# An Advanced Semantic Feature-Based Cross-Domain PII Detection, De-Identification, and Re-Identification Model Using Ensemble Learning

Poornima Kulkarni[1], Cauvery N K[2], Hemavathy R[3]
Department of ISE, RV College of Engineering, Bengaluru, India[1, 2]
Department of CSE, RV College of Engineering, Bengaluru, India[3]

***Abstract*—The digital data being core to any system requires communication across peers and human machine interfaces; however, ensuring (data) security and privacy remains a challenge for the industries, especially under the threat of man-in-the-middle attacks, intruders and even ill-intended unauthorized access at warehouses. Almost all digital communication practices embody personally identifiable information (PII) like an individual's address, contact details, identification credentials etc. The unauthorized or ill-intended access to these PII attributes can cause major losses to the individual and therefore it is inevitable to identify and de-identify aforesaid PII elements across digital platforms to preserve privacy. Unfortunately, the diversity of PII attributes across disciplines makes it challenging for state-of-arts to perform PII detection by using a predefined dictionary. The model developed for a specific PII type can't be universally viable for other disciplines. Moreover, applying multiple dictionaries for the different disciplines can make a solution more exhaustive. To alleviate these challenges, in this paper a robust ensemble of ensemble learning assisted semantic feature driven cross-discipline PII detection and de-identification model (EESD-PII) is proposed. To achieve it, a large set of text queries encompassing diverse PII attributes including personal credentials, healthcare data, finance attributes etc. were considered for training based PII detection and classification. The input texts were processed for the different preprocessing tasks including stopping-word removal, punctuation removal, website-link removal, lower case conversion, lemmatization and tokenization. The tokenized text was processed for Word2Vec driven continuous bag-of-word (CBOW) embedding that not only provided latent feature space for analytics but also enabled de-identification to preserve security aspects. To address class-imbalance problems, synthetic minority over-sampling techniques like SMOTE, SMOTE-BL, SMOTE-ENN were applied. Subsequently, the resampled features were processed for the feature selection by using Wilcoxon Rank Sum Test (WRST) method that in sync with 95% confidence interval retained the most significant features. The selected features were processed for Min-Max Normalization to alleviate over-fitting and convergence problems, while the normalized feature vector was classified by using ensemble of ensemble learning model encompassing Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier as base classifier. The proposed model performed a consensus-based majority voting ensemble to annotate each text-query as PII or Non-PII data. The positively annotated query can later be processed for dictionary-based PII attribute masking to achieve de-identification. Though, the use of semantic embedding serves the purpose towards NLP-based PII detection, de identification and re-identification tasks. The simulation results reveal that the proposed EESD-PII model achieves PII annotation accuracy of 99.77%, precision 99.81%, recall 99.63% and F-Measure of 99.71%.***

***Keywords*—PII Detection; machine learning; natural language processing; artificial intelligence; de-identification***

## I. INTRODUCTION

The last few years have witnessed significantly high-pace rise in digital data and allied communication. Increasing internet uses has broadened the horizon for digital data communication to serve socialization, business communication, networking etc. [1]. In fact, the modern world with exponentially rising population can't be hypothesized to exist without digital data. The digital data obtained by means of Email, Blogs, reviews, diagnosis details, business communication etc. can have personally identifiable elements including phone number, email ID, bank account details, social security numbers, diagnosis or ailments, insurance details, vehicle number, passport number etc. [1], [2]. The unauthorized and ill-intended intentional or unintentional use of these personally identifiable elements can cause both privacy breaches as well as losses [3-7]. For instance, an intruder can misuse insurance number and diagnosis details to cause cyber frauds and financial losses. Similarly, the unauthorized access to the bank account number, user identity number can give passage to the cyber criminals to cause digital crimes. Even the disclosure of house number to an unexpected intruder might broaden the horizon for stalking etc. [3], [4], [7]. To alleviate these issues, ensuring abstraction or masking of aforesaid personally identifiable elements can be of great significance. Though, the organizations claim to use personally identifiable information (PII) for constructive decisions including market-segmentation, profiling, marketing, business communication [8], [9]; however, the likelihood of PII misuse within organization or outside the mainstream mechanism can't be ruled out [10-13]. The unauthorized users or intruders might exploit one's online digital behaviour including texts, reviews, Email, personal chats etc. to misuse foresaid PII elements. To ensure privacy preserved communication identifying aforesaid PII is really a challenging task [5-7]. It becomes even more difficult to detect PII elements from unstructured and heterogenous data, which is unavoidable in modern digital world [6], [7].

Though, the classical approaches apply dictionary-based methods to detect and identify PII elements in text; however, preparing a large set of dictionaries for each discipline is near

infeasible or highly exhaustive. Moreover, the PII elements vary across the disciplines. For example, the PII elements pertaining to the user's personal identify might differ from the one from healthcare domain. In this case, generalizing the dictionary from one discipline is infeasible for the other and hence a model trained over one type of dictionary can't be used for other [14]. It limits efficacy of these solution towards real-time PII identification tasks. On the other hand, it is really difficult to prepare dictionary for each discipline distinctly. To alleviate these challenges, though a few machine-learning (ML) based solutions are proposed in the past [14]; however, almost all state-of-arts have been prepared for single domain and hence their scalability towards other discipline remains suspicious. To alleviate it, recently, the organization named European Unions suggested General Data Protection Regulation (GDPR) define PII attributes over broader spectrum to enable PII named identify representation (NER) for the different domains or disciplines [7]. However, ensuring both PII element heterogeneity, contextual details and optimal computing remained challenge for industry to ensure PII detection, de-identification and re-identification optimal. It can limit the performance of solution over large heterogenous data environment [15-17]. The lack of contextual details over large heterogenous and unstructured data makes PII detection challenging [15][16][18][19] and complex [20]. Though, to reduce computational exhaustion fast search space method was proposed [20] where expression matching was done over text content to perform PII detection. Ontology methods too can lack contextual dependencies to perform PII detection [17]. Though, the natural language programming (NLP) methods possess greater significance with the ML classifiers; however, ensuring feature optimality is inevitable to achieve better performance [21]. Despite efforts the diversity of PII elements amongst the different discipline such as healthcare [18][19], legal text documents [22], user browsing data and email [23] and academic or non-academic publication [24], make major state-of-arts confined and generalizing a solution over other discipline texts can yield low accuracy. This is because almost all ML-driven solutions perform PII detection for a specific data type or discipline [24], [25]. It clearly indicates that a model designed to perform PII detection in financial data can't be applicable in EHR-related dataset [75]. It motivates researchers to design a cross-domain learning environment, where one can take the inputs text data from the different domains (i.e., healthcare, business communication, financial data etc.) and train a model to have broader knowledge ability for PII detection and classification. Ironically, none of the state-of-art could address this problem so far. Though, the use of NLP with better feature engineering, contextual learning and robust classification can enable a cross-domain PII detection and classification. However, it requires training a robust learning model with maximum possible features encompassing samples from the different domains. Despite numerous studies in the past, none of the solutions could address class-imbalance problem while using ML or NLP methods for PII detection and classification. On the contrary, it is inevitable that the real-time systems might have text queries relatively higher for non-PII than the PII containing queries or sentences. Such skewed data and allied learning might force model(s) yielding false positive or false negative. Therefore, an NLP driven PII detection and classification model requires addressing class-imbalance problem to improve learning and classification results. Additionally, processing a training model over most representative PII text samples with embedded latent or semantic details can make learning more efficient and accurate. These key inferences can be considered as the predominant motivation behind this study.

Considering aforesaid motivations and allied scopes, in this paper a novel and robust ensemble of ensemble learning assisted semantic feature driven cross-discipline PII detection and de-identification model (EESD-PII) is proposed. To achieve it, a large set of text queries encompassing diverse PII attributes including personal credentials, healthcare data, finance attributes etc. were considered for training based PII detection and classification. The input texts were processed for the different pre-processing tasks including stopping-word removal, punctuation removal, website-link removal, lower case conversion, lemmatization and tokenization. The tokenized text was processed for Word2Vec-CBOW embedding that not only provided latent feature space for analytics but also enabled de-identification to preserve security aspects. To address class-imbalance problem, synthetic minority over-sampling techniques like SMOTE, SMOTE-BL, SMOTE-ENN were applied. Subsequently, the resampled features were processed for the feature selection by using WRST algorithm that in sync with 95% confidence interval retained the most significant features. The selected features were processed for Min-Max Normalization to alleviate over-fitting and convergence problems, while the normalized feature vector was classified by using ensemble of ensemble learning model encompassing Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier as base classifier. The proposed ensemble learning model performed consensus-based majority voting ensemble to annotate each text-query as PII or Non-PII data. The positively annotated query was later processed for a dictionary-based PII attribute detection and masking to achieve de-identification. The simulation results reveal that the proposed EESD-PII model achieves PII annotation accuracy of 99.77%, precision 99.81%, recall 99.63% and F-Measure of 99.71%.

The remaining sections in this paper are divided as follows. Section II discusses the related work, while the problem formulation and questions are given in Section III and Section IV, respectively. The overall proposed model is discussed in Section V, while the proposed model is given in Section VI. Conclusion and allied inferences are given in Section VII, while Section VIII discusses the future scope. The references used are given at the end of the manuscript.

