Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique

Poornima Kulkarni[1]
Assistant Professor, Department of ISE
RV College of Engineering, Bengaluru, India

Cauvery N K[2]
Professor, Department of ISE
RV College of Engineering, Bengaluru, India

***Abstract*—Personally Identifiable Information (PII) has gained much attention with the rapid development of technologies and the exploitation of information relating to an individual. The corporates and other organizations store a large amount of information that is primarily disseminated in the form of emails that include personnel information of the user, employee, and customers. The security aspects of PII storage have been ignored, raising serious security concerns onindividual privacy. A significant concern arises about comprehending the responsibilities regarding the uses of PII. However, in real-time scenarios, email data is regarded as unstructured text data, detecting PII from such an unstructured large text corpus is quite challenging. This paper presents an intelligent clustering approach for automatically detecting personally identifiable information (PII) from a large text corpus. The focus of the proposed study is to design a model that receives text content and detects possible PII attributes. Therefore, this paper presents a clustering-based PII Model (C-PPIM) based on NLP and unsupervised learning to address detection of PII in the unstructured large text corpus. NLP is used to perform topic modeling, and Byte mLSTM, a different approach of sequence model, is implemented to address clustering problems in PII detection. The performance analysis of the proposed model is carried out existing hierarchical clustering concerning silhouette and cohesion score. The outcome indicatedthe effectiveness of the proposed system that highlights significant PII attributes, with significant scope in real-time implementation. In contrast, existing techniques are too expensive to function and fit in real-time environments.**

***Keywords*—PII; natural language processing; word2vec machine learning; PII detection; security**

I. Introduction

The progressive digitization of functional domains of various processes in individual human and business contexts produces various data types. The data are generated in text format, audio format, video format, image format, and many more custom formats. The business objectives often demand to store or archive these data for longer, making it voluminous. The analogy of various formats of data and larger size of it is popularized in the recent past as verity and volume, respectively [1]. The ever-evolving business models at the pace of technological advancements have provided possibilities of innovative business applications in the healthcare industry, banking & finance, education, aviation, Défense etc. There are many contexts where certain information in these data is very private to a user or a system [2]. Providing necessary security to this private information becomes essential to mitigate associated risk due to system design vulnerabilities, potential threats, and attacks. The study of security towards this private information is popular as privacy preservation. The complexities and challenges of designing effective privacy preservation methods depend solely on the type of the data format, its size, and the data flow into the application. However, another popular term appears in the context of designing security models to preserve privacy – "Personally Identifiable Information" (PII). The data which can be used to identify a person is the higher layer meaning of PII [3]. Fig. 1 shows the relationship between PII and private information.

In the last decade, the increasing popularity of the Internet and PII collection has raised serious concerns about privacy policy. A surge in personal data breaches for an in-depth understanding of preferences, authenticating customers and employees has become a common occurrence in data-driven organizations. Even government organizations heavily rely on the collection of PII to carry out an important decision. Therefore, people are required to share their personal data. However, a significant concern arises towards comprehending the responsibilities regarding the uses of PII and its protection [4]. More specifically, the people are not knowing, how the government organization and corporates are handling individuals' PII. PII collection over the Internet is highly prone to identity theft, social engineering attacks and is most vulnerable to fraudulent criminals [5]. The evidence in the report provided by the Data Breach Level Index shows that 76.20% of data breaches in 2018 belonged to the social media industry, most of which were related to identity theft [6]. It is not surprising that PII has already turned out to be the new resource, and threat modeling for privacy is at the peak of the industry 4.0 revolution [7]. In addition, the cases of data breaches in the past recent years fascinated a compulsive concern towards PII in the research community. In response, significant efforts have been devoted in the existing literature to developing solutions against PII disclosure. The literature presented many solutions considering different contexts, such as PII on social networking sites, mobile applications, mobile network traffic, healthcare, corporate internal communications, and many more [8-10]. However, PII identification schemes in the existing literature do not provide comprehensive insight. Most of the existing schemes are based on rule-based approaches, limiting the scope and applicability of existing solutions in the real context. The rule-based approaches are based on the set of procedures and principles to represent knowledge from the structured data. Since, in real-world cases, the organization mostly maintains a large corpus which stores PII in the textual data format such as emails, contracts, IPv4 and MAC addresses, and telephone numbers. Among these textual data, emails contain more entropy compared to the other textual data. Apart from this, the textual corpus, especially email, is mostly unstructured since the information is presented in the native format, especially the email contents written with the different writing styles, contains the short subjective textual body. Therefore, the rule-based approaches are not much suitable for identifying PII from the unstructured large text corpus because they mainly deal with structured data formats. However, with the advent of machine learning (ML) models and advancement in natural language processing (NLP), PII of individuals from large unstructured text corpus can be efficiently identified, which cannot be addressed by applying the existing rule-based solution discussed so far [11-13]. In the existing literature, many efforts have been put forward by the researchers. But, the research work on detecting PII in unstructured text corpus is minimal. The existing literature does not focus on detecting full PII and its diversity that reflects user privacy and identity differently. A particular type of PII is easy to be identified, but it is crucial to consider the sources or categories of the content that reflects highly vulnerable PII. Therefore, the existing study lacks topic modeling and also suffers from huge computational overhead.

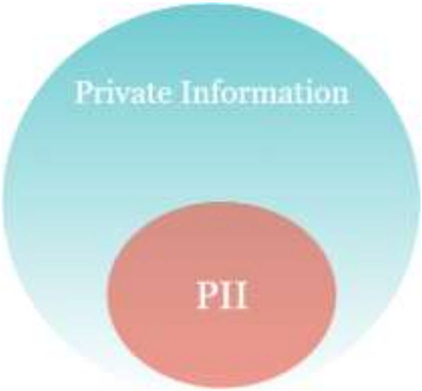


Fig. 1. Privacy Information and Personally Identifiable Information (PII ⊂ PI).

In the current study, the email dataset is considered an unstructured text corpus for the model evaluation. The email is mainly composed of low-quality text and one of the potential sources of the PII disclosure. The proposed research considers PII detection from the unstructured text corpus as a text mining and clustering problem. It introduces a clustering-based PII detection model (C-PIIM) using natural language processing (NLP) and an unsupervised learning algorithm. NLP is employed to perform topic modeling using Word-to-Vec and Bag of Word models. On the other hand, Byte-mLSTM as an unsupervised learning algorithm is used to detect full-PII from the text corpus. The performance validation of the proposed system C-PIIM is carried out concerning clustering performance metrics and PII probability in the text corpus. The significant contribution of the proposed work can be summarized as follows:

- The proposed system addresses the problem of automatically detecting and classifying the possibly included PII attributes from the large text data.
- The study also addresses the problem of precise feature extraction from low-quality textdata by introducing an effective data modeling and processing mechanism.
- A topic modeling is done to determine a set of contexts that show which category of text document has the most probable vulnerable PII.
- The hybrid nature of the deep learning technique is employed to achievehigher accuracy inclassifying the PPI from the email or text data.

