applied sciences

MDPI

Article

Utility of Features in a Natural-Language-Processing-Based Clinical De-Identification Model Using Radiology Reports for Advanced NSCLC Patients

Tanmoy Paul [1,2], Humayera Islam [2,3,4], Nitesh Singh [2,3,5], Yaswitha Jampani [2,3,5], Teja Venkat Pavan Kotapati [5], Preethi Aishwarya Tautam [5], Md Kamruz Zaman Rana [2,3,5], Vasanthi Mandhadi [1,2,3], Vishakha Sharma [6], Michael Barnes [6], Richard D. Hammer [7] and Abu Saleh Mohammad Mosa [1,2,3,4,5,*]

1 Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA
2 NextGen Biomedical Informatics Center, University of Missouri, Columbia, MO 65211, USA
3 University of Missouri School of Medicine, University of Missouri, Columbia, MO 65211, USA
4 Institute for Data Science and Informatics, University of Missouri, Columbia, MO 65211, USA
5 Department of Health Management and Informatics, University of Missouri, Columbia, MO 65211, USA
6 Roche Diagnostics, F. Hoffmann-La Roche, Santa Clara, CA 95050, USA
7 Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, MO 65211, USA
* Correspondence: mosaa@health.missouri.edu

Abstract: The de-identification of clinical reports is essential to protect the confidentiality of patients. The natural-language-processing-based named entity recognition (NER) model is a widely used technique of automatic clinical de-identification. The performance of such a machine learning model relies largely on the proper selection of features. The objective of this study was to investigate the utility of various features in a conditional-random-field (CRF)-based NER model. Natural language processing (NLP) toolkits were used to annotate the protected health information (PHI) from a total of 10,239 radiology reports that were divided into seven types. Multiple features were extracted by the toolkit and the NER models were built using these features and their combinations. A total of 10 features were extracted and the performance of the models was evaluated based on their precision, recall, and F_1-score. The best-performing features were n-gram, prefix-suffix, word embedding, and word shape. These features outperformed others across all types of reports. The dataset we used was large in volume and divided into multiple types of reports. Such a diverse dataset made sure that the results were not subject to a small number of structured texts from where a machine learning model can easily learn the features. The manual de-identification of large-scale clinical reports is impractical. This study helps to identify the best-performing features for building an NER model for automatic de-identification from a wide array of features mentioned in the literature.

Keywords: protected health information; natural language processing (NLP); named entity recognition (NER); de-identification; conditional random field (CRF)

Citation: Paul, T.; Islam, H.; Singh, N.; Jampani, Y.; Kotapati, T.V.P.; Tautam, P.A.; Rana, M.K.Z.; Mandhadi, V.; Sharma, V.; Barnes, M.; et al. Utility of Features in a Natural-Language-Processing-Based Clinical De-Identification Model Using Radiology Reports for Advanced NSCLC Patients. *Appl. Sci.* **2022**, *12*, 9976. https://doi.org/10.3390/app12199976

Academic Editor: Konstantinos E. Psannis

Received: 24 August 2022
Accepted: 29 September 2022
Published: 4 October 2022

1. Introduction

The electronic health record (EHR) is a collection of patients' health information in a digital format. Text-based medical records are an important resource for the EHR and an enriched knowledge source for medical research. One of the major limitations of the large-scale use of the EHR is the privacy of information in medical corpora. The Health Insurance Portability and Accountability Act (HIPAA) in the United States defines 18 types of protected health information (PHI) that need to be removed from medical records before circulating these for secondary usage. The PHI items encompasses name, phone number, geographic location, medical record number, social security number, etc. [1,2]. It

is impractical to de-identify large-scale data manually, since this can be expensive, time-consuming, and prone to error. Therefore, a reliable automated de-identification system is highly desired.

Healthcare and biomedical research has been significantly impacted by the utilization of natural language processing (NLP). Named entity recognition (NER) is a basic functionality of clinical NLP. It is defined as the identification of desired entities from texts. The main task of NER is to identify and classify specific words and meaningful phrases [3]. NER from clinical texts has been an area of increasing interest in recent years in the medical domain and drive clinical decision support (CDS) to enable healthcare providers to make personalized patient care decisions. Medical reports consist of both coded and unstructured texts. Although coded data can easily be de-identified, it is extremely challenging to de-identify unstructured texts.

