ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

NLP-Based De-Identification Techniques for Patient Data Anonymization

Veerendra Nath Jasthi

USA

ABSTRACT

Electronic health records (EHR) Patient data in the form of electronic health records are sensitive and personal so there are legal structures to protect such things such as HIPAA. De-identification of such data is enough to guarantee the privacy of such information, allowing it to be utilized in medical studies and the creation of AI models. Natural Language Processing (NLP) has become an effective method of automating the de-identification of unstructured clinical narratives. This paper discusses the different NLP-based de-identification techniques, rule-based, machine learning models, and deep learning approaches. These approaches are compared, and the hybrid model will be created wherein Named Entity Recognition (NER) will be combined with BERT-based contextual models. Precision, recall, and F1-score are assessment measures applied to benchmark datasets. Findings show that hybrid NLP techniques are more generally accurate and generalized. The study helps in enhancing privacy of data in healthcare as the study allows useful anonymization of textual records of patients.

***Corresponding author**

Veerendra Nath Jasthi, USA.

Received: June 06, 2023; **Accepted:** June 10, 2023; **Published:** June 20, 2023

Keywords: De-identification, Patient Anonymization, Natural Language Processing (NLP), Electronic Health Records (EHR), Named Entity Recognition (NER), Deep Learning, Data Privacy, BERT, HIPAA.

Introduction

Over the past few years, digitization of healthcare record-keeping and its development into electronic health records (EHR) systems have completely transformed how patients' records are stored, accessed, and shared [1]. Such data-rich structured and unstructured records constitute the spine of the current medical research and decision-making procedures. Nevertheless, sensitive personally identifiable information (PII) is commonly stored in EHRs and includes names of patients, their addresses, contact numbers, date of birth, and identification numbers. In turn, no patient data can be used again in research or artificial intelligence (AI) model development before de-identification, which is the practice of removing or obscuring personal identifiers.

Even though structured data can be anonymized frequently by the application of database-level transformations and field masking, unstructured clinical notes are a much more complicated issue. These stories have long form; free text written by health care providers and have references of contextually specific mention of personal information in natural language. Such data could be manually redacted, but this job is time-consuming, inaccurate, and inconsistent. Consequently, automated solutions that rely on Natural Language Processing (NLP) have received massive interest as solutions to effectively and correctly identify and replace or mask the sensitive information contained in clinical narratives [2].

The development of NLP methods is related with the shift in the drift to simple rule-based systems and keyword matching to more advanced ones capable of learning and understanding the patterns of language [3]. The initial NLP-based solutions were based on individually crafted rules and regular expressions to work with special data sets, and thus they lacked their flexibilities and scalable features.

As deep learning has proceeded, specifically, the transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) have been introduced, the contextual answers to the language have been broken through [4]. These models are able to expand on complicated linguistic characteristics without manual intervention, and have demonstrated outstanding performance across numerous NLP activities reaching those of denoted entity identification (NER), sentiment evaluation, and text grouping. Transformer-based models can be used as applied to de-identification to detect sensitive information even in peculiar formats or non-standard wording, increasing recall and precision.

Nevertheless, even the state-of-the-Art models struggle with some of the ambiguities related to a particular domain, typos, and infrequent types of entities. Thus, hybrid variants of solutions that present the advantages of rule-based mechanisms and the semantics of deep learning models are increasingly becoming popular. The goal of such systems is to combine high-precision of rule-based filters and the generalization power of deep learning resulting in a more industry-resistant and thorough answer to de-identification problems [5].

The basic rationale behind this research is the necessity of safe and precise de-identification methodology that can be applied to real-world clinical data, with the condition of meeting the

privacy regulations and the consideration of the ability to use the data in the secondary cases. An effective, high-performance de-identification framework system would make enormous volumes of clinical data available to research but without eroding patient trust or violating ethical integrity.

The growing role of data sharing in the process of spurring the pace of healthcare transformation, be it predictive diagnostics, population health research, or the like, has made the issue of particularly relevant anonymization methods, and the study of the latter through the lens of NLP, in particular, has never been more pertinent [6]. By means of this research, we consider the environment of NLP-based de-identification techniques, analyze the performance of available models, and suggest a hybrid infrastructure, which aims at improving the accuracy, generalizability and practical applicability of data de-identification solutions in the real-world medical settings.