## II. Related Work

The authors in study [26] applied conditional random field (CRF) and SVM to learn the pre-annotated PII lexical and embedding features to classify PII query. On the other hand, the authors [27] infused the phrase embedding concept to yield semantic details, which was later trained over machine learning to perform named entity detection and classification. On the other hand, the allied challenges in named entity recognition were discussed in study [28], where the authors concluded that the use of text-specific semantic details trained over better learning environment can perform more accurate PII identification. However, the authors failed to indicate their

suitability over the different sets of text inputs. The authors [29] used long and short-term memory (LSTM) deep network to extract and learn contextual features from the input queries to perform PII detection. The recurrent neural networks (RNNs) were applied in [30], [31] as well where the authors used pre-trained transformer-based language models for PII prediction. These approaches couldn't address large PII heterogeneity amongst queries from the different disciplines. Moreover, merely applying cross-validation learning can't generalize the solution for cross-discipline PII detection and classification [32]. In study [33], PrivacyBot was designed, especially to perform privacy sensitive PII detection over unstructured text input. The authors applied machine learning models to train over the text content for classification. Yet, their generalizability over large unstructured muti-source inputs remains suspicious [34]. In [35] deep learning was applied to classify annotated text corpus for PII classification. The authors in study [36] made use of the convolutional neural network to perform named entity recognition in unstructured text data. However, it was completely depending on the local hierarchical features, while a non-lineal environment demands a model to have contextual learning capabilities to yield accurate prediction results. In study [37], the conceptual relationship amongst the unstructured text was applied to perform PII detection from input texts. More specifically, an association extraction approach was applied on Twitter data to assess the relatedness of each word with PII attributes by using pointwise mutual information statistical association rule. The authors in study [38] exploited privacy information across privacy-related topics, including plan for the vacations, alcohol and healthcare conditions to perform PII detection and classification. Despite ML classifier, guaranteeing scalability of solution over large heterogenous inputs makes it suspicious. In [39], a semi-supervised ML [17] was applied for PII detection in healthcare text. The authors in study [17] fused topic modelling, privacy ontology and sentiment score, which was later processed by Naïve Bayes (NB) classifier to detect PII attribute in Twitter data. In study [40], PII dictionary was applied to annotate text elements as PII or non-PII. In study [41], the authors used ontologies with rule-based method to perform PII detection in text. The authors in study [42] designed a text feature learning driven PII detection interface for software-defined networks (SDN) that classifies each query as PII or non-PII. Though, it failed addressing numerous issues including feature optimality, class-imbalance and low accuracy problem. In [43], the authors used dictionary and intra-query structural details to perform PII detection and classification. RenCon [44] was proposed to detect PII in mobile network traffic, where supervised ML was used to annotate each query as PII or Non-PII. A similar work was done in study [5] where the authors applied ML algorithm to learn PII attributes personal identification (ID), phone number, social graph, email, location and biometric ID detail for PII detection. In [45] as well ML was applied where the Person of interest (PoI) and PII attributes were detected from the email dataset. The authors suggested that unlike standalone classifier, the use of ensemble learning method by using decision tree, SVM, neural network and random forest can yield better accuracy. PII detection over healthcare data was done in study [46], yet, despite BigData problem the authors failed in addressing the problem of class-imbalance [47]. The authors in study [48] used random forest algorithm with Simpson index that measured the diversity amongst the PII elements to annotate them as PII or non-PII element. Ontology method was applied in [49], [50] to detect PII and its masking. In [50], ML was applied to solve PII detection as an NLP problem.

In study [70], the authors focused on developing a new Korean dialogic dataset, especially designed for the PII de-identification. For PII detection, the authors used text anonymization benchmark and network intrusion detection dataset, based on which a new de-identification dataset was prepared. Unfortunately, there are very less efforts made towards text data de-identification; though there exists certain works related to the face de-identification in video or image dataset. For instance, in study [71] the authors developed a disentangled representation learning for multiple attributes preserving face de-identification. More specifically, they proposed replacing and restoring variational autoencoders (R 2 VAEs) that disentangle the identity-related factors and the identity-independent factors so that the identity-related information can be obfuscated, while they do not change the identity-independent attribute information. In [72], UU-Net was developed for the reversible face de-identification in visual surveillance video data. Here, the proposed UU-Net model learns jointly by optimizing a public module that receives the raw data and generates the de-identified stream, along with a private module, especially designed for security authorities. The second module receives the public stream and regenerate the original data by exploiting semantic and contextual details, disclosing the actual IDs of the subjects in a scene. This method made use of the conditional generative adversarial network to achieve synthetic faces by preserving pose, lighting, background information and even facial expressions. In study [73], the authors developed de-identification method by applying both human perceptions derived as semantic information and the face recognition models. This method explored the tradeoff between a user misidentifying the original identity with a well-known celebrity and a facial recognition model that tries to identify the original identity. It generated caricature faces of the de-identified faces to ensure that the manipulated faces can be distinguished effortlessly. In study [74], an attribute-preserving face de-identification framework called Enhanced Embedded Auto Encoders was developed. The proposed model embodied three components, privacy removal network (PRN), feature selection network and privacy evaluation network. Here, PRN made the model capable to discard information involving identity privacy and retaining desired face attributes for certain prediction applications. The other modules focused on exploiting global or contextual features to improve learning-based classification accuracy. Despite efforts, ensuring de-identification in text data, especially carrying critical details like electronic healthcare records (EHR) remains a challenge [75]. It broadens the horizon for the researchers to develop a scalable and robust PII detection, classification and deidentification and re-identification [75] model.

## III. PROBLEM FORMULATION

This research hypothesizes that the inclusion of PII texts from the different disciplines for improved NLP analytics can enable cross-discipline PII detection and classification. This hypothesis considers optimization on both data as well as

computational front, where initially it embodies PII queries from the different domains which are processed for the different pre-processing tasks including stopping-word removal, punctuation removal, website-link removal, lower case conversion, lemmatization and tokenization. This approach can strengthen the proposed PII detection model to address data heterogeneity and unstructured-ness and hence can make it suitable for the real-world applications. In addition, it can make the proposed model suitable for cross-discipline PII detection and classification. Moreover, unlike syntactical term matching based solutions, to retain sufficiently large contextual details Word2Vec CBOW method is proposed that can provide large latent information to make cross-domain PII detection. It also helps in achieving de-identification by transforming input texts into equivalent embedding matrix, which can later be used to perform de-identification and re-identification. Realizing the at hand computational challenges like local minima and convergence, redundant computation, in addition to the aforesaid data optimization the proposed work focuses on computational aspects as well, where at first WRST significant predictor test is applied over the embedded metrics which retains a sufficiently large set of samples with high representativeness. It can reduce computational cost decisively while ensuring that the performance remains high. The selected samples have been processed for resampling by using SMOTE resampling methods including SMOTE, SMOTE-BL and SMOTEENN methods. Noticeably, these resampling methods intend to alleviate any probable class-imbalance problem and skewed data problem without undergoing iterative hotspot creation as witnessed with random sampling and up-sampling methods. The resampled data has been later processed for Wilcoxon Rank Sum Test (WRST) method that in sync with 95% confidence interval retained the most significant features. Subsequently, Min-Max normalization is performed that maps each input data in the range of 0 to 1, and thus alleviates the problem of over-fitting problem. Finally, the normalized feature vector was classified by using ensemble of ensemble learning model encompassing Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier as base classifier. The proposed ensemble learning model performed consensus-based majority voting ensemble to annotate each text-query as PII or Non-PII data. In this manner, it performs two-class classification which classifies each query as PII or Non-PII. Once annotating query as PII, the predefined dictionaries can be applied to perform search-based method to detect and mask PII element. Though, the proposed model applied word-embedding method that transformed original input text into equivalent embedding metrics and hence preserved privacy and/or de-identification to the data while retaining data-sanity for analytics.

## IV. RESEARCH QUESTIONS

In reference to the aforesaid formulations, this research defines certain questions, whose justifiable answers can put foundation for a robust PII detection and classification system. These research questions (RQ) are:

***RQ1***: Can the use of multi-disciplinary PII text with multi-aspects pre-processing tasks like stopping-word removal, punctuation removal, website-link removal, lower case conversion, lemmatization and tokenization enable NLP to achieve cross-discipline PII detection and classification?

***RQ2***: Can the use of Word2Vec CBOW embedding, WRST significant predictor test, SMOTE resampling, and Min-Max normalization provide semantically rich feature vector for PII detection and classification?

***RQ3***: Can the use of ensemble-of-ensemble learning environment encompassing Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier as base classifier to perform PII detection and classification?

***RQ4***: Can the strategic amalgamation of the solutions RQ1-RQ4 provide a robust cross-discipline PII detection and classification system?

As indicated in RQ1, the use of the text data and allied PII variables (say, artefacts) can enable an AI tool to learn more effectively. Here, training any machine learning or deep model over heterogenous PII elements belonging to the different categories of text inputs can improve learning ability and knowledge. Undeniably, it can improve learning; however, the data heterogeneity over the different input sources, data nature, diversity of presentation etc. can impact overall learning. To address this problem, there is the need of transforming input raw data into corresponding processed structured data for further feature extraction and learning. In this reference, this research defines research question (RQ1) which assesses whether the pre-processing tasks like stopping-word removal, punctuation removal, website-link removal, lower case conversion, lemmatization and tokenization enable NLP to achieve cross-discipline PII detection and classification. In the same manner, to improve scalability even with the reduced computational costs, this research intends to use contextual or latent semantic information, which can enhance learning and classification accuracy. To achieve it, this research makes use of the semantic embedding technologies like Word2Vec embedding. This method intends to transform input (pre-processed text data) into corresponding latent embedding matrix (low-dimensional matrix). To further improve computational efficacy and reliability over non-linear, imbalanced data environment, it performs SMOTE resampling (to alleviate class-imbalance problem), and WRST significant predictor test-based feature selection. These methods altogether can improve feature environment and hence can alleviate any likelihood of class-imbalance, premature convergence, local minima etc. In this reference, this research defines research question RQ2, whether aforesaid methods can improve overall performance towards a run-time PII detection and classification solution. Unlike traditional machine learning methods like decision tree (DT), support vector machine (SVM), naïve bayes (NB), etc. the ensemble learning methods like random forest (RF), AdaBoost, extra tree classifier, XGBoost etc. have performed better for the different data and image classification problems. Being consensus driven prediction solutions, it yields higher accuracy and reliability towards PII detection and classification. Considering it as motivation, this research intends to use only ensemble learning methods, such as the Bagging, Boosting, RF, ETC, XGBoost to perform maximum voting ensemble-based learning and classification. In this method, a total of five base classifiers (here, ensemble classifiers) are considered. This is because, out of five base classifiers, if three of the base classifiers classifies an input as PII and annotates it as 1, it would

be classified as PII. On the other hand, if three base classifiers annotate an input as 0, the final prediction is made as non-PII. Being consensus driven approach or maximum voting-based approach, it can be more reliable towards run-time significances. These aspects are defined in terms of the research question RQ3. Since, the at hand problem is a challenge of text classification, this work considers standard classification related performance parameters such as the accuracy, prediction, recall and F-Measure as performance variable. In this reference, this work defines research question RQ4. Thus, the overall research intends to achieve justifiable answer for these questions (RQ1-RQ4).

## V. System Model

This section discusses the overall proposed method and allied implementation. The overall research method encompasses the following key steps:

*a) Data Acquisition,*
*b) Pre-processing,*
*c) Word2Vec CBOW Embedding,*
*d) Significant Predictor Test,*
*e) Resampling,*
*f) Normalization, and*
*g) Ensemble(E2E) MVE Classification.*

A snippet of the overall proposed model is given in Fig. 1.