The design and development of the proposed system are carried out in such a manner that it can meet corporate production requirements by automatingmonitoring and detecting PII and ensuringagreement.

The remaining sections of this paper are arranged in the following manner. Section II discusses the related work on personnel data leakage and PII detection; Section III presents the proposed system C-PIIM design and its implementation, followed by systematic data modeling and discussion; Section IV presents the result and discussion; finally, Section V concludes the entire work of this paper.

II. Related Work

The incidents of information breaches and openings attracted the wide attention of security agencies and government, particularly if it encompasses PII. Corporate companies, healthcare industries, and social networking sites are the most attractive targets for the attackers looking for PII, which can be used for identity theft, and even unauthorized access to sensitive data. Over the years, personal data leaking has been widely studied in the existing literature. A recent study by Go et al. [14] raises serious concern on the opening and un-intended disclosure of information on social media platforms. The authors suggest an intelligent and flexibly centralized model based on software-defined networking with virtualization to counteract unintended PII opening on the network traffic. This work improves the previous work done by Liu et al. [15], where automatic detection of PII is carried out by employing a set of systematic expressions and dictionary-based methods. Another work in this context by Ren et al. [16] proposes a model, namely ReCon, to address the discloser of PII in the mobile network traffic using the supervised learning model. ReCon leverages the decision tree to identify and block the open PII in the mobile network traffic.

The application of the learning technique, as in the work of Noever [17], proposes a method to identify the Person of interest (PoI) and PII from the email dataset of the corporate company. This method uses a mechanism of ensembled learning that considers decision trees, support vector classifier, random forest, and neural network. The PoI and PII are identified from the text analysis, such as financial records and emails. This study also analyses the sentiments of several employees regarding corporate crises. Zaeem and Barber [18] focus on the increasing PII misuses over the Internet. This study reveals an interesting statistic concerning lacking user privacy preservation incorporate companies in North America. The author collected data from the stock exchange platform and labeled the collected information with different rating scores. This study provides an effective direction so that corporate industries can improvise their privacy law.

In critical applications such as healthcare, researches are carried out to find useful information from the medical data or records to aid emergency management. However, the healthcare records usually consist of patient information, and the researchers often encounter PII that needs to be protected. In this regard, Michael et al. [19] discussed securing PII for the human participant and health research using a big data tool. Similarly, Alnemari et al. [20] focus on protecting PII available in healthcare data. The authors have presented an interesting discussion on the existing techniques for protecting PII. The study findings showed that multiple attribute workload distribution is more effective than the traditional anonymization and differential privacy approaches to obscure PII while allowing the researchers to conduct analysis efficiently.

The work of Onik et al. [21] presented an intelligent risk classification model based on accumulated mobile application permission data to classify vulnerable PII associated with the mobile owners. The authors have developed a google-play API to collect permission data of several android applications. They adopted different ML classifiers to detect the most significant PII such as contact number, social graph, email, location, biometric ID, and a unique ID. This study has presented significant work, but it has achieved less accuracy than similar existing research works due to training the model with fewer data.The work of Majeed et al. [22] suggested an improved vulnerability-aware PII anonymization scheme to ensure user privacy. The author used random forest mechanisms to identify the identity of the most vulnerable PII and used the Simpson index to calculate the diversity to reduce the risk of PII disclosure.

The work of Venkatanathan et al. [23] examined the impact of public and private data opening on social media toward disclosure of the PII to strangers. This study has demonstrated that the wall posts and extensive descriptions of individuals on their profile pages in social networks trade significant privacy leakage to the world. The authors have also presented an analytical design for privacy and PII masking. The work of Tesfay et al. [24] studied the challenges and issues in discovery of PII from textual data using ontologies, NLP and learning approaches. This study has considered the both PII and privacy sensitive information with all relevant definitions, and terminologies. A systematic approach is carried out to explore different types of the problem considering the information which needs to be regarded. Liu et al. [25] provided significant work on the identification of PII based on the information related to the user behavior in the network traffic. Firstly, the authors have discussed on the role of application and internet service provider that collects data user information to enhance their quality of experience, traffic control and improve security services. Further, the authors have discussed the challenges in identifying PII from the massive network traffic. The authors then presented an efficient algorithm, namely TPII, to address the detection of PII in the massive and complex nature of traffic data. A concept of a decision tree with an optimization approach is used to perform the classification of the PII. This study uses a dataset of real data gathered from a university network with more than 10k users. The work of Vishwamitra et al. [26] suggested a collaboratively controlling mechanism for protecting PII for photo sharing over the social network. The presented scheme is designed based on the multiparty access control mechanism and policy specification-based detection process.

III. Research Problem

In literature, a number of research works have been presented for privacy security concerning PII. This section highlights some significant issues associated with the existing approaches as follows:

- It has been analyzed that research works on detecting PII in unstructured text corpus using ML and NL is quite limited.
- Most of the existing works do not providea comprehensive insight into data modeling and its processing regarding PII discovery.
- However, they have achieved a good result for structured data, but the existing techniques may show poor performance for low-quality data, like email,which is often unstructured as it contains many acronyms, short-text, and errors.
- The existing studies lack topic modeling, which is important when dealing with private information discovery.
- It has also been noticed that the previous prediction-based approaches based on the traditional machine learning and statistical techniqueslack efficiencyandhence, suffer from huge computational overhead to achieve higher accuracy in the privacy information detection and classification process.

Hence, the problems mentioned above are addressed in the proposed study. The next section details the proposed system and strategy adopted in its implementation.

IV. System Design and Implementation

The proposed research study focuses on detecting PII from the large unstructured text corpus to advance privacy practices and create better regulations in the communication processes. The study believes that the chain of custody of the data, i.e., information in the internal communication, must not be broken and openly disclosed. The PII in the data must be retrieved only when it is necessary.Therefore, an effective and automated system is developed to identify PII from the large text corpus that consists of emails and text messages. The current work does not intend to detect quasi PII. Instead, the method is devised to identify full PII to check if there is a direct identifier such as name, email address and social security numbers etc., based on the subject and contents of the text corpus. The introduced model adopts the application of ML and NLP for clustering the full-PII identifiers and non-PII identifiers from a large unstructured text corpus. The schematic architecture of the proposed model is described in Fig. 2.