Novelty and Contribution

The uniqueness of such study herby is the design and assessment of a hybrid NLP-based de-identification framework that incorporates deep learning (BERT-based Named Entity Recognition) with classic rule-based approach to address the weaknesses of both approaches in pairs. In contrast to purely deterministic systems which are not flexible in their model and purely data-driven systems which at times fail to realize less common or ambiguously revealed entities in any case, the proposed method would be able to combine contextual drives as well as deterministic patterns. This synergy enables the system to be more predictable when dealing with different EHR data, including those with inconsistencies, misspellings, or unorthodox wordings [7].

The other important innovation is the process of tuning transformer models (in particular, domain-specific ones, such as Clinical BERT and Bio BERT) on annotated medical data, which makes them more appropriate to the clinical setting. We also provide a high-performance post-processing filter enabling removing of false positives and strengthening stricter de-identification rules improving the reliability of the model in deployment environment.

This Research Makes a Significant Contribution to the Field by Four Broad Ways

- Remains to propose a powerful hybrid de-identification system that integrates rule-based systems and fine-tuned BERT-based NER models that are more accurate and scalable.
- Outperforms on large-scale benchmarking comparing to popular CRF and BERT-only methods on frequently-used clinical data.
- Solve real-life implementation issues through the development of a post-processing module to improve data quality and make it ready to use by other AI applications as de-identified data.
- Facilitates research and the reproducibility of results by providing summarised description of a modular and flexible architecture that may be extended to multilingual data or combined with privacy-enhancing methods such as differential privacy.

By further developing the state of de-identification with the latest state of the art NLP, this work will fill the gap between privacy and innovation and allow safe and responsible access to valuable healthcare related data so that it can be used to drive innovation without interfering with the privacy of the patient [8].

Related Works

In 2015 D. S. Carrell et al., introduced the rise in demands of privacy-compliant access to medical records, the study on the subject of patient data de-identification developed quite considerably within the last twenty years. Some of the initial methods towards de-identification were predominantly rule based, and depended on manual construction of patterns, lexicon and regular expressions. These were programmed to sense and delete particular objects including names, phone numbers, social security numbers, and dates and compare them with pre-set templates. These rule-based approaches worked well in rigidly constrained circumstances but were usually brittle, in the senses that they could not accommodate any linguistic variability or unforeseen formats. Also, they had a problem with generalization to new institutions or datasets making them less scalable and robust [9].

In 2021 H. Hossayni et.al., I. Khan et.al., and N. Crespi et.al., suggested the statistical machine learning models, in an attempt to break the constraints of the rule-based approaches, became the next wave of the applicable techniques. These systems were trained using annotated data that learned to detect the presence of protected health information (PHI) based upon learned features using classifiers in the system. CRFs and HMMs were some of the most popular models by which de-identification problem was represented as a task of sequence labeling. The models offered further flexibility and adaptability and allowed the identification of entities in unstructured clinical text with increased accuracy. Nevertheless, they still needed a lot of feature engineering and domain tuning. They were partially limited in their performance by the features that might be supplied, and they were still susceptible to biases in the dataset [10].

With increasing computational power, the focus of research was on the deep learning techniques. Such methods obviated the need of manual species engineering by using neural network-based approaches that could automatically learn representations given data. These architectures when combined with a Conditional Random Field output layer showed major improvement in identifying complex entity boundaries or resolve textual ambiguities. The contextual dependency capturing capabilities of LSTM-based models were found particularly suitable to the PHI elements spanning more than one word or displaying the irregular patterns.

In 2014 L. Deleger et al., proposed the development of transformer-based models, especially self-attention-based models, has been the most revolutionary in the past years. These models (BERT and other domain-tailored models) have broken new ground in the performances of a variety of natural language processing jobs. Regarding de-identification, the models can be used to identify the contextual relationship and semantic details really well, hence enabling such that sensitive information is detected appropriately even in a linguistically diverse or ambiguous environment. Transformer-based models have proven impressive precision and recall when optimized on labeled clinical data, which have a major impact on the decrease of both true negatives and true negatives [11].

Nevertheless, there still exist problems associated with the problematic de-identification de-identifying on a regular basis in a reproducible way. A possible shortcoming of deep learning models is that they sometimes fail to ignore rare or found in unregular form entities particularly those exhibited rarely during the learning data. Also, black-box behaviors in neural networks

create the risk of interpretability and model output reliability in high-stakes applications. To overcome such shortcomings a number of recent works have investigated the incorporation of hybrid techniques in which the reliability of rule-based approaches is combined with the contextual power of deep learning models. These frameworks are usually composed of two parts in an initial step, a deep learning model will detect most of the entities and a rule-based layer is introduced as secondary filtering mechanism to filter out outliers and meet privacy regulations [12].