The NER techniques can be divided into four categories: lexicon-based, heuristic-based, machine learning, and hybrid techniques [4–6]. The majority of the primitive NER systems applied lexicon- and heuristic-based techniques. These systems deployed rules derived from syntactic-lexical patterns as well as information lists to classify and identify named entities (NE) [7–10]. Since these approaches exploit language-related knowledge, these are considered to be highly efficient [11]. However, there are a few limitations of these techniques, as they are domain-specific, expensive, and involve human expertise in that domain. Due to these limitations, researchers have shifted their interests towards machine-learning-based techniques.

There have been numerous efforts to improve the performance of clinical NER systems by undertaking various strategies to exploit the existing infrastructure of machine learning algorithms. Machine learning techniques can be either unsupervised, semi-supervised, or supervised. Some studies used an ensemble of multiple machine learning methods [12,13]. Hybrid machine learning models with high confidence rules have also been applied [14]. Multiple studies have used unsupervised models using clustering algorithms [15,16]. The existing literature shows that supervised machine learning algorithms have been incorporated in most of the best-performing NER systems.

Multiple algorithms are used to build supervised NER models, such as conditional random fields (CRF), maximum entropy (ME), and structured support vector machines (SVMs) [17–19]. These algorithms build NER models by exploiting the predefined multidimensional feature sets from text datasets. The performance of a supervised model depends largely on the selection of proper features. In this study, we investigated the performance of a CRF-based NER model by using multiple features and their combinations.

The objective of this study was to identify the best-performing features for a de-identification NER model. We did not aim to build a high-performance model. In fact, our aim was to find the features which work the best to build such a model. The major contribution of this study is that it helps to identify the best possible features from the wide range of features mentioned in the existing literature. This article first discusses the data, preparation of the gold standard repository, and the experimental setup in the Materials and Methods section. The performances of the NER models are presented and evaluated in the Results section to identify the best-performing features. Finally, the key finding of the experiments were discussed and summarized in the Discussion and Conclusions sections.

2. Materials and Methods

2.1. Dataset

In this study, 7 types of radiology reports were acquired from the EHR of the University of Missouri Healthcare. The total number of reports was 10,239. Table 1 shows the number of reports in each type. All the patients were diagnosed with IIIB or greater stages of cancer and the diagnoses were made between 2010 and 2018. In this study, we identified five types of PHI items from these unstructured clinical texts. The PHIs were NAME, DATE, HOSPITAL, LOCATION, and ID (patient visit number).

Table 1. Description of dataset.

Report Type	Abbreviation	Number of Reports
Interventional Radiology	IR	273
Mammography	MA	167
Magnetic Resonance Imaging	MRI	1010
Nuclear Medicine Technique	NM	655
Ultrasound	US	644
Computed Tomography	CT	2741
X-ray	XR	4749

2.2. NLP Toolkit

To annotate the PHIs in the radiology reports, we used an NLP software called MITRE Identification Scrubber Toolkit (MIST) [20]. All of the reports passed through a two-step annotation process for tagging the HIPAA-defined PHI items. At the first step, each report was manually annotated by a data annotator. It was reviewed by a second annotator at the second step and necessary correction was made if an error was found. A completely annotated gold standard repository was created through this two-step annotation process. Multiple supervised NER models were built by using this gold standard dataset.

The NER model was built using a different NLP toolkit called Clinical Language Annotation, Modelling and Processing (CLAMP) [21]. The feature extraction module of CLAMP enabled us to develop a conditional random field (CRF)-based model using various sets of features. These features could be used both individually and in a combination.

2.3. Performance Metrics

In this study, performances of the models were evaluated using three performance metrics, such as precision (P), recall (R), and F_1-score (F1). Precision is a measure that indicates how many of the positive predictions are actually correct. Recall indicates how many positive cases were predicted out of all the positive cases. F_1-score is a measure that combines both precision and recall. Precision is an important measure when there is a high cost associated with false positive. On the other hand, recall is important when the cost of false negative is high. F_1-score is a measure that is used when a balance between precision and recall is expected. It is usually described as the harmonic mean of these two metrics. The metrics are calculated as below:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F_1\text{–}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

There were two stages of the experiment: Identification of the best feature set and determination of the minimum size of the training set to produce the best performance. Figure 1 depicts the workflow of the entire experiment.