The implementation of the proposed model follows a systematic data modeling to achieve a suitable and precise feature vector for the clustering process using unsupervised learning for PIIidentification from the unstructured text corpus. The system design consists of a total of five core modules, namely, i) Dataset and its importance, ii) Dataset visualization, iii) pre-processing, iv) Topic modeling using NLP, and v) mLSTM implementation for PII identification.

A. Dataset and its Importance

This section presents a brief description of the dataset and its importance in the context of PII detection in corporate companies. A collaborative consortium of the 22 institutions under the flagship of the Artificial intelligence center of SRI, international has initiated a project, namely, "for Cognitive Agent that Learns and Organizes (CALO)". The "Enron email dataset" was collected during the project CALO. There is an interesting background behind the creation of this dataset an investigation study by the Federal Energy Regulatory Commission (FERC) behind the malpractices followed within the eco-system of the Enron corporation – a flagship company that has revenue up to 101 billion USD till the year 2000 and became bankrupt in the year 2001. The FERC made the Enron-email dataset public for the first time, containing email communication data from 150 senior management with approximately half a million messages. The various acquisitions and transitions in the dataset have taken place from time to time. In this paper, a version of the dataset published by Carnegie Mellon University in 2015 is used, obtained from the Kaggle. The top contributors are the researchers majorly from the two countries, including the United States and India. The statistics in Table I show its current popularity status.

B. Dataset Visualization

In this section, exploratory analysis is carried to visualize the characteristics and attributes of the data set to comprehend the need for pre-processing over the text dataset for further processing in the topic modeling. Since this is the initial phase of the proposed system implementation, the first step is to import the dataset (DS) available in the form of .csv to the data frame (DF) of the computing environment, which represents the entire structure of DS in tabular representation with rows (R) and columns (C). The structure of DF is illustrated with text samples in Table II.

In Table II, visualization of DF is carried out that represents a total of 517401 categorical data ∈ R corresponding to 2 headers fields such that {file, message} ∈ C, where the file (F) refers to the actual placewhere all the mail is stored that contains user information (U_I), type of inbox (T_{inb}) and subfolder (S_F) on the mail, such that $F \in C \supseteq \{U_I, T_{inb}, S_F\}$. On the other hand, message (M_{sg}) consists of the email content (body) and header (H) of the email. A sample view of $M_{sg}$is shown in Fig. 3.

Further, the analysis on the 'file'∈ C is carried out towards identifying its uniqueness, and it is found that it does not associate with any repetitive and duplicate entries. The next section presents pre-processing operation to split the header and body of the M_{sg} for the PII detection.

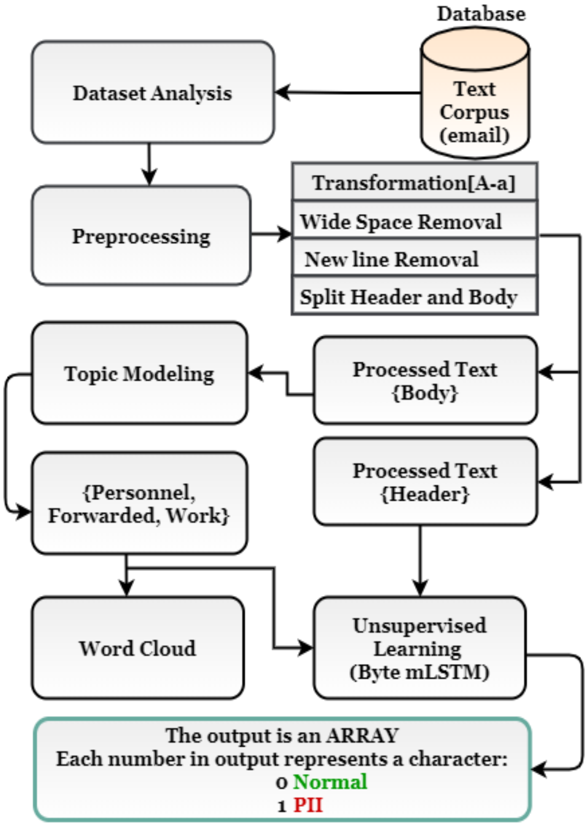


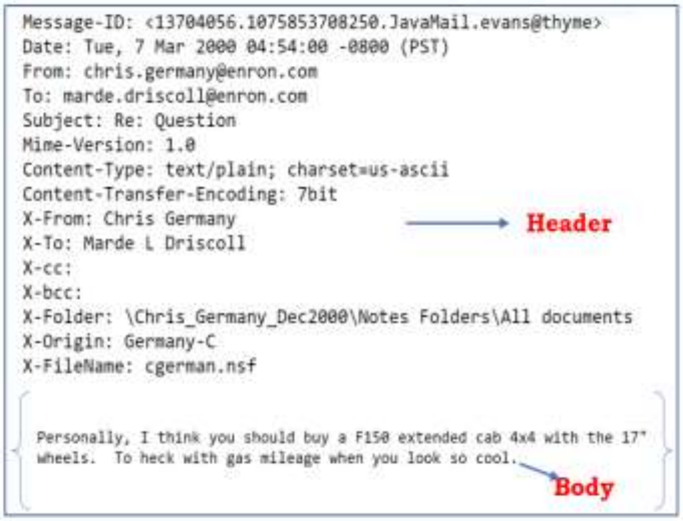
Fig. 2. Schematic Architecture of the Proposed Model.

TABLE I. POPULARITY INDICATOR OF THE ENRON-EMAIL DATASET

Views	Downloads	Download per view ratio	Total unique contribution
245000	29300	0.12	180

TABLE II. VISUALIZATION OF DF

SI. No	file	message
0	allen-p/_sent_mail/1.	Message-ID: <18782981. 1075855378110.JavaMail. e
1	allen-p/_sent_mail/10.	Message-ID: <15464986. 1075855378456.JavaMail. e
2	allen-p/_sent_mail/100.	Message-ID: <24216240. 1075855687451.JavaMail. e
3	allen-p/_sent_mail/1000.	Message-ID: <13505866.1075863688222.JavaMail.e
⋮	⋮	⋮
517400	zufferli-j/sent_items/99	Message-ID:<28618979. 1075842030037.JavaMail. e



```
Message-ID: <13704056.1075853708250.JavaMail.evans@thyme>
Date: Tue, 7 Mar 2000 04:54:00 -0800 (PST)
From: chris.germany@enron.com
To: marde.driscoll@enron.com
Subject: Re: Question
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Chris Germany
X-To: Marde L Driscoll
X-cc:
X-bcc:
X-Folder: \Chris_Germany_Dec2000\Notes Folders\All documents
X-Origin: Germany-C
X-FileName: cgerman.nsf

Personally, I think you should buy a F150 extended cab 4x4 with the 17"
wheels.  To heck with gas mileage when you look so cool.
```

Fig. 3. Illustration of the Message Field in the Dataset.