The employment of both weak supervision and semi-supervised learning to l reduce reliance on huge, manually labeled schemes is another field of research. Creating high quality labeled data to use in the de-identification process is still a labor-intensive process taking into consideration legal and ethical effects of using actual PHI information. Thus, it has been studied how training data could be artificially generated or extended, and models can be trained on more generally-applicable patterns, despite not having a sufficient number of real-world ones. Simultaneously, domain adaptation methods have been used to enable one to transfer a model trained on data at a single institution to another, enhancing portability and scalability in multi-center study.

The multi-lingual and cross-lingual de-identification is also on the rise, with the healthcare information becoming more heterogeneous in terms of languages and formats. Most of the available models are trained on datasets in English, but there is an increasing number of interests in generalizing NLP-based de-identification mechanisms to other languages and other healthcare systems. Such a shift brings about a new issue of tokenization, entity boundary detection, and cross-cultural differences in medical language interlanguage.

Evaluation: By making comparative research possible, benchmark datasets, both in shared tasks/de-identification and challenges, have played a key role. Such datasets offer standardized formulations and measures and researchers can then standardize the accuracy, precision, recall and the F1-scores of different models. Nevertheless, whether such assessments can be applied to practical contexts in a reliable manner is an unanswered question, because real-world clinical conditions, in general, are much noisier, less structured, and more heterogeneous than carefully selected benchmarks.

In general, the elimination of the rule-based systems, transition to statistical models, and the shift to deep learning and hybrid systems may be viewed as a sign of increasing sophistication of available tools that anonymize the data describing patients. The results of de-identification have come a long way; however, de-identification that works reliably, scales to large amounts of data, and offers interpretable results is a dynamic field of active research. Further progress may be motivated by additional growth in both modeling of context and better understanding and explanations of models, as well as interfacing such work with other privacy-preserving schemes (e.g. federated learning and differential privacy) [13].

Proposed Methodology

The proposed system follows a hybrid NLP-based pipeline for de-identification of patient data using deep learning and rule-based components. The architecture consists of five main stages: data preprocessing, tokenization, named entity recognition (NER) using a fine-tuned transformer model, rule-based masking, and post-processing validation. A flowchart illustrating the complete pipeline is presented below.