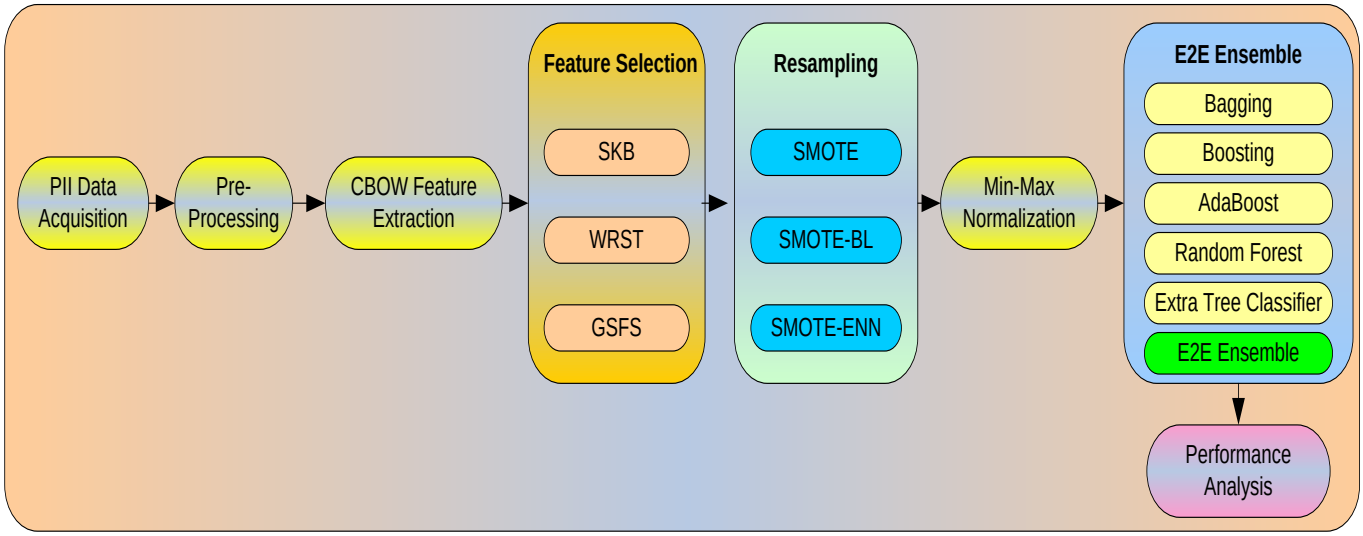


Fig. 1. Proposed EESD-PII model.

The detailed discussion of the overall proposed work is given as follows:

### A. Data Acquisition

Considering real-world heterogenous and unstructured text data available across digital communication platforms such as Email, Blogs, social media, data warehouses etc., in this research primary data was synthesized. Moreover, since there exists very limited data available towards cross-discipline PII detection problem, we synthesized own primary data encompassing sufficiently large text queries possessing PII elements from the different domains, including personal credentials, healthcare data, financial data etc. In sync with targeted PII detection and classification problem, Python Faker tool was taken into consideration where a total of 12000 text (reviews) queries were synthesized. Realizing the diversity of PII attributes across different disciplines, a set of PII elements were prepared for each discipline. A snippet of the different PII attributes used for the different discipline is given in Table I.

As depicted in Table I, text queries from three different domains including personal credentials, healthcare credentials and financial credentials were prepared. Here, personal credentials encompassed the key PII attributes as “Name”, “Address”, “Email”, “Aadhar Number”, “PAN”, “SSN”, “Vehicle Number”, and “Phone Number”. Similarly, three different PII credentials including “Health Insurance Number”, “Diagnosis Details”, and “Life Insurance Number” were considered. The other attributes like “bank Account Number”, “Loan Account”, “Balance Details”, “OTP”, “Access PIN” and “Credit Card Number”, were prepared towards financial credentials. We synthesized 4000 queries from each discipline and thus a total of 12000 queries were obtained. To synthesize these reviews Python’s Faker package was applied that embodies different callable functions to generate a large number of unstructured texts. The aforesaid text reviews were generated arbitrarily in such way that it embodies aforesaid PII credentials (Table I) in the different texts and allied disciplines. In other words, each of the queries generated possessed the different PII elements pertaining to the personal credentials, healthcare and finance. Thus, the queries generated provided data heterogeneity while ensuring data privacy and integrity. This approach also serves the purpose of PII privacy preserving and anonymized data preparation. Recalling the fact that the faker generates queries with the aforesaid credentials randomly and hence guaranteeing uniform proportion of each PII attribute remains difficult. In this manner, the data can be hypothesized to be unbalanced and hence can give rise to the skewed learning. It eventually can give rise to the false positive or false negative performance. To alleviate this problem, applying resampling can

be of great significance. The resampling can improve data distribution uniform throughout and hence training a model over such balanced sample can improve learning and classification accuracy.

TABLE I. PII ATTRIBUTES

| Discipline | PII Attributes | No. of Samples |
|---|---|---|
| Personal Credentials | Name<br>Address<br>Email<br>Aadhar Number<br>Permanent Account Number<br>Social Security Number<br>Vehicle Number<br>Phone Number | 4000 |
| Healthcare Credentials | Health Insurance Number<br>Diagnosis Details<br>Life Insurance Number | 4000 |
| Financial Credentials | Bank Account Number<br>Loan Account<br>Balance Details<br>One Time Password<br>Access PIN<br>Credit Card Number | 4000 |

### B. Pre-processing

In reference to the real-world data scenarios where the likelihood of data-heterogeneity and unstructured-ness can't be ruled out, strengthening data with the different pre-processing tool can be of great significance. Additionally, since the proposed method intends to exploit semantic word-embedding metrics (feature) improving data with certain set of pre-processing tasks can be of great significance. With this motive, we performed the different pre-processing tasks including the removal of stopping words, punctuation marks, Emoji, lower case conversion, lemmatization etc. We apply the following pre-processing tasks to ensure optimal data presentation and input towards targeted PII detection, de-identification and identification tasks.

*a) Missing value removal*
*b) Unicode normalization*
*c) Emoji removal*
*d) Website link removal*
*e) Punctuations and Stopping word removal*
*f) Converting to the lower case*
*g) Removing stop-words*
*h) Lemmatization*
*i) Tokenization.*

A snippet of these pre-processing tasks is given as follows:

*1) Missing Value Removal:* In real-world data environment, the text queries or reviews, especially collected from email, online Blogs, reviews, social media etc. can have certain broken sentences or incomplete words. There can be the set of words or strings with incomplete sentence representing the missing data element. Any NLP and allied AI-based solution, especially the text-analysis methods might yield inaccurate prediction results with aforesaid missing elements. In fact, the lack of contextual relatedness amongst aforesaid incomplete words or missing elements can impact overall performance. Learning over their allied features can impact learning and classification results. To address this problem, removing broken words and sentences from the input text corpuses can be vital. To achieve it, we applied standard NLTK Python library to remove the incomplete sentences. The overall process of missing value removal ensured that the proposed method examines and retains associations amongst the PII attributes and incomplete sentence(s).

*2) Unicode Normalization:* Realizing diversity amongst the reviews where the different users can have written the reviews with the different word-constitution, formats, (personalized) writing method, skills and texting designs, the proposed method performed Unicode normalization on each input query or sentence. The aforesaid data heterogeneity (amongst the word constituents and writing style) might impact NLP solution and hence transforming raw input text into a uniform structure is must. To achieve it, we performed Unicode normalization that transferred the input text into the single norm data output for further semantic feature extraction and corresponding de-identification tasks.

*3) Emoji Removal:* Though, the considered Faker tool didn't introduce any emoji element; however, in real-world data systems the likelihood of emoji (Ex. 😊, 😵) can't be ruled out, especially the modern mobile based text communications. Though, there exists no concrete literature which could prove their efficacy or significance towards NLP problems; however, their presence in raw data can impact learning adversely. In fact, readability of such elements can be difficult and can introduce data ambiguity that eventually can impact overall performance. Considering these facts, we performed emoji removal by using NLTK libraries and rule-based functions. This ability can make the proposed model robust in real-time applications where the targeted PII detection tool can be applied as an interface to detect, de-identify and re-identify PII attributes.

*4) Hashtag and URL Removal:* In present day digital world, where social media, web-media etc. have been playing decisive role in digital data promotions and allied marketing (say, communication). The users often try to mention the different URLs or web-links pertaining to certain subject matter. In addition, there can be user defined or organization driven hashtag information (i.e., #). The aforesaid URLs and #Hashtag helps the individual or organization gaining broader market space or audiences. Despite such significances, such web-URLs, links and/or #Hashtag don't have any significance towards PII detection and classification. The presence of such elements in texts can impact NLP learning and prediction (accuracy or) efficiency. It makes it inevitable to remove such elements before executing feature extraction and learning. In this paper, we applied rule-based methods to remove different hashtag and web-URLs. The rule-based method helped in identifying the URL components including www, https:// etc.,

that eventually helped in removing aforesaid contents or attributes. We applied Python URL removal function to remove hashtag and URL elements from each input query.

*5) Punctuation and Stopping Words Removal:* In text data the presence of the non-word symbols or the punctuation (i.e., ",", "?", "!",";", ":", etc.) can't be ruled out. Though, such elements have decisive significance towards contextual role or presentation; however, are not much related to the PII probability. Though, their presence in vicinity of a PII element, for instance, "AA0123456" can't be ruled out. Noticeably, though AA0123456 can be certain PII element; however, "" seems to be not related to the PII and just represents a highlighting element and/or remark. Therefore, removing such elements can be vital, especially when exploiting NLP for PII detection and classification. We applied NLTK library to remove punctuation marks from each text query or sentence. The deployed method applied standard expression and the rule-based approach to detect and remove punctuations from the input texts.

In general, the stop-words represent the terms that add no value to the sentence. In at hand NLP problem, the presence of these words can impact feature purity and its significance towards contextual (or semantic) learnability. Considering this fact, we dropped aforesaid stop-words from the input text corpus without making any change to the original intend behind the text phrase.

*6) Lower Case Conversion:* The ML algorithms can be case-sensitive and therefore the diversity amongst the case might impact learning efficiency. To cope up with the NLP solution, the relatedness and significance amongst the terms like SMART and smart can be differ and hence can cause contextual ambiguities in the extracted features. To alleviate this problem, the input texts were processed for lower case conversion, where each word or term was converted into lower case. We used Python's inbuilt lower-case conversion function to perform lower-case conversion.

*7) Lemmatization:* Once converting text data into equivalent lower-case sequence, each query was processed for lemmatization. In this method, the extended word was transformed into the respective root form. It assessed the expected component of the text without losing the original sense behind the word. Noticeably, stemming and lemmatization performs similar task; however, differs the way it (i.e., lemmatization) assesses the context first, which is then followed by the transformation of the extended word to the root word. On the contrary, in stemming the extended characters like "s", "es" etc. at the end of the word are removed that loses the actual intend of the word. The following illustrations depicts the process of stemming and lemmatization methods.

Studies → Lemmatization→ Study,

Studies → Stemming→ Studi.