C. Dataset Preprocessing

Based on the exploratory analysis, it has been analyzed that the email dataset is associated with a different form of text representation; some are small, and some texts are capital in between the sentences. It has also been identified that most email bodies have wide spaces and new line characters, which may introduce uncertainty and ambiguity in the learning model. Therefore, the study considers removing wide spaces, newline characters, and transforming capital letters into small letters. Further, the algorithm performs a data splitting operation to separate the header and body of the messages filed of the email dataset. The above-mentioned algorithm demonstrates pre-processing operation carried out for removing irrelevancy in the dataset, which after processing provides Header and Body of message field as a separate vector. However, the data in M_{sg} also consists of stop words and punctuations with low entropy, but in the current study, it is considered significant in topic modeling.

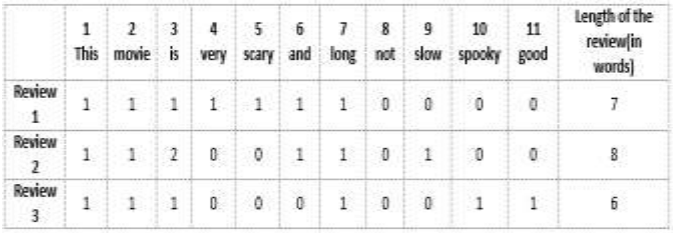
Algorithm-1email Data Pre-processing

```
Input: $M_{sg}$
Output: Header (H) and Body (B)
Start:
Init H →[ ], B →[ ]
                Load → DF[$M_{sg}$]
        def function: email_prepros($M_{sg}$)
text_lower = re.findall($M_{sg}$, ′(. ||[a − z]||[A − Z]) ))do
                $M_{sg}$ = text.lower($M_{sg}$)
        $M_{sg}$ = text.replace_arg($M_{sg}$, :\n + ′ )
[$M_{sg}$]←Text.split($M_{sg}$)
                    return = H, B
                For each R from $M_{sg}$ do
                DF$_p$ = email_prepros($M_{sg}$)
H←append.Header
B←append.Body
End
```

D. Topic Modelling

The PII has different nature thatreflects user privacy and identity differently. The identification of a certain type of PII is not a much challenging task. However, identifying the most vulnerable PII in unstructured text corpus is a challenging task. If such vulnerable PII is identified or exposed in the data, their presence may increase privacy risks.Once compromised, attackers can use the data to obtain PII to harm users by social engineering attacks or identity theft. Therefore, organizations or individuals must assess vulnerable PII to reduce the risk of privacy leakage. Therefore, the proposed study performs topic modeling to examine a set of documents and extract the categories and groups that reflect probable vulnerable PII. The proposed study considers topic modeling as an NLP problem and adopts the Word2vec model [27], an unsupervised text clustering model. The model takes the document text or email body (B) as input and gives a topic vector as an output. The document is converted to a vector with the help of the bag of words (BoW) technique, which is a word matrix that represents the count of each word in B. The sample representation of the word matrix for text data is illustrated in Fig. 4 where each row (R) of the matrix represents a B and the column represents tokenized text contained in the B of the email.

The word matrix obtained from the BoW is then fed to the word2vec algorithm to group emails into a cluster based on the topics needed for the study. However, the number of clusters can be limited, and in the current study, it has been limited to three clusters: personnel emails, work emails, and forward emails. Fig. 5 illustrates the working procedure of the word2vec algorithm m.



	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

Fig. 4. A Bag of Words Algorithm

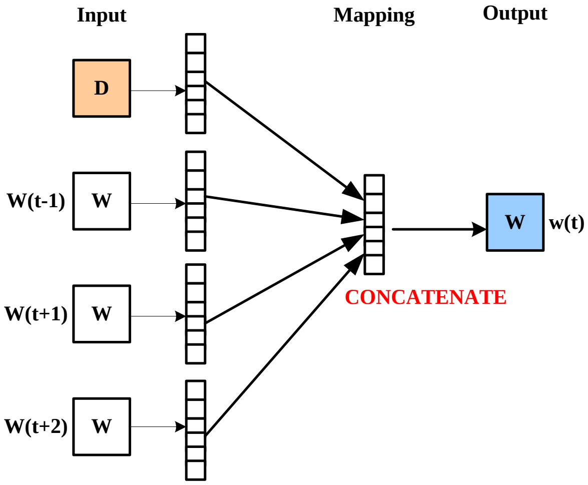


Fig. 5. Architecture of Word2vec Algorithm.

Once the model is trained, it can be used for text partitioning. Text partitioning is a collective term used to assign a topic to the text corpus and separate the corpus based on the topics extracted. In this case, the text is separated based on whether the email is personal or work or forwarded. This classification is preferred since, logically speaking, personal mails may reveal more PII than work email and forwarded. It should be noted that the work email contains more sensitive and confidential data than personal email. However, the scope of this study is to protect individual privacy rather than protecting the company's data. Other modules will be in a position to prevent data leakage. With this, the topics will be assigned to respective clusters, and new data is obtained that willbe further provided as an input to the next algorithm for PII detection. The output of the word2vec for topic modeling is demonstrated using word cloud as follows.

The word cloud of personal email is shown in Fig. 6, which reveals that the words are mostly related to personal topics like vacation planning or ranting about the employers etc. This is the place where most of the quasi and full PII can be exposed since people may speak about their family and friends in such emails. The next word cloud is shown for the work email.

In Fig. 7, the word cloud is shown for work emails that describe most of the words repeated here are related to the company's fuel and energy, which is the company's main domain. Since most of the discussion is about the company work, the chances of exposure of PII are significantly less. However, the people speak about their colleagues and employers, and hence there will be PII exposure in these emails as well. For example, the word 'Pallen'shown in the word cloud is an important person in the company.