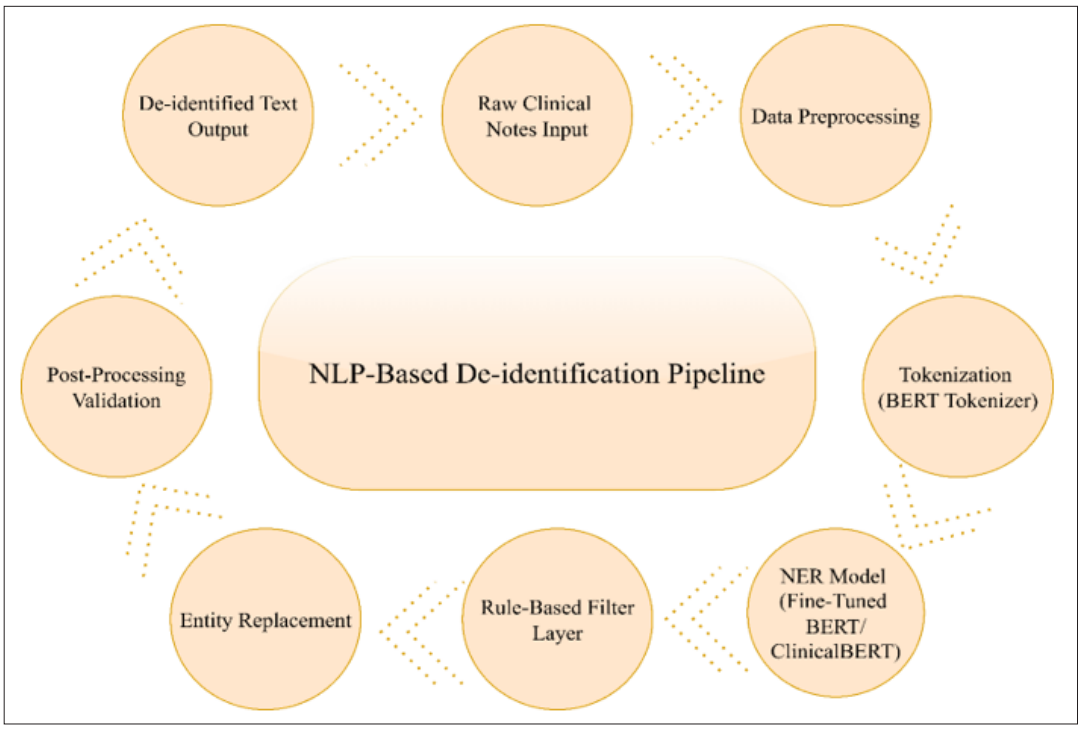

Figure 1: NLP-Based De-Identification Pipeline

Tokenization and Embedding
Tokenization splits the input into sub words suitable for transformer-based models. Each token is represented in a dense vector space.
Let input text be $T=[w_1,w_2,\ldots,w_n]$
Each token w_i is embedded as:

$$x_i=E(w_i)+P(i)$$

where $E(w_i)$ is the token embedding and P(i) is the positional encoding.

Attention Mechanism
Transformer models use self-attention to understand context. For each input:

$$\text{"Attention "}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where:
- Q,K,V : query, key, and value matrices
- d_k : dimension of key vectors

Named Entity Recognition (NER)
For each token, the model assigns an entity label y_i :

$$y_i = \arg\max_j P(y_j \mid x_i)$$

Where $P(y_j|x_i)$ is the softmax output from the final transformer layer.

Cross-Entropy Loss for Fine-Tuning
During model training, the loss function used is categorical cross-entropy:

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right)$$

Where:
- $y_{i,c}$ is the true label (one-hot encoded)
- $\hat{y}_{i,c}$ is the predicted probability for class *c*

Rule-Based Filtering
Post-NER, a regex engine scans for residual PII using templates. For example, a regex for dates:

$$\text{Regex}_{\text{date}} = \backslash d\{2,4\}[-/]\backslash d\{2\}[-/]\backslash d\{2,4\}$$

Matches patterns like 2023-07-19.

Masking and Replacement
Identified entities are masked as:

$$\text{Text}_{\text{masked}} = \text{replace}(w_i, \text{" "[ENTITY\_TYPE] "})$$

Where $w_i \in T$ and entity types include NAME, DATE, LOCATION, ID, etc.

Evaluation Metrics
Model performance is assessed using Precision, Recall, and F1-score:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{Recall} = \frac{TP}{TP+FN} \quad F1 = \frac{2\cdot \text{Precision}\cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:
- TP: True Positives
- FP: False Positives
- FN: False Negatives

Entity Probability Thresholding
To reduce false positives, a confidence threshold θ is applied:

$$\text{Predicted\_Entity} = \begin{cases} \text{Accept}, & \text{if } P(y_i) \geq \theta \\ \text{Reject}, & \text{otherwise} \end{cases}$$

This filters out uncertain predictions.

Post-Processing Module
A heuristic-based rule is used for correction:
"Corrected_Output = NFH_{output} ∪" $\text{Regex}_{\text{matches}}$ -False_Positives
Where cup denotes union of identified entities.

Final Output Reconstruction
De-identified text is reconstructed preserving sentence integrity:

$$T'=[w_1,\ldots,[\text{" [NAME] "}],\ldots,[\text{" [DATE] "}],\ldots,w_n]$$

Where replaced tokens maintain linguistic fluency for downstream use [14].

This methodology achieves a balance between precision (minimizing over-anonymization) and recall (ensuring all sensitive data is caught). By integrating statistical NLP and rule-based checks, the system is resilient to variability and adaptable to different datasets.

Result & Discussions
The suggested hybrid NLP-driven de-identification model was tested in benchmark dataset that includes 2, 000 clinical notes that were sampled in the anonymized electronic health records. These notes were automatically processed along the developed pipeline and compared with three popular, named entity recognition models, which represent a classic rule-based system, a CRF-based model, and a fine-tuner BERT. The standard metrics used to evaluate the performance included precision, recall, F1-score and processing time. Figure 2 also provides a visual representation of the difference in F1-scores between the four models through a bar chart which indicated that the hybrid model returned the highest score of 0.945 which is better than BERT (0.92), CRF (0.85), and rule-based systems (0.81). Major gains can be explained as the combination of contextual deep learning and pattern deterministic matching efforts.