As depicted above, to retain original intend or originality, lemmatization was performed where each word was transformed into respective original form.

*8) Tokenization:* Once performing lemmatization of the input text or queries, tokenization was performed that transformed each query into certain set of tokens. In this work, the tokenization method at first transformed input query or sentence into corresponding set of tokens, which were later processed for semantic feature extraction and learning to perform feature learning and classification.

### *C. Semantic Feature Extraction*

Though, in the past a few ML and NLP-based efforts have been made towards PII detection and classification; however, most of the state-of-arts fail in address or exploit contextual feature to perform learning and classification. Additionally, unlike term-matching based approaches the use of latent information or semantic features can improve efficacy, especially in NLP-based solutions including at hand PII detection and classification. Unlike traditional PII detection models, in this work the focus was made on exploiting depth or semantically enriched features or latent features from the input queries is hypothesized. In addition, to presence data sanity, originality with privacy preserving (say, de-identification), word-embedding can be of great significance. More specifically, the use of Word2Vec based Continuous Bag of Words (CBOW) model can yield low-dimensional semantic feature to perform learning and classification for PII detection and classification. Unlike other approaches such as TF-IDF or N-skip gram with $n = 1$, CBOW yields low-dimensional features to perform learning and classification towards PII-detection. The use of Word2Vec CBOW model can be even effective over the data encompassing PII attributes from the different disciplines, including personal credentials, healthcare and finance. With such data diversity, the proposed Word2Vec CBOW model can be vital. Moreover, in the considered dataset there can be limited text elements amongst the non-PII attributes and therefore retaining allied contextual details can provide more insight towards NLP-based PII detection and classification. Training a ML solution with aforesaid contextual and semantic (or latent) information can make PII detection more efficient and accurate over unannotated or minimally annotated inputs. Though, the literatures suggest different semantic extraction methods including TF-IDF, Skip-Gram, Glove, etc.; however, CBOW provides more intrinsic feature with easier implementation over large inputs. Additionally, it yields relatively low-dimensional features with intact feature vector or contextual significance, which can provide suitable feature environment for ML to predict PII in input sentence or query. In reference to these inferences, we applied Word2Vec CBOW method to perform word-embedding over the tokenized inputs for each query. In this method, for each query it generates set of embedding metrics, which is subsequently used for optimization and learning to perform PII detection. A brief of the CBOW method used in this work is given as follows:

In Word2Vec CBOW, we performed word-embedding on the input tokenized inputs (say, text sequences per query). More specifically, Gensim Word2Vec method was applied that transformed the set of tokens into corresponding low-

dimensional semantic embeddings (say, embedding vector). To achieve it, we designed Word2Vec with dual-layer neural network having two hidden layers. It helped achieving more contextually rich but sparser features so as to improve computational aspects. In CBOW method, let, $W_{i-1}$, $W_{i-2}$, $W_{i+1}, W_{i+2}$ be the context words retrieved from the sentence or review text. Then, CBOW predicts $W_i$ which is related with the other tokens available in that specific query. The predicted embedding outputs are always related to the target token $W_i$. The detailed discussion of CBOW embedding is given as follows:

In general, the CBOW method contains two sets of word-embedding vectors, say, "Source-side" and "Target-side" vectors, signifying $v_w, v'_w \in \mathbb{R}^d$ for each sentence or query's tokens. Noticeably, in our proposed method, we obtained $w \in V$ as the tokenized vocabulary once performing tokenization for each text input or sentence. Being Gensim-based embedding, a text window within the input review or sentence contains central token $w_0$ which generates respective context embedded vector $w_1, \dots, w_C$. For the above stated conditions, (say, text-window), the CBOW loss is measured as in Eq. (1).

$$v_c = \frac{1}{C}\sum_{j=1}^{C} v_{w_j} \tag{1}$$

$$\mathcal{L} = -\,log\,\sigma\left(v'_{w0}\,T_{v_c}\right) - \sum_{i=1}^{k} log\,\sigma\left(-v'_{n_i}\,T_{v_c}\right) \tag{2}$$

In (2) $n_1, \dots, n_k \in V$ be the negative examples obtained from the noise distribution $P_n$ over the input vector $V$. In Eq. (2), the parameter $\mathcal{L}$ gradient is estimated with respect to the target value $v'_{w0}$, negative target value $v'_{n_i}$ and average context source $(v_c)$.

$$\frac{\partial \mathcal{L}}{\partial v'_{w0}} = \left(\sigma\left(v'_{w0}\;T_{v_c}\right) - 1\right)v_c \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial v'_{n_i}} = \left(\sigma\left(v'_{n_i}\;T_{v_c}\right) - 1\right)v_c \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial v_c} = \left(\sigma\left(v'_{w0}\;T_{v_c}\right) - 1\right)v'_{w0} + \sum_{i=1}^{k}\left(\sigma\left(v'_{n_i}\;T_{v_c}\right) - 1\right)v'_{n_i} \tag{5}$$

Now, deploying Chain-rule method over the source context embedding, the gradient of the predicted word vector, also called the context vector is obtained as per Eq. (6).

$$\frac{\partial \mathcal{L}}{\partial v_{w_j}} = \frac{1}{C}\left[\left(\sigma\left(v'_{w0}\;T_{v_c}\right) - 1\right)v'_{w0} + \sum_{i=1}^{k}\left(\sigma\left(v'_{n_i}\;T_{v_c}\right) - 1\right)v'_{n_i}\right] \tag{6}$$

To address the problem of inaccurate context vector update, in this work context word normalization was performed. To achieve it, we sampled the context window's width arbitrarily in the predefined range of 1 to $C_{max}$ for each target word. Thus, applying this method the embedded metrics for each query was obtained for further processing or eventual analytics task(s). Noticeably, this method not only provided the embedded metrics as semantic feature for learning and classification; but also transformed the original text including PII elements or attributes into equivalent pseudonymized data and hence fulfils the goal of deidentification, without losing data intend and originality towards further analytics. Unlike data masking, our proposed model can provide a suitable reversible data form with certain rule-based approaches. Though, in the proposed work, such tasks are not required as the PII element itself has been transformed into anonymized embedding vector or numeric value.

### *D. Ensemble Feature Selection*

This is the matter of fact that in real-time applications, the volume of data inputs (here, text reviews) can be gigantic and hence training a ML method over such humongous data might give rise to the local minima and convergence and hence can yield inaccurate prediction result. Additionally, it can make overall processing exhaustive. The severity of such problems can be high in NLP domain, especially over the large text inputs. To alleviate these problems, retaining the set of most-significant samples having decisive impact on the (PII) prediction results can be vital. Though, there exists a number of state-of-art feature selection methods; however, in this work we examined applied three different methods, named Select K-Best, Wilcoxon Rank Sum Test and Gini-Score based significant predictor test to retain the optimal set of features, which were horizontally concatenated to yield an ensemble feature vector for further learning and classification. A brief of these methods is given as follows:

*1) Select k-Best Feature Selection:* The Select k-Best (SKB) method identifies the top-k most significant samples or allied features. We applied Chi-Squared method to perform Select K-Best based most significant feature selection. The applied Chi-Squared feature selection approach measures the level of significance of each feature to retain the top-K features. It is mainly achieved by estimating the value of $\chi^2$statistics in reference to the target class. In this mechanism, each feature is examined separately to assess the relationship between the feature and target class. It acts as a non-parametric test method which compares the different variables for an arbitrarily chosen data. As an independence-test approach, it identified the difference and independence amongst the different arbitrarily selected variables. Thus, it measured a value on the basis of the association between the feature instance and the class it must belong to. For output 0, it states that there exists no association between the feature element and the class. The greater association refers stronger relationship between the feature (say, sample) and the probable class. The use of Chi-Squared method was done by using scikit-learn library which helped retaining the set of most significant samples. In the developed approach, the Chi-square method performed in reference to the information-theoretic feature selection paradigm, where it measured the intuition that the best terms $t_k$for a class $c_i$ are the one distributed amongst the set of positive and negative examples of the class $c_i$. Chi-square's test can be derived as per Eq. (7).

$$Chi - Square(t_k, c_i) = \frac{N(AD-CB)^2}{(A+C)(B+D)(C+D)} \tag{7}$$

In Eq. (7), $N$ represents the total samples or features available in the input feature vector (for the considered text corpus), and $A$ be the elements in class $c_i$ with $t_k$. The feature elements carrying $t_k$ in other classes is defined as $B$. The number of features in class $c_i$ that don't possess any term of $t_k$

is given by $C.D$ on the other hand, represents the number of feature elements with no term $t_k$ in other classes. Thus, by using above Eq. (7), the proposed method assigned a score value for each feature towards a class (i.e., PII or Non-PII). Eventually, the scores are amalgamated to yield a final score, given as Eq. (8).

$$max\big(Chi - Square(t_k, c_i)\big) \tag{8}$$

Thus, with the obtained score as shown in Eq. (8), it retained the top-K features (or samples) to perform further processing.

*2) Wilcoxon Rank Sum Test:* It is also called as the significant predictor test which belongs to a type of non-parametric test with independent samples. Functionally, it measures the level of significance of each embedding metrics value and allied probability towards PII or Non-PII query. Here, WRST compares the location of the two samples by using two matched samples. More specifically, Signed Rank Sum test approach was applied to measure the level of significance. Consider, the paired data contains the samples $(X_1, Y_1), \dots, (X_n, Y_n)$, where each sample signifies a pair of measurements. These measurements are converted to the real-numbers, while the paired sample test is converted to the one-sample test by replacing each pair of numbers $(X_i, Y_i)$ by respective difference $X_1 - Y_1$. This method enables ranking of the difference in between the pairs. However, it needs that the data remains on an ordered metric scale. Consider the data for one-sample test be $X_1, \dots, X_n$, then assuming that the samples carry different absolute values, and no sample is equal to zero, then it applies the following steps to examine the level of significance of each data in the sample.

- ✓ Estimate the value $|X_1|, \dots, |X_n|$
- ✓ Sort $|X_1|, \dots, |X_n|$ and use the sorted-list to assign ranks $R_1, \dots, R_n$, where the rank of the smallest measurement remains unit value (i.e., 1). And the rank of the subsequent smallest becomes two and so on.
- ✓ Consider $sgn$ be the sign-function stating $sgn(x) = 1$, if $x > 0$ and $sgn(x) = -1$ if $x < 0$. It provides signed rank sum value T by using following formulation [see Eq. (9)].

$$T = \sum_{i=1}^{N} sgn\,(X_i) R_i \tag{9}$$

- ✓ Retrieve a $p - value$ by estimating $T$ to its distribution under the null-hypothesis.