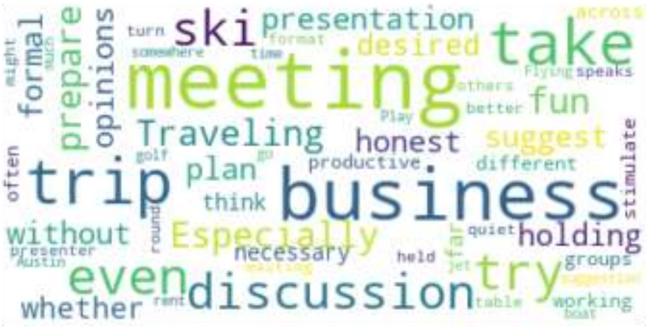


Fig. 6. Word Cloud for Personal Email.

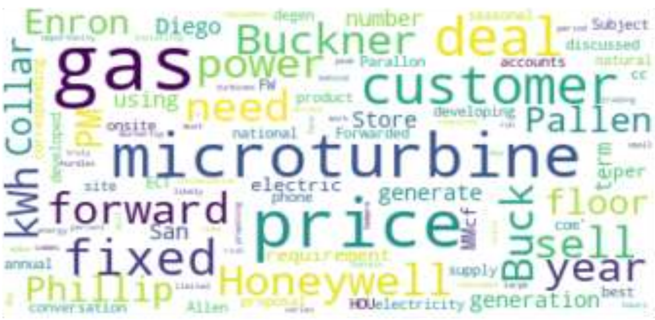


Fig. 7. Word Cloud for Work Email.

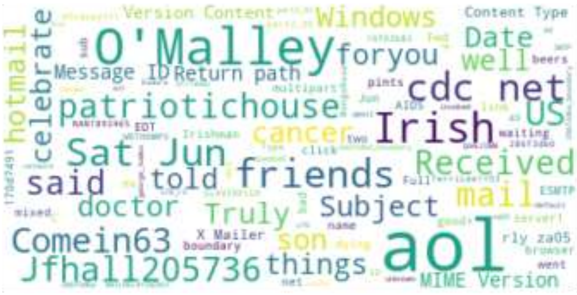


Fig. 8. Word Cloud for a Forwarded Email.

The word cloud for forwarded email isdepicted in Fig. 8 that showsdiscussions were more about the general context, such as jokes or some commonly shared information. These types of emails may contain very little PII. The next section presents an unsupervised learning model, namely byte mLSTMalgorithm, for detecting PII in the text data Corpus. Byte mLSTM is an unsupervised approach for NLP or text processing [2]. In the present study, it is being used to detect PII.

E. Byte-mLSTM

Long-short-termmemory (LSTM) is an improved version of Recurrent Neural Network (RNN) model, suitable for sequence classification problems. In the proposed study, a different form of LSTM, i.e., byte mLSTM algorithm, is used to perform the clustering of text into normal text and PII. The mLSTM is given by Krause et al. [28] designed based on the joint approach of LSTM and multiplicative RNN architecture. The detection of PII in the proposed study using Byte m-LSTM can be described as clustering problem of n samples such that $X = \{x_1, x_2 x_2 \cdots L\}$ into K groups described by c_i using a nonlinear mapping function $f_\alpha: X \rightarrow Y$, where x denotes input sequences of length L where $X \in R^{d\times 1}$ and $c_i$denotes centroid and $i = \{1,2,3 \cdots K\}$, α is the training parameters, and $Y \in R^K$is the embedded feature space from the model. For a given input sequence X to the learning model Byte m-LSTM, the encoder $E_i \epsilon R^c$ at time t_i and LSTM cell c and Decoder combinedly trained to reconstruct input feature vector space X in inverse order by reducing the following objective as follows:

$$\sum_{X\epsilon N_t} \sum_{i=1}^{L} ||x_i - x_i'||^2 \ldots. \quad (1)$$

Where, N_t refers to the training set sequences, and L output state of the encoder is used as the initial input x_i to the decoder part of the the-LSTM model. The decoder part then gets a state $S \epsilon D^{i-1}$ and maps to the expected output space x_{i-1}' corresponding to the target x_{1-1} vector space. In the training phase, the model employed CAHL (Clustering assignment hardening loss) function, which is explicit and most suitable for the clustering mechanism. The proposed study considers stabilized resemblances between data points and c_i as a soft assignment (q_{ik}) of k with statistical estimation technique to measure the resemblance between learned features y_i and c_i. The probability of assigning sample i to k can be numerically represented as follows:

$$q_{ik} = \frac{\left(1+\left|\left|y_i - c_k\right|\right|^2/\varphi\right)^{-\varphi+1/_2}}{\Sigma_k\left(1+\left|\left|y_i - c_k\right|\right|^2/\varphi\right)^{-\varphi+1/_2}} \quad (2)$$

Where y_i corresponding to $x_1 \epsilon X$ after feature representation and φ denotes the degree of freedom. In the proposed study,φ is constant, which is considered equal to 1. In the next step of the resemblance between the distributions is evaluated using KLD by reducing the distance between q_{ik} and the supplementary distribution P and probability distribution Q as follows:

$$KL(P||Q) = \Sigma_i \Sigma_j p_{ij} \log {}^{p_{ij}}/_{q_{ij}} \quad (3)$$

Where KL denotes relative entropy, p_{ij} and q_{ij} ϵP and Q, respectively, are enhanced and adjusted through the backpropagation mechanism of the neural network, resulting inthe reduced distance between the data samples corresponding assigned k towards a better quality of clusters. However, the data points of a cluster closer to the average value of another cluster may adversely impact the loss. Therefore, q_{ik} is further normalized based on the frequency for each cluster K and raised by power factor 2 numerically expressed as follows:

$$p_{ij} = \frac{{}^{q_{ij}^2}/_{f_i}}{\Sigma_j {}^{q_{ij'}^2}/_{f_{i'}}} \quad (4)$$

where, f_i denotes q_{ik} frequency

The proposed architecture of mLSTM for text clustering is given in Fig. 9, consisting of two blocks, i.e., encoder and decoder LSTM. The encoder parts consist of multiple LSTM cells in parallel to the decoder part.