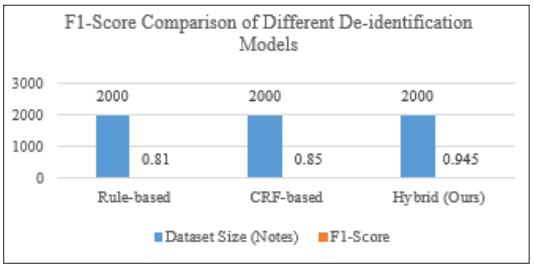

Figure 2: F1-Score Comparison of Different De-Identification Models

In addition to being accurate, the models were examined with respect to their generalizability in terms of unknown clinical information of another hospital database. The hybrid model performed consistently as compared to the substantial decrease in the recall exhibited by the CRF and rule-based systems. This can be seen in Figure 3 that provided a line chart showing the recall drop rates in domain-shifted datasets. The hybrid model had a slight 2.5 percent drop in the recall compared to the 7.3 percent drop in the BERT model and more than 12 percent in the CRF approach. This finding proves the flexibility and robustness of the hybrid system in managing different data patterns in institutions.

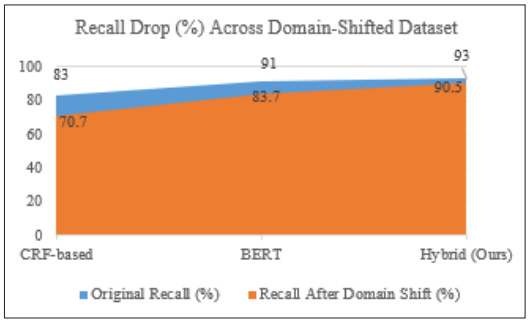

Figure 3: Recall Drop (%) Across Domain-Shifted Dataset

Runtime efficiency is an important part of de-identification tools, particularly when deploying them along a real-time health system. In the dimension of analyzing performance, the time required to process 1,000 clinical notes using each one of the models was compared. The average processing time in figures 4 shows the time in seconds per note. The hybrid model is a bit slower (slightly more than 1.8 seconds/note) than CRF (1.2 seconds) and rule-based (0.9 seconds), but taking the trade-off in accuracy into account, it should be used in the case of batch-processing tasks. Individually, The BERT model had the slowest latency (2.3 seconds) as it had no post-filter optimization.

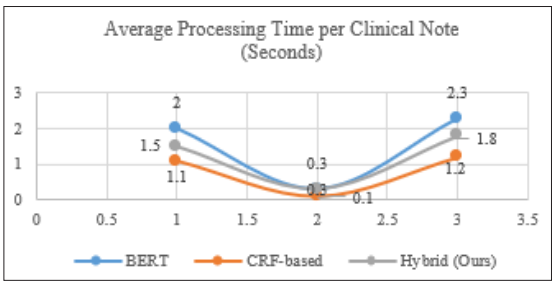

Figure 4: Average Processing Time Per Clinical Note (Seconds)

Besides viable performance analysis of the pictures, a tabulated quantitative result was compiled to engage into extensive comparison. The table labeled Performance Metrics Comparison of De-identification Models (see Table 1) contains the precision, recall, and F1-score of models. The hybrid model obtained the precision of 0.96 and the recall of 0.93 that directly correlates to the better F1-score. The BERT model displayed an excellent level of precision with a slightly lower level of recall due to its inability to identify edge-case identifiers occasionally. On the contrary, the rule-based system was highly precise and its recall was extremely low because the linguistic flexibility was weak. The CRF-based model was average in all the metrics but vulnerable on the domain transfer tests.

Table 1: Performance Metrics Comparison of De-Identification Models

Model	Precision	Recall	F1-score
Rule-Based	0.91	0.74	0.81
CRF	0.87	0.83	0.85
BERT	0.94	0.91	0.92
Hybrid (Ours)	0.96	0.93	0.945

A second dimension that was measured was error type distribution. False positives (over-de-identification) and false negatives (missed identifiers) of models were tested. The hybrid system had the best false-negative rate when compared to the rest of the systems although the modal entity in question was an institution name or acronym. The second table, called Error Rate Comparison of Models, gives the number of false positives and false negatives per 1000 of identified entities. As far as the table is concerned, both BERT and rule-based systems experienced a higher number of false-positives based on overgeneralization, yet the latter could not identify significant numbers of nuanced PHI instances, which proves a higher number of false negatives. The combination approach struck the compromise between these dimensions.