Here, the ranks can be assigned in such manner that the value $R_i$ remains the number of $j$ for which $|X_j| \leq |X_i|$. Additionally, for $\sigma: \{1, \dots, n\} \rightarrow |X_{\sigma(1)}| \leq |X_{\sigma(n)}|$, then $R_{\sigma(1)} = i$ for all element $i$. With the obtained $p - value$ for each sample, the level of significance is measured and each sample was annotated as significant or insignificant. Thus, applying this approach, the significant feature elements were retained for further computation.

*3) Gini Score based significant feature selection (GSFS):* Gini Score measures the level of impurity in a data by applying a function as shown in Eq. (10).

$$m(s) = \sum_{i \neq j} \widehat{P_{si}}\, \widehat{P_{sj}} = 1 - \sum_j \widehat{P_{sj}} \tag{10}$$

This method generalizes the variance impurity, signifying the variance of a distribution belonging to the two class labels $i$ and $j$. GSFS is also referred as the expected error rate when the class label is selected randomly from the input feature distribution (here, CBOW embedding metrics). The impurity criterion applied to be highly peaked at the same likelihood in comparison to the entropy-based approaches that makes Ginni Index based feature selection suitable for our targeted PII detection and classification problem. In this work, the Eq. (11) was applied to measure the likelihood of a feature (or sample) to be retained for further computation.

$$GSFR(S) = 1 - \sum_{i=1,\dots,m} P_i^2 \tag{11}$$

In Eq. (11), $P_i$ states the likelihood that a tuple in feature set $S$ belongs to the class $C_i$. Here, we obtained the value of $P_i$ as per $|C_i, S|/|S|$.

Once performing aforesaid feature selection methods over the input embedding metrics, the retained features were horizontally concatenated to yield a composite or fused (say, ensemble) feature vector for further learning and (PII) prediction.

### *E. Resampling*

As discussed in the previous section, in real-world scenario or dataset, the frequency of PII attributes over a large text input can be smaller. In other words, the number of non-PII elements can be higher than the PII elements, signifying class-imbalance problem. The similar problem can be unavoidable in at hand PII classification problem as well. Since, in this work, the total samples comprised a massive 12000 queries, containing 4000 samples from each discipline (i.e., personal credential, health credential and financial credential). However, the number of queries or sentences (say, review) containing PII element were almost 1100, which is smaller than even 10% of the original data size, which is 12000. Though, the data combination is prepared intentionally, emulating or representing the real-time data condition where the PII elements can be of minority class than the non-PII elements (signifying majority class). This data imbalance problem can skew the learning process and hence can yield inaccurate performance. To alleviate this problem, resampling method can be applied which can make effort to retain sufficient balance between the samples belonging to the PII data as well as non-PII queries. Though, in the past, different classical approaches like up-sampling, random sampling methods have been applied; however, such approaches might often result iterative hot-spot problem (i.e., iterative class-imbalance). Noticeably, the classical up-sampling method intends to increase the minority samples to leverage the imbalance, while in down-sampling the majority class samples are reduced. On the other hand, in random sampling technique, the additional samples are appended by selecting samples from the original feature space (i.e., original minority samples). An inappropriate selection of samples might give rise to the aforesaid hot-spot or imbalance data conditions, iteratively. To address this problem, in this work three different improved resampling methods are applied. More specifically, we have applied three different variants of the synthetic minority over-sampling method (SMOTE), named SMOTE, SMOTE-

Boundary line (SMOTE-BL), and SMOTE with Edited Nearest Neighbor (SMOTE-ENN). Unlike traditional resampling techniques, SMOTE method generates synthetic samples depicting most correlated features, without impacting original sample distribution or allied (sample) ratio. Though, in exceedingly high non-linear data or feature space, SMOTE can yield better performance than the classical random sampling or up-sampling methods; however, it undergoes a scenario where there can be the multiple data instances belonging to both minority as well as majority class (it can happen due to feature relatedness or ambiguity). The classical SMOTE doesn't address this problem. Though, it has been solved by using SMOTE-BL. Considering the efficacy of SMOTE-BL method over classical SMOTE, we applied Python Imbalance Data library's SMOTE-BL method to resample the selected features (by using, Select K-Best, WRST and GSFS, distinctly).

In addition to the above discussed SMOTE and SMOTE-BL resampling method, we applied SMOTE-ENN as well that used minority samples as input to generate the synthetic samples. The generated samples were subsequently processed with k-Nearest Neighbor (k-NN) classifier, where Euclidean distance-based k-NN algorithm was taken into consideration. This approach formed a vector between the current samples and the one from obtained k-neighbors. The estimated vector was multiplied with a random number existing in between 0 to 1, which was appended to the original sample to obtain the final synthetic sample as output. Unlike classical SMOTE resampling method where defining class-boundaries can be challenging as there can be certain synthetic minority samples undergoing cross-over or overlap with the majority class. The severity of such events can be more frequent over the large non-linear features. This problem can mis-label the synthetic samples inappropriately, while training over such data might impact overall classification results or can yield false positive or false negative results. To address this problem, we applied an improved SMOTE algorithm called SMOTE-ENN. It possesses an additional computing ability in conjunction with the Edited Nearest Neighbor (ENN) classifier in which the label of each synthetic sample is compared in reference to the vote of its k-NNs neighbors. If it finds any inconsistency between the input sample and corresponding k-NNs, it drops that sample from the synthetic sample set, else it retains the same. Here, the higher $k - value$ enables stringent cleaning and therefore appends original data with the optimal set of synthetic samples to alleviate class-imbalance problem. It also provided consistent set of input features for further learning and classification.

### *F. Min-Max Normalization*

In sync with huge non-linear features, to alleviate any probable over-fitting and convergence problem, we performed feature normalization by using Min-Max Normalization method. The proposed Min-Max normalization method mapped input features in the range of [0 1]. Noticeably, Min-Max normalization was done over each feature set retained after feature selection. Mathematically, we applied equation (12) to perform Min-Max normalization over the input features. As depicted in (12), the variable $x_i$ be the feature element, where $x_i \in N$, which is mapped to the corresponding normalized value signifying $x_i'$. Here, $x_i'$ is obtained in the range of 0 to 1. In (12), the parameters $min(X)$ and $max(X)$ states the minimum and the maximum values of $X$, respectively.

$$Norm(x_i) = x_i' = \frac{x_i - min(X)}{max(X) - min(X)} \quad (12)$$

### *G. Ensemble of Ensemble Classification*

Unlike traditional ML-based solutions where the authors have applied merely single or standalone ML algorithm to perform learning and classification, we have designed a robust ensemble of ensemble (E2E) learning method that applies five different ensemble learning methods including Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier algorithms. In the developed E2E model, each classifier performs learning over the normalized features, and provides label for each sentence as 1 and 0, for PII and Non-PII element respectively. Finally, applying the concept of the maximum voting ensemble (MVE), it labels each input query as PII or Non-PII. Noticeably, unlike traditional machine learning methods like decision tree (DT), support vector machine (SVM), naïve bayes (NB), etc. the ensemble learning methods like random forest (RF), AdaBoost, extra tree classifier, XGBoost etc. have performed better for the different data and image classification problems. Being consensus driven prediction solutions, it yields higher accuracy and reliability towards PII detection and classification. Considering it as motivation, this research intends to use only ensemble learning methods, such as the Bagging, Boosting, RF, ETC, XGBoost to perform maximum voting ensemble-based learning and classification. In this method, a total of five base classifiers (here, ensemble classifiers) are considered. This is because, out of five base classifiers, if three of the base classifiers classifies an input as PII and annotates it as 1, it would be classified as PII. On the other hand, if three base classifiers annotate an input as 0, the final prediction is made as non-PII. Being consensus driven approach or maximum voting-based approach, it can be more reliable towards run-time significances. A brief of the different ensemble base classifiers used to design E2E ensemble is given as follows:

*1) AdaBoost:* It represents a kind of adaptive boosting ensemble learning method that possesses superior instance-wise analysis ability. In this work, to implement AdaBoost method, the allied prerequisite tests were doled-out to the equivalent weight that enables generation or forming certain weak learners. For each cycle of computation, it measures the error rate for the aforesaid weak classifier and eventually the weight for the accurately classified sample is improved, while the weights for the inaccurately classified samples are reduced. Finally, the weak learner becomes the strong learner and classifies each sentence as PII or PII-Free and labels them as "1" and "0", respectively. The gradient boosting method is an improved Boosting ensemble in which the weights for the accurately classified samples are varied gradually, rather with a fixed update. It solves convergence problem.

*2) Bagging:* In this work, we applied bagging ensemble learning method with two different kernels. More specifically, we applied k-NN algorithm and MNB classifier as the two base classifiers to design Bagging ensemble learning algorithms.

*3) Random Forest (RF*): It is one of the most efficient ensemble-learning methods that embodies multiple tree-based classifiers. Being a tree-based learning model, each tree generates its own predicted output or vote for the most probable class (for each sentence or allied embedding feature vector). Consider that the input training samples be $N$, then a sample possessing $N$ cases are selected arbitrarily from the input data (say, input feature vector). The selected samples are subsequently used as training set to constitute a new tree. Let, there be the $M$ input variables, then the best split is applied over $m$ to split the node. Noticeably, in this work the value of $m$ was fixed as constant while performing forest development. Thus, the proposed method develops each tree to the possible largest level or size. Noticeably, unlike other machine learning algorithms, RF requires relatively smaller number of parameters to be calculated during learning and classification and hence serves a lightweight and computationally efficient classifier (solution). In sync with above discussed forest growth process, the RF algorithm can be defined as the amalgamation of the different tree-structures, given as Eq. (13).

$$\{h(x, \theta_k), k = 1,2, \dots i \dots\} \quad (13)$$

In Eq. (13), $h$ represents the classifier function, while the arbitrary vector generated unformal across tress is given by $\{\theta_k\}$. Here, each tree embodies a unit vote for its most probable class towards a unit query or sentence (say, input $x$). To be noted, the ability to employ multiple DTs where each DT performs respective classification makes proposed RF method more suitable for learning and classification. We applied a bootstrapped subset of training samples to train each tree across the formed forest that uses 70% of the training data, while the remaining data elements are labeled as the out-of-bag samples, which are subsequently applied for inner cross-validation to predict output. Thus, applying this method, we classified each input query or sentence (say, corresponding embedded feature vector) as PII or PII-Free, which were subsequently labelled as "1" or "0", correspondingly.