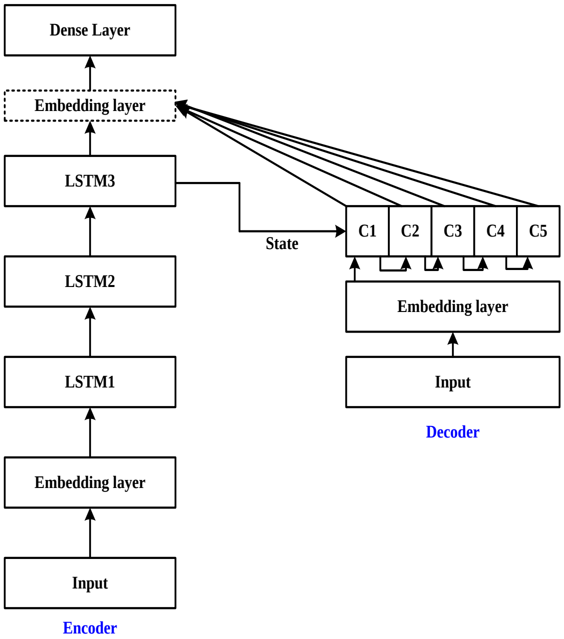


Fig. 9. The Schematic Architecture of mLSTM.

It can be observed that the LSTM-1 layer of the decoder network is placed parallel to LSTM 3 of the encoder layer. In the proposed study, the input to the learning model is carried out in the sequence of ASCII code of each letter in the input text. The encoder is an LSTM layer that accepts delayed input in the form of a vector forwards it to LSTM one by one in a sequence. The training of implemented learning model is carried out by loss function, namely, CAHL.The model takes input as H, clusters obtained from previous Algorithm word2vec, i.e., personnel (P), work (W), and forwarded (F) with the help of an encoder at the input layer. The encoder takes the input {H,P,W, F}and converts every character into ASCII code, which is fed one by one in a sequence. The model considers a sequence encoder instead of a window function, as the sequence encoder always works better in NLP problems. At the same time, the Decoder is takingthe output of the previous algorithm, i.e., word2vec, and also header information which is non-sequential data. In this process, layer 3 of the encoder and layer 1 of the Decoder functions parallelly. The output of the Decoder LSTM-1 layer is multiplied by the output of encoder LSTM-3. The multiplication operation in m-LSTM is carried out using an embedding layer that acts as a multiplier, letting the encoder input pass or stop based on the output of the decoder layer. The decoder LSTM also changes its output according to the input given in the encoder, and due to this reason,the LSTM-3 layer of the encoder is made aware ofthe LSTM1 layer of the Decoder.

V. Discussion and Perspective

The management of private information and its legislation is one of the important issues for organizations today. For instance, email is one of the main sources of communication in most corporate companies. However, this is also one of the potential sources of privacy leaks. In this regard, companies need an effective technology that can automate the process of managing personal information, assisting in monitoring PII leaks in the workflow process.Although the previous works have shown considerable efforts in private information discovery, to date, none of the existing techniques are more effective at identifying potentially vulnerable PII and its source. The proposed study attempted to address this issue by designing an effective mechanism of topic modeling that shows different categories of text documents containing sources and different aspects of the vulnerable PII. The proposed work focuses on detecting full PII followed by clustering operations that allow to determine and derive valuable patterns between text elements that help to offers important meaning to distinguish structures of PII. Unlike previous techniques, the proposed study uses hybrid nature of deep learning mechanism that detects and highlights the presence of PII in the given text data without depending on other external resources and rule-based approaches. The advantage of the adopted learning mechanism is that it achieves better generalization by taking advantage of both LSTM and multi-placative RNN (m-RNN) algorithms. The LSTM is good at handling natural language as sequence classification problems. With the implementation of a suitable embedding and encoding layer, LSTM can generalize the actual meaning in the text data. The advantage of m-RNN is that it generalizes long-term dependencies between the input text feeds and achieves stability in the overall learning process.Therefore, the learning model of unsupervised nature employed in the proposed system is efficient in addressing the problem of modeling text elements under a variable context. Another significance of the proposed model is that it can support and handle large and unstructured text corpus and faster responsive features. The model's design is user interactive, flexible to different lengths of data usage, and offers a better discovery of private information than the existing system.

VI. Result and Performance Analysis

The design and implementation of the proposed system are carried out using the Python programming language. The current study's entire work followed systematic data modeling and a clustering approach to identify PII from the unstructured text corpus.The outcome obtained for the proposed C-PIIM based on NLP and unsupervised learning employed in the study yielded promising results. This section discusses the results and performance analysis of the proposed C-PIIM concerning the type of mails, message head, and body. Also, the model's effectiveness is justified based on the comparative analysis of the proposed clustering method with the existing hierarchal clustering method [29] regarding two clustering performance metrics such as silhouette and cohesion score. The performance metric cohesion can be defined as a performance indicator that shows how fit the class methods are identical to each other. The performance metric silhouette can be described as a performance indicator that measures the quality of a clustering technique, numerically expressed as follows:

$$\text{silhouette} = \frac{(q-p)}{\max(p,q)} \quad (5)$$

The analysis from Fig. 10 shows thata maximum number of texts are within 10,000 – 20,000 characters. The total number of characters includes the characters in the header as well as the body and whitespaces. There are some mails whose length goes up to 2,00,000 characters upon further inspection; such lengthy mails are regarding knowledge transfer. However, a significant observation can be made from the analysis that most PII is being exposed in shorter mails. The following analysis in Fig. 11 is carried out regarding PII identification (%) in various mails determined from the topic modeling.

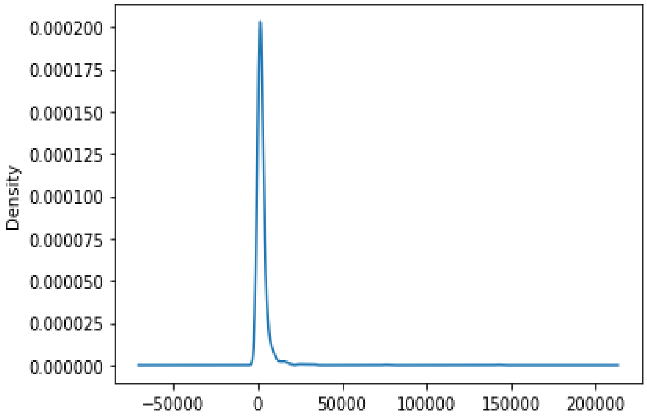


Fig. 10. Histogram for the Length of Emails in a Number of Characters.

