang K., Chen R., and Yu P. S., Privacy preserving data publishing: A survey on recent developments, in ACM Computing Surveys, 2010, pp. 14:1 – 14:53

Gardner J. and Xiong L., An integrated framework for deidentifying heterogeneous data, in Proc. Data and Knowledge Engineering, 2009, pp. 1441-1451.

General Data Protection Regulation [Online]. Available: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

Industrial-Strength Natural Language Processing [Online]. Available: https://spacy.io/

Li N., Li T., Venkatasubramanian S., t-Closeness: Beyond k-Anonymity and l-Diversity, in IEEE 23rd International Conference on Data Engineering, 2007.

Liu C., Mittal P. Linkmirage: Enabling privacy preserving analytics on social relationships, in NDSS, 2016, pp. 21-24.

Liu K., Das K., Grandison T., and Kargupta H. preserving data analysis on graphs and social networks, In H. Kargupta, J. Han, P. Yu, R. Motwani, and V. Kumar, editors, Next Generation Data Mining. CRC Press, 2008.

Machanavajjhala A., Gehrke J., Kifer D., Venkitasubramaniam M., l-diversity: Privacy beyond kanonymity, in Proc. 22nd International Conference on Data Engineering (ICDE). IEEE Computer Society, 2006.

Mehta B., Rao U., Privacy preserving unstructured big data analytics – issues and challenges, in Proc. International Conference on Security and Privacy, Nagpur, India, 2015, pp. 120-124.

Mendes R., Vilela J.P, Privacy-preserving data mining: Methods, metrics, and applications, in IEEE Access, 2017, pp. 10562–10582.

Number of social media users worldwide from 2010 to 2021 (in billions) [Online]. Available: https://www.statista.com/statistics/278414/number-ofworldwide-social-network-users/

Samarati P., Sweeney L., Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and cell suppression, Technical report, SRI International, 1998.

Thavavel V., Sivakumar S., A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment, in International Journal of Computer Science Issues, 2012, pp. 434-441.

The “localization” of Russian citizens’ personal data [Online]. Available:https://home.kpmg/be/en/home/insights/2018/09/the-localisation-of-russian-citizens-personal-data.html

11. Language Resource References

Twitter Sentiment Analysis [Online]. Available: https://www.kaggle.com/paoloripamonti/twitter-sentiment-analysis