*4) Extra Tree Classifier (ETC):* This algorithm forms a cluster of unpruned DTs as per the conventional top-down paradigm. Unlike RF method, ETC algorithm comprises randomization of attribute and cut-point selection when performing node-split. It is also able to form overall randomized trees containing structures which are independent of the outputs of the training sample. In fact, it distinguishes itself from the other tree-based ensemble algorithms because of the two key factors. The first is that it splits nodes by choosing cut-points arbitrarily and the second it applies the complete training sample to for forest (say, tree growth). The predicted output from the encompassing trees is amalgamated to yield final PII prediction result by using MVE method. The key approach behind the ETC method is that the complete randomization of the cut-point and attribute altogether with ensemble averaging that minimizes the variance in more efficient way than the weaker randomization methods applied in other machine learning approaches. Additionally, the use of the original training samples minimizes the probability of bias and therefore accomplishes higher accuracy for classification. Applying this ensemble approach each sentence was classified as PII and PII-Free, which was labelled as "1" and "0", respectively.

Once obtaining labelled output from each base (ensemble) classifier, for each query or sentence, the proposed MVE method estimated the consensus, signifying the maximum vote for each sentence (i.e., PII or Non-PII). As the proposed model applied five different ensemble ML classifier as base classifier, a query with minimum three 1 was classified as PII and the one with three 0 was annotated as the non-PII query. In this manner, the proposed model applied consensus (say, Maximum Voting) score to predict each query as PII or non-PII.

The use of this approach can help identifying a set of those queries having PII elements, which can not only enable further PII element detection and identification, but can also make computation fast due to reduced search space (or queries). Once identifying a query as PII, dictionary-based technique(s) can be applied to perform specific PII attribute detection from each sentence or query. Though, this research considers it a scope for future efforts, and hence is beyond the research scope. Simulation results and allied inferences are given in the subsequent section.

## VI. Results and Discussion

In this work, the key emphasis is made on developing a robust cross-discipline PII detection and classification system for privacy preserved digital data transmission and storage. In this work, the PII detection and classification task is solved as an NLP problem, where it exploits semantic features from the input queries to learn and predict PII elements in each query. The overall proposed model emphasizes on addressing almost major challenges in at hand PII detection problem. In other words, we made effort to address both data challenges (feature engineering) as well as classification so as to ensure a reliable and optimized solution. In this reference, this work contributed a robust ensemble of ensemble learning assisted semantic feature driven cross-discipline PII detection and de-identification model (EESD-PII). Being a machine learning driven NLP solution, we have tried to exploit PII attributes from the different disciplines, which provide sufficiently large feature space to learn and hence predict PII queries in real-world application. Moreover, realizing class-imbalance, convergence, local minim and over-fitting problems, different computational optimization measures were taken into consideration that ensured optimal data (or feature) as well as computational component to achieve a robust cross-discipline PII detection and classification solution. In sync with the real-world data conditions where the likelihood of data heterogeneity, unstructured-ness ss etc. can' be ruled out, we at first performed pre-processing over the input cross-discipline datasets. Noticeably, we synthesized a total of 12000 text queries containing PII attributes from three different categories including personal credentials ("Name", "Address", "Email", "Aadhar Number", "PAN", "SSN", "Vehicle Number", and "Phone Number"), healthcare credentials ("Health Insurance Number", "Diagnosis Details", and "Life Insurance Number") and financial credentials ("Bank Account Number", "Loan Account", "Balance Details", "OTP", "Access PIN" and "Credit

Card Number"). Here, the key motive was to strengthen the proposed PII detection model to detect and classify each input text query irrespective of its subject matter or allied discipline. To address aforesaid problem of data heterogeneity and unstructured-ness different pre-processing tasks were performed including missing value removal, Unicode normalization, removal or emoji, website link, punctuations, stop-words, and lower-case conversion. Additionally, lemmatization and tokenization were performed, where the earlier one helped in retaining the root-intends or contextual information, while the use of tokenization helped tokenizing each query to exploit corresponding semantic features for learning and prediction. The tokenized outputs were passed to the CBOW feature extraction that obtained semantically enriched feature vector to perform learning and prediction. Noticeably, we applied CBOW as Word2Vec embedding method to retain maximum possible contextual feature even in low-dimensional feature which helped retaining sufficiently large contextual feature vector even with the low computational cost. Subsequently, CBOW feature was processed for feature selection, where the different algorithms including Chi-Square based Select K-Best (SKB) method, WRST algorithm and Ginni Score based feature selection (GSFS) methods. Unlike traditional feature selection methods, we designed an ensemble feature model by concatenating the selected features from the SKB, WRST and GSFS methods. Thus, the combined feature vector was later applied for feature resampling. In this work, we applied three different resampling methods including SMOTE, SMOTE-BL and SMOTE-ENN. Noticeably, in the proposed model, we executed these three resampling methods individually over the selected features. It helped addressing class-imbalance problem. Subsequently, the resampled features were then processed for Min-Max normalization that mapped each input in the range of 0 to 1 and thus alleviated any likelihood of convergence and over-fitting problem. The normalized outputs were passed to the E2E classifier which was designed by using five different ensemble learning methods named Bagging, Boosting, AdaBoost, Random Forest, and Extra Tree Classifier. We applied consensus oriented MVE method to perform learning and classification.

The proposed model was developed using Python Notebook, and the simulation was done on Google Co-laboratory platform, which helped reducing implementation complexity and cost. The simulation was done on a central processing unit configured with Microsoft Window operating system operating with 8 GB RAM and Intel i5 processor. To quantify the performance of the proposed PII detection and classification system, we retrieved confusion metrics in terms of true positive (TP), true negative (TN), false positive (FP) and false negative (FN). These parameters were used to measure performance in terms of accuracy, precision, recall and F-Measure, which were obtained by using following equations (Table II).

TABLE II. PERFORMANCE PARAMETERS

| Parameter | Mathematical Expression |
|---|---|
| Accuracy | $\frac{(TN+TP)}{(TN+FN+FP+TP)}$ |
| Precision | $\frac{TP}{(TP+FP)}$ |
| Recall | $\frac{TP}{(TP+FN)}$ |
| F-measure | $2.\frac{Recall.Precision}{Recall+Precision}$ |

Observing the implementation schematic (Fig. 1), it can be found that the overall proposed model encompasses the different feature selection, resampling and classification models; though, the proposed model defines the term "proposed" with the CBOW feature extraction, ensemble feature selection, resampling, Min-Max normalization and E2E MVE-based classification. It signifies that the proposed model encompasses sub-algorithms for the different phases, including feature selection, resampling and classification. Considering this fact, we at first examined the performance with the different feature selection model, feature resampling and classifiers. Subsequently, with the best performing model the comparison was done with the existing state-of-arts. In this reference, the overall performance characterization is done in terms of intra-model characterization and inter-model characterization. The detailed discussion of the results and allied inferences is given in the subsequent sections.

### *A. Intra-Model Characterization*

In this section, the performance comparison with the different feature selection methods, feature resampling and classifiers is given. Noticeably, in this work we applied single Word2Vec CBOW feature extraction method, which is subsequently processed for the different feature selection, feature resampling, and ensemble classifiers. In this reference, this research performs relative performance with the different algorithms. The key purpose of this assessment is to identify the best performing set of algorithms towards targeted PII attribute detection and classification solution. The detailed discussion of the results obtained and allied inferences is given as follows:

In the proposed model, the CBOW features are fed as input to the three different feature selection methods, including SKB, WRST and GSFS method. Table III presents the simulated performance outcomes with the different feature selection methods. The outputs obtained reveal that with the CBOW input feature, SKB feature selection model exhibits the accuracy of 94.91%, while WRST and GSFS method yielded PII prediction accuracy of 95.03% and 94.89%, respectively. Similarly, the F-Measure outputs were measured as 94.93%, 95.08% and 95.10%, for SKB, WRST and GSFS feature selection methods, correspondingly. The simulation results signify that the intrinsic features driven approaches whether WRST and GSFS methods yield better performance towards PII prediction. The precision performance by SKB, WRST and GSFS methods were obtained as 94.86%, 94.97% and 95.21%. It clearly indicates that the WRST method or GSFS can be suitable towards at hand text-review driven PII prediction solution. Noticeably, these results (Table III to Table V) have been obtained by using proposed E2E MVE ensemble learning classifier.

TABLE III. PERFORMANCE WITH THE DIFFERENT FEATURE SELECTION METHODS

| Feature | Feature Selection | Performance (%) | | | |
|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F-Measure |
| CBOW | SKB | 94.91 | 94.86 | 95.01 | 94.93 |
| | WRST | 95.03 | 94.97 | 94.82 | 95.08 |
| | GSFS | 94.89 | 95.21 | 95.00 | 95.10 |

Table IV presents the efficiency with the different resampling methods. More specifically, three different resampling methods were applied towards targeted PII detection and prediction model. The simulation results reveals that the CBOW features with SMOTE, SMOTE-BL and SMOTE-ENN methods exhibited the accuracy of 97.77%, 97.93% and 99.35%, respectively. The simulation results also reveal that the SMOTE, SMOTE-BL and SMOTE-ENN methods achieves precision of 99.63%, 99.81% and 99.88%, respectively. These algorithms exhibit the recall of 99.37%, 99.63% and 99.94%, correspondingly. The depth assessment also exhibits that the F-Measure performance by SMOTE, SMOTE-BL and SMOTE-ENN resampling method exhibited the F-Measure of 99.49%, 99.71% and 99.90%, respectively. The simulation clearly depicts that the proposed SMOTE-ENN method exhibits the superior performance than the other approaches. Recalling the theoretical aspects where SMOTE-ENN methods possess superior efficacy than the classical random sampling, up-sampling and even classical SMOTE variants. The results obtained confirm efficacy of the proposed SMOTE-ENN model towards at hand PII detection and prediction model.

TABLE IV. PERFORMANCE WITH THE DIFFERENT FEATURE RESAMPLING METHODS

| Feature | Feature Resampling | Performance (%) | | | |
|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F-Measure |
| CBOW | SMOTE | 97.77 | 99.63 | 99.37 | 99.49 |
| | SMOTE-BL | 97.93 | 99.81 | 99.63 | 99.71 |
| | SMOTE-ENN | 99.35 | 99.88 | 99.94 | 99.90 |

The simulation results with the different ensemble learning classifiers including the proposed E2E MVE ensemble model are given in Table V. Here, the key motive is to assess relative efficacy of the different base classifiers and assess whether the proposed E2E ensemble with MVE concept can yield superior performance or not. In this reference, the accuracy, precision, recall and F-Measure performance obtained are given in Table V. As depicted in Table V, Bagging ensemble method exhibits the accuracy of 97.89%, while Boosting, AdaBoost, Random Forest and Extra Tree Classifier algorithms exhibit the accuracy of 96.68%, 96.72%, 98.67% and 98.81%, respectively. On the contrary, the proposed E2E MVE ensemble method which applied aforesaid ensemble learning methods as base classifier achieved the PII prediction of 99.77%. The precision performance by Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier algorithms was measured as 94.37%, 96.61%, 94.99%, 94.99% and 96.01%, respectively. The precision performance by the proposed E2E MVE method exhibited the recall of 99.63%, which is higher than other standalone ensemble learning classifiers. The F-Measure performance by Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier algorithms were obtained as 95.45%, 96.69%, 95.88%, and 96.065, respectively. The proposed E2E MVE method exhibited the F-Measure of 99.71%, which is superior than any other approaches.