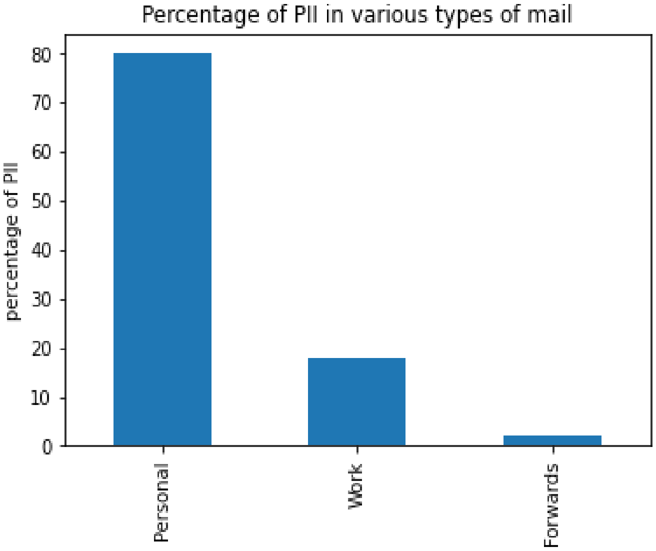


Fig. 11. Percentage of PII in Various Types of Mail.

The graph trend from Fig. 11 exhibits that personnel mail has a higher percentage of PII exposer than the work category and forward category of mails. The outcome shows 80% of the PII is exposed in personal mail, 18% is exposed in work emails, and only around 2% is exposed in the forwarded emails. Therefore, from the analysis, most of the PII is exposed in the personal mails.

In Fig. 12, analysis is carried out concerning message contents. The graph trend exhibits, the headers contain more PII compared to the body contents. This comes as no surprise as the header always contains the email and name of both sender and receiver. However, this will be prevented by the system for exposure of PII. The main focus is the body of the mail, where 20% of the PII is exposed. It exposes vital data which may be used to cause a financial loss to the individual. The next analysis presents the proposed system's comparative analysis with the existing hierarchical method regarding culturing performance metrics.

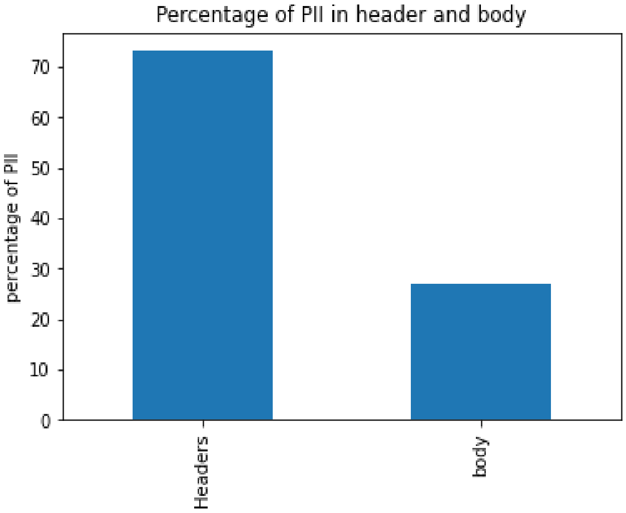


Fig. 12. Percentage of PII in Header and Body.

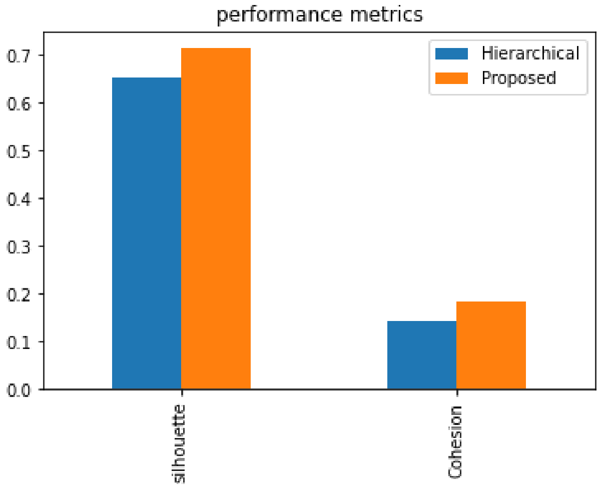
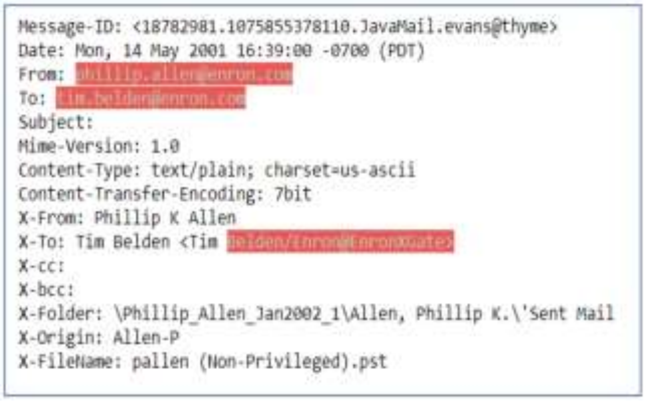


Fig. 13. Comparative Analysis.

Fig. 13 exhibits that the proposed C-PIIM outperforms the existing method in terms of silhouette and cohesion metric. From the analysis, the proposed system achieved 0.7156 and 0.1832 silhouette and cohesion scores, respectively. In contrast, the existing method has achieved 0.1421 and 0.1832 silhouette and cohesion scores, respectively, proving the proposed model'seffectiveness and applicability for real-time implementation. The next Fig. 14 presents the outcome achieved by the proposed model C-PIIM.



```
Message-ID: <18782981.1075855378110.JavaMail.evans@thyme>
Date: Mon, 14 May 2001 16:39:00 -0700 (PDT)
From: phillip.allen@enron.com
To: tim.belden@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Tim Belden <Tim Belden/Enron@EnronXGate>
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Jan2002_1\Allen, Phillip K.\'Sent Mail
X-Origin: Allen-P
X-FileName: pallen (Non-Privileged).pst
```

Fig. 14. Detection of PII.

Fig. 14 exhibits the outcome of the proposed C-PIIM that highlights the presence of PII in the email (text corpus) of the user, that contains addresses unique to sender and receiver as well as it also reflects the email address of some senior person of the company, which may be vulnerable to social engineering threat or identity theft.