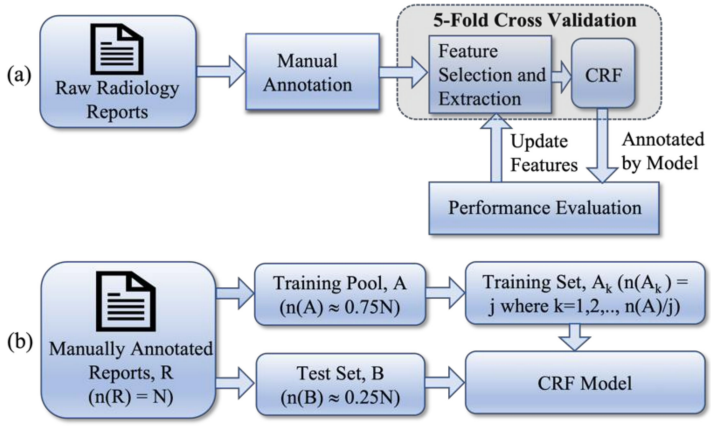

Figure 1. Workflow of the experiment: (**a**) Identification of the best feature set, (**b**) determination of the minimum size of the training set to obtain the best performance.

2.4. Identification of the Best Feature Set

At this stage of the experiment, our goal was to determine which features make the NER model perform the best in annotating the PHI mentioned above. The workflow of this stage of the experiment is illustrated in Figure 1a. A total of 10 features were extracted and a model was built for each type of report using each of these 10 features. The performance of each model was evaluated by applying a 5-fold cross-validation technique. Performance was evaluated by calculating precision, recall, and F_1-score. Based on the performance of models trained by individual features, the best features were selected for the second stage of the experiment. Once we found out the best-performing features, we built two more NER models for each type of reports. One model was built with the best four features (M4) and the other one was built with all of the 10 features (M10). Moreover, similar to these two models, we built two more models using all of the radiology reports. The objective of the latter two models was to find out whether the performance of the features varies significantly when trained with multiple types of reports.

The performance of a supervised NER model may vary significantly with the feature selection. Although existing literature provides numerous features used in supervised NER, the utility of the features is domain-specific. Hence, not all the features prove to be effective depending on the nature of the data. Here, we present all the features which were explored in this study.

Brown Clustering (BC) is a hierarchical clustering language model [22]. It cluster words to maximize the mutual information of bigrams. In this model, a word class can be selected at various levels of the hierarchy. As a result, poor clusters with small number of words can be compensated. N-gram is a feature that utilizes the co-occurrence of neighboring items of a named entity. [23] For example, in a 2-g model, the frequency of each left-right pair of neighbors of a tagged entity is calculated. Similarly, the prefix and suffix of a named entity are represented by the prefix-suffix feature.

Random Indexing (RI) is a type of dimensionality reduction technique that is widely used in natural language processing. It is a random projection method that approximates similarities in sets of co-occurrence weights by incrementally generating distributional representation [24]. An index is associated with each document or context which forms a multidimensional sparse random vector. Another multi-dimensional vector of integers named distributional vector is associated with each word. Distributional vectors are initiated at zero. It is updated by adding a index vector whenever the associated word is found in the context. Finally, the semantic relationship between words are evaluated based on the similarity among respective distributional vectors.

Section (S) feature represents the section where a named entity is encountered. Sentence pattern (SP) utilizes built-in rules by CLAMP and recognizes the pattern of a sentence. Word Embedding (WE) is similar to brown clustering and random indexing since it is a

representation of word distribution that is produced on unlabeled data. In discrete word embedding (DWE), a distributed representation enables the extraction of character level of features at word level. Moreover, it does not require any syntactic knowledge. [25]. Word Shape (WS) identifies whether a word starts with a number, English letter, etc., or not. Word regular expression (WRE) feature is the regular expression patterns of words that may represent a specific type of named entity.

2.5. Determination of the Minimum Size of the Training Set

At the second stage of the experiment, our objective was to observe how the performance of a model varies with the number of training data and in turn, to determine the minimum size of training set to achieve the best performance of the model. At first, all the reports (n(R) = N) of each type were divided into two groups: training pool (A) and test set (B). The test set was selected randomly and it consisted of 25% of all the reports of that type. The rest of the reports were included in the training pool. Using the best feature sets derived from the first stage of the experiment, multiple models were built by using variable number of training set from the training pool. Figure 1b shows all these steps for training set size j where:

- N = total number of reports of a type
- $n(A) \approx 0.75N$ = number of reports in the training pool
- $n(B) \approx 0.25N$ = number of reports in the test set
- $n(A_k)$ = j = 25, 50, 100, 200, ... , n(A) = number of reports in each training dataset
- k = 1, 2, 3, ... , n(A)/j = number of iterations for each training dataset, A_k.

For each type of report, we first selected varying size of training set where the size was denoted by j and j = 25, 50, 100, 200, ... , n(A). For each value of j, a total of n(A)/j training sets (A_k) were created where the size of each training set, $n(A_k)$, was j and the sets were mutually exclusive. For each training set, A_k, a model was created using the best feature set and the performances of all such models were evaluated using the same test set (B) which was already set aside. The mean F_1-score of all the models, for each training set size j, was selected as the metric to evaluate the performance for that training set size. Finally, variation of the models' performances with the varying size of training set were investigated.