TABLE V. PERFORMANCE WITH THE DIFFERENT CLASSIFIERS

| Feature | Feature Resampling | Performance (%) | | | |
|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F-Measure |
| CBOW | Bagging | 97.89 | 94.37 | 96.57 | 95.45 |
| | Boosting | 96.68 | 96.61 | 96.69 | 96.69 |
| | AdaBoost | 96.72 | 94.99 | 96.80 | 95.88 |
| | Random Forest | 98.67 | 94.99 | 96.80 | 95.88 |
| | Extra Tree Classifier | 98.81 | 96.01 | 96.13 | 96.06 |
| | **Proposed E2E** | **99.77** | **99.81** | **99.63** | **99.71** |

Taking into consideration of the overall results and allied inferences, it can be quantified that the proposed model encompassing CBOW semantic features, WRST feature selection, SMOTE-ENN resampling, and E2E MVE ensemble method exhibits the accuracy of 99.77%, precision 99.81%, recall 99.63% and F-Measure of 99.71%. Here onwards we call this method as the proposed model and thus with this specific performance, we have performed the relative performance comparison with other state of arts, which is given in the subsequent sections.

### *B. Inter-Model Characterization*

In sync with the results obtained and allied inferences, we define proposed model as the PII detection solution encompassing CBOW feature extraction, WRST feature selection, SMOTE-ENN resampling, Min-Max normalization and E2E MVE ensemble classifier. Here onwards, we define accuracy, precision, recall and F-Measure performance of the proposed model as 99.77%, 99.81%, 99.63% and 99.71%, respectively. This section discusses the relative performance with the other state-of-arts, as given in Table VI.

This is the matter of fact that in the very few efforts have been made for PII detection, where the different algorithms have been applied for feature extraction, and classification (say, learning). A few approaches have used deep learning as well to learn over the local deep features pertaining to each PII related query to perform (PII) detection and classification. However, their inability to address class-imbalance problems, convergence and local minima etc. make then inferior towards a robust and optimal PII detection and classification solution. Moreover, the different existing approaches have applied the different datasets

and hence generalizing solutions for the same data becomes difficult, especially when the data availability is a challenge. Despite this we have compared the efficacy of our proposed model with other state of arts. Noticeably, in sync with the above discussions (i.e., Inter-Model Assessment), we have identified CBOW feature extraction followed by CCRA feature selection, SMOTE-ENN resampling, Min-Max normalization and ETC classifier as the optimal solution towards the targeted PII detection and classification system. And therefore, respective performance (say, accuracy 99.77%, precision 99.81%, recall 99.63% and F-Measure of 99.71%) is considered as the proposed method towards targeted PII detection and classification system. In this reference, a comparison between the proposed method and other state-of-arts is given in Table II. Noticeably, in the past the different state-of-arts have applied the different performance variables for efficacy analysis, we considered a common performance variable F-Measure to assess relative efficiency. Furthermore, the considered state-of-arts represent the different approaches towards NER (Named Entity Recognition) analysis that possess similar intends as that of PII detection.

TABLE VI. INTER-MODEL PERFORMANCE COMPARISON

| Reference | Method | F-Measure (%) |
|---|---|---|
| [51] | Bi-LSTM-CRF +CNN | 93.5 |
| [52] | Bi-LSTM-CRF+ELM (Extreme Learning Machine) + BERT+ Flair | 93.38 |
| [53] | Bi-LSTM + Flair embeddings | 93.09 |
| [54] | BERT+ BI-LSTM | 92.80 |
| | BERT Base | 92.4 |
| [55] | Bi-LSTM + multi-task | 92.61 |
| [56] | Bi-LSTM + BERT + ANN-LM | 92.22 |
| [57] | Word-Embedding + ANN-LM | 91.93 |
| [58] | Bi-LSTM + CRF +Auto-encoders+ Lexical Features | 91.89 |
| [59] | Bi-LSTM + CRF + Lexical Features | 91.73 |
| [60] | LM- LSTM +CRF | 91.71 |
| [61] | Bi-LSTM+ CRF | 91.62 |
| [62] | Hybrid Semi-Markov +CRF | 91.38 |
| [63] | IXA-Pipes | 91.36 |
| [64] | CCNN+ WLSTM+ CRF | 91.35 |
| [65] | CRF +GRU | 91.26 |
| [66] | BI-LSTM +CRF | 90.94 |
| [67] | Glove | 88.30 |
| [68] | CBOW | 88.20 |
| [69] | Bi-LSTM + Glove + Flair | 85.51 |
| **Proposed** | **CBOW + WRST + SMOTE-ENN + Min-Max Normalization + E2E MVE** | **99.71** |

In sync with the results obtained (Table VI), it can be found that the major at hand efforts towards NER or applied PII detection methods have exploited deep learning methods with the different feature embedding techniques such as CBOW [68], Glove [67], Flair [69]. Though, the major state-of-arts have directly applied varied deep learning methods such as LSTM, Bi-LSTM, CNN etc., to perform two-class classification by using Softmax classifier. However, most of these methods fail in addressing contextual details within the input text corpus. Moreover, none of these approaches could address inherent challenges of NLP including unstructured data, class-imbalance, local minima and convergence. Such limitations might make any solution suspicious, especially when the solution is supposed to be applied over real-world BigData. Unlike aforesaid approaches, in this work, we emphasized on addressing above stated issues so as to contribute a generalizable optimal PII detection and classification system. Moreover, a few approaches like [56][57] tried to extract local features from the sequential text embeddings for further learning and classification by using machine learning methods (i.e., ANN-LM); however, the highest F-Measure observed was 92.2%, which is almost 7% lower than the proposed method. Here, the impact of feature engineering such as class-imbalance (SMOTE-ENN), feature selection (WRST) and normalization can't be ruled out. Noticeably, these feature improvement methods helped our proposed model to achieve features with minimum redundant data or non-redundant feature elements. It helped in alleviating pre-mature convergence. SMOTE-ENN helped in addressing class-imbalance problem, while the use of Min-Max normalization resolved over-fitting problem. Thus, the improvement in data or allied feature enabled our proposed method to exhibit superior over other state-of-arts. The overall results confirm superiority of the proposed method over other state-of-arts available in PII detection domain. The depth assessment affirms positive response or answers for the research questions (RQ1-RQ4), as discussed in Section IV.

## VII. CONCLUSION

Unlike existing PII detection approaches where the authors have either applied predefined dictionaries to detect and classify (PII) text for a specific dataset or discipline, in this work a robust multi-disciplinary PII detection method was developed. To achieve it, the PII detection, de-identification and re-identification tasks were solved as an NLP problem. In this reference, unlike predefined dictionary or syntactical learning-based solutions, the texts encompassing PII and normal queries related to the different disciplines including personal credential, healthcare, business communication details etc. were processed for NLP processing that eventually enabled a cross-discipline PII detection system. Additionally, this research made effort to alleviate at hand computational limitations including class-imbalance, local minima and convergence and over-fitting. More specifically, novel and robust ensemble of ensemble learning assisted semantic feature driven cross-discipline PII detection and de-identification model (EESD-PII) was developed in this work. In sync with cross-discipline PII detection and classification task, a large set of text queries

possessing PII elements belonging to the personal credentials, healthcare data, financial texts and queries etc. were used for training. The collected cross-discipline texts were processed for pre-processing tasks like the removal of stopping-words, punctuations, URL-link, lower case conversion, lemmatization and tokenization. The tokenized text-sequences were processed for continuous bag-of-word (CBOW) word-embedding that provided semantic feature space for further learning. The use of CBOW embedding provided contextually-rich feature vector to ensure better learning and PII detection (and/or classification). Realizing real-world scenarios where the frequency of PII terms is very small in comparison to the non-PII elements (signifying severe class-imbalance problem), SMOTE resampling methods including SMOTE, SMOTE-BL and SMOTE-ENN were applied. This approach helped in alleviating the class-imbalance problem and hence curse of dimensionality, thus helping the model to achieve higher learning accuracy. Subsequently, Wilcoxon Rank Sum Test (WRST) method was applied to retain the most representative samples in reference to the 95% confidence interval. The retained features were processed for Min-Max Normalization which helped alleviating any over-fitting and convergence problems. Finally, a robust ensemble of ensemble (E2E) learning classifiers was designed by using five different ML algorithms encompassing Bagging, Boosting, AdaBoost, Random Forest and Extra Tree Classifier as base classifier. This approach not only helped in achieving more reliable outcome but also alleviated performance diversity problem. Noticeably, the use of tokenization and word-embedding helped achieving de-identification goals as well, without losing data essence for further learning and classification. The proposed E2E model performed majority voting ensemble to annotate each text-query as PII or Non-PII data. The simulation results reveal that the proposed EESD-PII model achieves PII annotation accuracy of 99.77%, precision 99.81%, recall 99.63% and F-Measure of 99.71%. In future, the authors can design AI driven multi-disciplinary dictionary with PII-sensitive masking approach to improve de-identification; though, the proposed model retains aspects of de-identification and re-identification to serve real-time decisions.

## VIII. Future Scope

This is the matter of fact that the proposed method has yielded superior and generalizable performance towards PII detection and classification; however, it was mainly based on exploiting semantic features trained over ensemble machine learning classifiers. The use of deep features can also be explored. The recent development in advanced deep structures, especially designed to address long-term dependency problems like multi-attention-based Bi-LSTM, Bi-GRU, residual networks etc. can be applied as ensemble feature structure. It can help to exploit more contextual details with no probability of any gradient vanishing or gradient explode over large deep structure running over non-linear data space. In this reference, the researchers can focus on applying deep (hybrid) feature models for feature extraction and learning to achieve better outputs. The use of GAN can also be considered in future to improve de-identification and word-restructuring.

## References

[1] A. Acquisti, L. Brandimarte, G. Loewenstein, “Privacy and human behavior in the age of information”, Science 347(6221), 2015, pp. 509-514.

[2] M. Callahan, “Us dhs handbook for safeguarding sensitive personally identifiable information. Washington, DC, 2012.