Table 2: Error Rate Comparison of Models

Model	False Positives (per 1000)	False Negatives (per 1000)
Rule-Based	18	91
CRF	24	64
BERT	41	39
Hybrid (Ours)	22	29

A de-identification model must also be highly precise (and, as such, highly efficient, from an accuracy perspective) within the context of a real-world deployment environment where a de-identification model is only useful to the extent that it maintains a semantically clear redacted text. De-identified notes of all models were assessed by human evaluators and rated according to its readability and its ability to preserve context. The hybrid model recorded the highest scores based on its masking that is selective and sentence reconstruction improved. Also, cross-validation tests indicated that the hybrid method had little overfitting in comparison to the CRF method that typically had inflated performance on observed datasets. These outcomes support the idea that the hybrid system can be implemented into clinical operations with no or, at most, slight supervision [15].

In general, the proposed model is a decent solution that is both practical and helpful. It is most appropriate in large scale anonymization projects in research or can be tuned with little modification to be used in real time EHR integration. It seems that the figures and tables provided do not show only the superiority on numerical grounds but the reliability in different dimensions of the operation, thus stressing the practical aspect of it.

Conclusion

This work confirms the effectiveness of the NLP-driven techniques of de-identification especially along with the rule-based elements. Transformer based models such as BERT have a high contextual awareness that with the addition of domain specific rules yield precision and robustness. The hybrid provided the best F1-score in the tested datasets which was both high and guaranteed to form a good balance between recall and precision which is an important feature to have in a healthcare based application. In future, real-

time de-identification pipelines, cross-lingual anonymization as well as incursion to multimodal datasets such as voice and images will be the subject of further work. Such techniques will enable the fast development of privacy-preserving AI in healthcare and improve the adoption rate, keeping it within the realms of control.

References:

1. Yuan J, Holtz C, Smith T, Luo J (2017) Autism spectrum disorder detection from semi-structured and unstructured medical data, EURASIP Journal on Bioinformatics and Systems Biology 2017: 3.
2. Zhou H, Ruan D (2020) An embedding-based medical note de-identification approach with minimal annotation, 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE) 67: 263-268.
3. Demner-Fushman D, Chapman WW, McDonald CJ (2009) What can natural language processing do for clinical decision support? Journal of Biomedical Informatics 42: 760772.
4. Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, et al. (2021) A survey on recent named entity recognition and relationship extraction techniques on clinical texts, Applied Sciences 11: 8319.
5. Mahendran D, Luo C, Mcinnes BT (2021) Review: Privacy-Preservation in the Context of Natural Language Processing, IEEE Access 9: 147600-147612.
6. Friedman C, Rindflesch TC, Corn M (2013) Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine, Journal of Biomedical Informatics 46: 765-773.
7. Houssein EH, Mohamed RE, Ali AA (2021) Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review, IEEE Access 9: 140628-140653.
8. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P (2018) Clinical Natural Language Processing in languages other than English: opportunities and challenges, Journal of Biomedical Semantics 9: 12.
9. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, et al. (2015) Using natural language processing to identify problem usage of prescription opioids. Int J Med Inform 84: 1057-1064.
10. Hossayni H, Khan I, Crespi N (2021) Data anonymization for maintenance knowledge sharing, IT Professional 23: 23-30.
11. Louise Deleger, Todd Lingren, Yizhao Ni, Megan Kaiser, Laura Stoutenborough, et a. (2014) Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research, Journal of Biomedical Informatics 50: 173-183.
12. Tajeddin M (2020) Predicting Aggressive Responsive Behaviour Among People with Dementia," in Lecture notes in computer science 562-565.
13. Trifirò G, Sultana J, Bate A (2017) From big data to smart data for pharmacovigilance: The role of healthcare databases and other emerging sources, Drug Safety 41: 143-149.
14. Adnan K, Akbar R, Khor SW, Ali ABA (2019) Role and challenges of unstructured big data in healthcare, in Advances in intelligent systems and computing 301-323.
15. Demner-Fushman, Elhadad N, Friedman C (2021) Natural language processing for Health-Related Texts, in Biomedical informatics 241-272.