3. Results

Figure 2 shows the results of the Analysis-1 (identifying the best feature set) in the form of heatmaps. There are seven types of reports and three different performance metrics. For each type of report, there are three heatmaps, each corresponding to a performance metric. Therefore, there are a total of 21 heatmaps. While describing the results, we will denote each heatmap with a corresponding report-metric pair (such as IR-Precision, MRI-Recall, etc.). The columns of the heatmap represent the PHI items and the rows represent the features. In the figure, LOCATION, DATE, HOSPITAL, PHONE, and NAME, were denoted by LC, DT, HS, PH, and NM. All of the heatmaps were generated based on the Red-Yellow-Green color scale. In each heatmap, the red color represents the highest value, and the green represents the lowest. It can easily be observed from the figure that the section and sentence pattern features failed to identify any of the PHI items as for all the reports their precision, recall, and F_1-score were N/A, 0.00, and 0.00, respectively. Therefore, we will exclude these two features from the further description of the results.

The IR heatmaps show no such feature that could outperform all other features across all of the performance metrics in identifying the PHI items. For example, the IR-Precision heatmap illustrates that the highest value of 0.73 was achieved by n-gram and prefix-suffix features in identifying LOCATION. For DATE, the highest precision was achieved by word shape with a value of 0.96. For HOSPITAL, the highest value of 0.71 was achieved by n-gram, prefix-suffix, and word shape features. Word embedding and prefix-suffix yielded the highest values for NAME (0.85) and ID (0.88), respectively. Therefore, in terms of precision, the best features were n-gram, prefix-suffix, word embedding, and word

shape. The IR-Recall heatmap reveals that for LOCATION, n-gram had the highest value of 0.69. For DATE, the highest value of 0.85 was achieved by discrete word embedding, n-gram, prefix-suffix, word embedding, and word shape features. Similar observation reveals that for HOSPITAL, NAME, and ID, the best-performing features were n-gram, word embedding, and word shape. Though discrete word embedding shared the highest value with other features for DATE, it performed poorly for NAME and had a value of 0.00 for ID. Therefore, we excluded it from the set of best features. IR-Recall heatmap showed that the best-performing features were n-gram, prefix-suffix, word embedding, and word shape. Moreover, a similar study of the IR-F_1 Score heatmap gave the same four features as the best-performing ones.