[3] Personally Identifiable Information (PII) Guidebook, Personally Identifiable Information Working Group of the Indiana Executive Council on Cybersecurity, January, 2021, pp. 1-33. [Accessed on 27 April 2022]

[4] Y. Pan, B. Stackpole, L. Troell, “Computer forensics technologies for personally identifiable information detection and audits”, ISACA, vol. 02, 2010, pp. 1-7. [Available http://scholarworks.rit.edu/article/999]

[5] P. Kulkarni and N. K. Kauvery, “Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique”, (IJACSA) International Journal of Advanced Comp. Science and Applications, Vol. 12, No. 9, 2021.

[6] X. Feng, Y. Feng and A. Asante, “A Systematic Approach of Impact of GDPR in PII and Privacy”, International Journal of Engineering Science Invention, 10 (1), pp. 05-14.

[7] ICO (2016) “Overview of the General Data Protection Regulation (GDPR)”. https://ico.org.uk/for-organisations/data-protectionreform/overview-of-the-gdpr/ [Accessed 14/3/2021].

[8] A. K. Makhija, “Deep Learning Application – Identifying PII (Personally Identifiable Information) to Protect”, Journal of Accounting, Finance, Economics, and Social Sciences Vol.5, No.3, 2020, pp.49-55.

[9] C. Cadwalladr and E. Graham-Harrison, “Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach", The guardian, 2018, pp. 17-22.

[10] Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of “personally identifiable information.” Communications of the ACM, 53(6), 24.

[11] Al-Zaben, N., Hassan Onik, M. M., Yang, J., Lee, N.- Y., & Kim, C.-S. (2018). General Data Protection Regulation Complied Blockchain Architecture for Personally Identifiable Information Management. 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE). doi:10.1109/iccecome.2018.8658586.

[12] B. T. Welderufael, S. Jetzabel, and P. Sebastian, “Challenges in Detecting Privacy Revealing Information in Unstructured Text”, pp. 1-8. [Accessed on 23 April 2023].

[13] H. C. Kum, S. Ahalt, “Privacy by design: understanding data access models for secondary data. American Medical Informatics Association (AMIA) Joint Summits on Translation Science and Clinical Research Informatics (2013).

[14] Shah, R., Valera, M.: Survey of sensitive information detection techniques: The need and usefulness of machine learning techniques [Accessed on 27 April 2023].

[15] C. Posey, U. Raja, R. E. Crossler, and A. J. Burns, “Taking stock of organisations’ protection of privacy: categorising and assessing threats to personally identifiable information in the USA”. European Journal of Information Systems, 2017, 26(6), 585–604.

[16] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, “EU General Data Protection Regulation: Changes and implications for personal data collecting companies”, Computer Law & Security Review, 2017, 34(1), pp. 134–153.

[17] A. C. Islam, J. Walsh, R. Greenstadt, “Privacy detective: Detecting private information and collective privacy behaviour in a large social network. In: Proceedings of the 13th Workshop on Privacy in the Electronic Society, 2014. [Accessed on 27 April 2022]

[18] L. Xiao-Bai and Q., “Anonymizing and sharing medical text records”, Information Systems Research, 2017, Vol. 28(2), pp. 332–352.

[19] C. A. Kushida, D. A Nichols, R. Jadrnicek, R. Miller, J. K. Walsh, and K. Griffin, Strategies for de-identification and anonymization of electronic health record data for use in multi-center research studies”, Medical care, 2012, Vol. 50(Suppl): S82.

[20] T. Aura, T. A. Kuhn, and M. Roe, “Scanning electronic documents for personally identifiable information. In Proceedings of the 5th ACM workshop on Privacy in electronic society, 2006, pp. 41–50.

[21] P. Ongsulee, “Artificial intelligence, machine learning and deep learning”, 2017 15th International Conference on ICT and Knowledge Engineering, 2007. doi:10.1109/ictke.2017.8259629.

[22] Arttu Oksanen, J Tuominen, E Mäkelä, M Tamper, Aki Hietanen, and Eero Hyvönen. 2019. Semantic finlex: Transforming, publishing, and using finnish legislation and case law as linked open data on the web. Knowledge of the Law in the Big Data Age, 317:212–228.

[23] Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2360–2369. Association for Computational Linguistics.

[24] Fadi Hassan, Josep Domingo-Ferrer, and Jordi Soria-Comas. 2018. Anonymization of unstructured data via named-entity recognition. In International conference on modelling decisions for artificial intelligence, pages 296–305. Springer.

[25] Filip Gralinski, Krzysztof Jassem, Michał Marci´nczuk, and Paweł Wawrzyniak. 2009. Named entity recognition in machine anonymization. Recent Advances in Intelligent Information Systems, pages 247–260.

[26] Gang Luo, Xiaojing Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint named entity recognition and disambiguation. In Proceedings of the 2015 Conf. on Empirical Methods in Natural Language Proc., page 879–888, USA. Association for Computational Linguistics.

[27] Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In Proceedings of the Eighteenth Conference on Comp. Natural Language Learning, pages 78–86.

[28] Lev Retinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), page 147–155, USA. Association for Computational Linguistics.

[29] Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. In Transactions of the Association for Computational Linguistics, volume 4, pages 357–370. MIT Press.

[30] Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using BERT-Bi LSTM-CRF for Chinese electronic health records. In 2019 12th international congress on image and signal processing, biomedical engineering and informatics, pages 1–5. IEEE.

[31] Kathleen C Fraser, Isar Nejadgholi, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. 2019. Extracting umls concepts from medical text using general and domain-specific deep learning models. arXiv preprint arXiv:1910.01274.

[32] Xin, Y. et al. (2018). Machine learning and deep learning methods for cybersecurity. IEEE Access, 6, 35365-35381.

[33] Tesfay, W. B., Serna, J., & Rannenberg, K. (2019). PrivacyBot: Detecting Privacy Sensitive Information in Unstructured Texts. 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). doi:10.1109/snams.2019.8931855

[34] Jinlan Fu, Pengfei Liu, and Qi Zhang. 2020. Rethinking generalization of neural models: A named entity recognition case study. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7732–7739.

[35] Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. 2018 10th International Conference on Cyber Conflict (CyCon).

[36] Tesfay, W. B., Serna, J., & Rannenberg, K. (2019). PrivacyBot: Detecting Privacy Sensitive Information in Unstructured Texts. 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). doi:10.1109/snams.2019.8931855.

[37] Wang, Z., Quercia, D., S_eaghdha, D.O.: Reading tweeting minds: Real-time analysis of short text for computational social science. In: Proceedings of the 24th ACM Conference on Hypertext and social media (2013)

[38] Mao, H., Shuai, X., Kapadia, A.: Loose tweets: An analysis of privacy leaks on twitter. In: Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society (2011)

[39] Jindal, P., Gunter, C.A., Roth, D.: Detecting privacy-sensitive events in medical text. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14 (2014)

[40] Gill, A.J., Vasalou, A., Papoutsi, C., Joinson, A.N.: Privacy dictionary: a linguistic taxonomy of privacy for content analysis. In: Proceedings of the SIGCHI conf. on human factors in computing sys. 2011, pp. 3227-3236.

[41] Zhang, N.J., Todd, C.: A privacy agent in context-aware ubiquitous computing environments. In: IFIP International Conference on Communications and Multimedia Security. Springer (2006)

[42] Pape, S., Serna-Olvera, J., Tesfay, W.: Why open data may threaten your privacy. In: Workshop on Privacy and Inference, co-located with KI (September 2015)

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

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

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

[51] Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Clozedriven pretraining of self-attention networks. In 2019 Conference on Empirical Methods in Natural Language Processing, 2019.

[52] Jana Straková, Milan Straka, and Jan Hajic. Neural architectures for nested NER through linearization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5326–5331, Florence, Italy, July 2019. Association for Computational Linguistics.

[53] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In COLING 2018: 27th International Conference on Computational Linguistics, pages 1638–1649, 2018.

[54] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 4171–4186, 2019.

[55] Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. Semi-supervised sequence modeling with cross-view training. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1914–1925, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[56] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[57] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi supervised sequence tagging with bidirectional language models. CoRR, abs/1705.00108, 2017.

[58] Minghao Wu, Fei Liu, and Trevor Cohn. Evaluating the utility of hand-crafted features in sequence labelling. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2850–2856, Brussels, Belgium, Oct-Nov 2018. Association for Comp. Linguistics.

[59] Abbas Ghaddar and Philippe Langlais. Robust lexical features for improved neural network named-entity recognition. CoRR, abs/1806.03489, 2018.

[60] Liyuan Liu, Jingbo Shang, Frank F. Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. CoRR, abs/1709.04109, 2017.

[61] Jason P. C. Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. CoRR, abs/1511.08308, 2015.

[62] Zhixiu Ye and Zhen-Hua Ling. Hybrid semi-Markov CRF for neural sequence labeling. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 2), pages 235–240, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[63] Rodrigo Agerri and German Rigau. Robust multilingual named entity recognition with shallow semi-supervised features. Artificial Intelligence, 238:63 – 82, 2016.

[64] Jie Yang and Yue Zhang. NCRF++: An open-source neural sequence labeling toolkit. In Proceedings of ACL 2018, System Demonstrations, pages 74–79, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[65] Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks. CoRR, abs/1703.06345, 2017.

[66] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. CoRR, abs/1603.01360, 2016.

[67] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014.

[68] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013.

[69] Carlos Jorge Augusto Pereira da Silva, “Detecting and Protecting Personally Identifiable Information through Machine Learning Techniques”, Faculdade De Engenharia Da Universidade Do Porto, July 27, 2020.

[70] L. Fei, Y. Kang, S. Park, Y. Jang, J. Lee and H. Kim, "KDPII: A New Korean Dialogic Dataset for the Deidentification of Personally Identifiable Information," in IEEE Access, vol. 12, pp. 135626-135641, 2024.

[71] M. Gong, J. Liu, H. Li, Y. Xie and Z. Tang, "Disentangled Representation Learning for Multiple Attributes Preserving Face Deidentification," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 1, pp. 244-256, Jan. 2022.

[72] H. Proença, "The UU-Net: Reversible Face De-Identification for Visual Surveillance Video Footage," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 2, pp. 496-509, Feb. 2022.

[73] L. Laishram, J. T. Lee and S. K. Jung, "Face De-Identification Using Face Caricature," in IEEE Access, vol. 12, pp. 19344-19354, 2024.

[74] J. Liu, Z. Zhao, P. Li, G. Min and H. Li, "Enhanced Embedded AutoEncoders: An Attribute-Preserving Face De-Identification Framework," in IEEE Internet of Things Journal, vol. 10, no. 11, pp. 9438-9452, 1 June1, 2023.

[75] B. Shickel, P. J. Tighe, A. Bihorac and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," in IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589-1604, Sept. 2018.