VII. Conclusion

As PII collection continues to increase, the costs of data breaches also increase, ranging from economic losses to reputation losses. For understanding the risks associated with privacy opening, several efforts have been made in the literature to detect PII disclosure and leaks. However, there are limited works regarding PII detection in the large unstructured text corpus. In this paper, NLP and Byte-mLSTM mechanisms are used to design an effective model for the purpose of PII leak detection.The proposed system comprises topic modeling for segmenting and grouping data storage categories that account for disclosure of potential or most vulnerable PII. Byte mLSTM as unsupervised learning is employed as an effective clustering mechanism to detect vulnerable PII in the text data. The study outcome proved the effectiveness of the proposed clustering-oriented PII detection compared to the existing hierarchical clustering approach. The proposed model can be used in real-time scenarios like in the corporates to warn their employees against sending PII included in the email. However, the model is limited to detection of full PII, cannot detect Quasi PII. The proposed study will be extended in future work considering the quasi PII and de-identification process.

References

[1] S. Chenthara, H. Wang, K. Ahmed, “Security and Privacy in Big Data Environment”, In: Sakr S., Zomaya A.Y. (eds) Encyclopedia of Big Data Technologies. Springer, Cham, 2019.

[2] M. Petrescu, A.S. Krishen, “Analyzing the analytics: data privacy concerns”, J Market, vol. 6, pp. 41–43, 2018.

[3] F. Alizadeh, T. Jakobi, J. Boldt, and G. Stevens, “Gdpr-reality check on the right to access data: Claiming and investigating personally identifiable data from companies”, In Proceedings of Mensch und Computer, pp. 811-814, 2019.

[4] Boyd JH, Randall SM, Ferrante AM. Application of privacy-preserving techniques in operational record linkage centres. Medical data privacy handbook. 2015:267-87.

[5] Pawlicka A, Choraś M, Pawlicki M. Cyberspace threats: not only hackers and criminals. Raising the awareness of selected unusual cyberspace actors-cybersecurity researchers' perspective. InProceedings of the 15th International Conference on Availability, Reliability and Security 2020 Aug 25 (pp. 1-11).

[6] “Data breach statistics.” https:// breachlevelindex.com/, Retrieved on 25th September 2021.

[7] M. M. H. ONIK, C. KIM and J. YANG, "Personal Data Privacy Challenges of the Fourth Industrial Revolution," 21st International Conference on Advanced Communication Technology (ICACT), pp. 635-638, 2019.

[8] A. Iyengar, A. Kundu, G. Pallis, “Healthcare informatics and privacy”, IEEE Internet Computing, vol. 22(2), pp. 29-31, 2018.

[9] D. Hiatt, Y.B. Choi, “Role of security in social networking”, International Journal of Advanced Computer Science and Applications, vol.7(2), 2016.

[10] D.R. Pope, Y.H. Hu, M.A. Hoppa, “A Survey on Securing Personally Identifiable Information on Smartphones”, Virginia Journal of Science, vol.71(3), 2020.

[11] P. Silva, C. Gonçalves, C. Godinho, N. Antunes, M. Curado, “Using natural language processing to detect privacy violations in online contracts”, InProceedings of the 35th Annual ACM Symposium on Applied Computing, pp. 1305-1307, 2020.

[12] R.N. Zaeem, M. Manoharan, Y. Yang, K.S. Barber, “Modeling and analysis of identity threat behaviors through text mining of identity theft stories”, Computers & Security, vol. 1, 65, pp. 50-63, 2017.

[13] S. Applebaum, T. Gaber, A. Ahmed, “Signature-based and Machine-Learning-based Web Application Firewalls: A Short Survey”, Procedia Computer Science, Jan 1;189:359-67, 2021.

[14] S.J. Go, R. Guinto, C.A. Festin, I. Austria, R. Ocampo, W.M. Tan, “An SDN/NFV-enabled architecture for detecting personally identifiable information leaks on network traffic”, In Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), pp. 306-311, 2019.

[15] Y. Liu, H. H. Song, I. Bermudez, A. Mislove, M. Baldi, and A. Tongaonkar, "Identifying personal information in internet traffic," in Proceedings of the 2015 ACM on Conference on Online Social Networks, COSN '15, (New York, NY, USA), pp. 59–70, ACM, 2015.

[16] J. Ren, A. Rao, M. Lindorfer, A. Legout, and D. Choffnes, "Recon: Revealing and controlling pii leaks in mobile network traffic," in Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, MobiSys '16, (New York, NY, USA), pp. 361–374, ACM, 2016.

[17] D. Noever "The Enron Corpus: Where the Email Bodies are Buried?", arXiv preprint arXiv:2001.10374, 2020.

[18] Z.R. Nokhbeh, K.S. Barber, "A study of web privacy policies across industries", Journal of Information Privacy and Security,vol.13(4), pp.169-85, 2017.

[19] M.D. Bader, S.J. Mooney, A.G. Rundle, "Protecting personally identifiable information when using online geographic tools for public health research", Am J Public Health, pp. 206-208, 2016.

[20] A. Alnemari, R.K. Raj, C.J. Romanowski, S. Mishra, "Protecting personally identifiable information (pii) in critical infrastructure data using differential privacy", In IEEE International Symposium on Technologies for Homeland Security (HST) , pp. 1-6, 2019.

[21] M.M. Onik, C.S. Kim, N.Y. Lee, J. Yang, "Personal Information Classification on Aggregated Android Application's Permissions", Applied Sciences, (19), pp. 39-97, 2019.

[22] A. Majeed, F. Ullah, S. Lee, "Vulnerability-and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data", Sensors, vol.17(5), pp.1059, 2017.

[23] J. Venkatanathan, V. Kostakos, E. Karapanos, J. Gonçalves, "Online disclosure of personally identifiable information with strangers: Effects of public and private sharing, Interacting with Computers, vol. 26(6):614-26, 2014.

[24] W.B. Tesfay, J.M. Serna, and S. Pape, "Challenges in Detecting Privacy Revealing Information in Unstructured Text", In PrivOn@ ISWC, 2016.

[25] Y. Liu, T. Song, L. Liao, "TPII: tracking personally identifiable information via user behaviors in HTTP traffic", Frontiers of Computer Science, vol. 14(3):1-4, 2020.

[26] N. Vishwamitra, Y. Li, K. Wang, H. Hu, K. Caine, G.J. Ahn, "Towards pii-based multiparty access control for photo sharing in online social networks', In Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies, pp. 155-166, 2017.

[27] Y. Hu, J. B-Graber, B. Satinoff, "Interactive topic modeling", Mach Learn 95, pp.423–469, 2014.

[28] B. Krause, L. Lu, I. Murray, S. Renals, "Multiplicative LSTM for sequence modelling", arXiv preprint arXiv:1609.07959,2016.

[29] F. Nielsen, "Hierarchical Clustering. In: Introduction to HPC with MPI for Data Science", Undergraduate Topics in Computer Science. Springer, Cham, 2016.