Feature	Precision					Recall					F1-Score					Report
	LC	DT	HS	NM	ID	LC	DT	HS	NM	ID	LC	DT	HS	NM	ID	
BC	0.69	0.91	0.67	0.54	N/A	0.62	0.84	0.68	0.25	0.00	0.65	0.88	0.68	0.34	0.00	
DWE	0.70	0.87	0.70	0.59	N/A	0.68	0.85	0.69	0.38	0.00	0.69	0.86	0.70	0.46	0.00	
NG	0.73	0.87	0.71	0.84	0.72	0.69	0.85	0.71	0.76	0.60	0.71	0.86	0.71	0.80	0.66	
PS	0.73	0.94	0.71	0.76	0.88	0.69	0.85	0.71	0.67	0.49	0.71	0.89	0.71	0.71	0.63	
RI	0.70	0.88	0.69	0.45	N/A	0.63	0.84	0.69	0.15	0.00	0.66	0.86	0.69	0.23	0.00	IR
S	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
SP	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
WE	0.69	0.85	0.69	0.85	0.67	0.64	0.85	0.68	0.77	0.45	0.67	0.85	0.69	0.81	0.54	
WS	0.71	0.96	0.71	0.78	0.73	0.68	0.85	0.71	0.67	0.60	0.70	0.90	0.71	0.72	0.66	
WRE	0.09	0.88	N/A	0.42	N/A	0.01	0.83	0.00	0.01	0.00	0.01	0.86	0.00	0.03	0.00	
BC	0.67	0.89	0.67	0.67	N/A	0.58	0.71	0.64	0.29	0.00	0.62	0.79	0.65	0.40	0.00	
DWE	0.66	0.78	0.66	0.56	0.98	0.58	0.64	0.62	0.30	0.32	0.62	0.70	0.64	0.39	0.48	
NG	0.67	0.86	0.67	0.86	0.91	0.59	0.79	0.65	0.79	0.91	0.62	0.82	0.66	0.82	0.91	
PS	0.68	0.87	0.66	0.84	0.87	0.60	0.78	0.64	0.72	0.74	0.64	0.82	0.65	0.78	0.80	
RI	0.68	0.84	0.63	0.57	N/A	0.59	0.71	0.62	0.19	0.00	0.63	0.77	0.63	0.29	0.00	MA
S	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
SP	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
WE	0.69	0.86	0.71	0.83	0.76	0.60	0.79	0.66	0.80	0.53	0.65	0.82	0.68	0.82	0.62	
WS	0.68	0.85	0.66	0.84	0.93	0.59	0.81	0.63	0.71	0.90	0.63	0.83	0.65	0.77	0.91	
WRE	0.33	0.80	0.33	0.66	0.97	0.01	0.64	0.02	0.13	0.32	0.03	0.71	0.03	0.23	0.48	
BC	0.80	0.88	0.80	0.73	N/A	0.77	0.74	0.74	0.21	0.00	0.78	0.80	0.77	0.32	0.00	
DWE	0.80	0.81	0.82	0.46	0.04	0.80	0.74	0.75	0.22	0.00	0.80	0.77	0.78	0.29	0.00	
NG	0.82	0.89	0.82	0.90	0.85	0.82	0.83	0.81	0.86	0.83	0.82	0.86	0.82	0.88	0.84	
PS	0.81	0.94	0.81	0.89	0.92	0.81	0.83	0.80	0.82	0.79	0.81	0.88	0.81	0.85	0.85	
RI	0.81	0.87	0.81	0.45	N/A	0.81	0.73	0.75	0.14	0.00	0.81	0.79	0.77	0.22	0.00	MRI
S	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
SP	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
WE	0.81	0.88	0.81	0.88	0.77	0.81	0.82	0.80	0.85	0.54	0.81	0.85	0.81	0.87	0.63	
WS	0.81	0.92	0.81	0.90	0.87	0.80	0.83	0.80	0.81	0.83	0.81	0.88	0.81	0.85	0.85	
WRE	0.55	0.85	0.54	0.65	0.58	0.24	0.77	0.24	0.09	0.09	0.33	0.81	0.33	0.15	0.16	
BC	0.80	0.91	0.82	0.83	N/A	0.72	0.72	0.82	0.13	0.00	0.76	0.80	0.82	0.22	0.00	
DWE	0.82	0.90	0.80	0.81	0.97	0.72	0.71	0.82	0.27	0.22	0.77	0.79	0.81	0.40	0.36	
NG	0.82	0.90	0.83	0.91	0.88	0.80	0.83	0.83	0.80	0.84	0.81	0.86	0.83	0.85	0.86	
PS	0.82	0.92	0.82	0.90	0.86	0.80	0.82	0.82	0.76	0.73	0.81	0.87	0.82	0.82	0.79	
RI	0.83	0.88	0.79	0.69	N/A	0.72	0.73	0.82	0.07	0.00	0.77	0.80	0.80	0.13	0.00	NM
S	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
SP	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
WE	0.79	0.90	0.83	0.89	0.60	0.78	0.80	0.82	0.77	0.54	0.79	0.84	0.83	0.83	0.57	
WS	0.82	0.91	0.82	0.90	0.89	0.80	0.83	0.82	0.74	0.85	0.81	0.87	0.82	0.82	0.87	
WRE	0.74	0.86	0.71	0.81	N/A	0.42	0.75	0.47	0.58	0.00	0.53	0.80	0.56	0.10	0.00	
BC	0.75	0.88	0.75	0.78	0.31	0.73	0.79	0.75	0.56	0.00	0.74	0.83	0.75	0.65	0.01	
DWE	0.76	0.93	0.77	0.78	0.94	0.75	0.80	0.75	0.83	0.03	0.75	0.86	0.76	0.80	0.06	
NG	0.77	0.90	0.78	0.97	0.86	0.76	0.86	0.77	0.94	0.85	0.76	0.88	0.78	0.95	0.85	
PS	0.76	0.95	0.77	0.94	0.86	0.75	0.85	0.76	0.89	0.73	0.75	0.89	0.77	0.91	0.79	
RI	0.76	0.86	0.77	0.71	0.47	0.76	0.79	0.76	0.56	0.02	0.76	0.83	0.76	0.63	0.05	US
S	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
SP	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
WE	0.77	0.89	0.78	0.96	0.77	0.76	0.84	0.77	0.94	0.49	0.76	0.87	0.78	0.95	0.60	
WS	0.76	0.94	0.78	0.94	0.86	0.75	0.85	0.76	0.87	0.84	0.76	0.89	0.77	0.91	0.85	
WRE	N/A	0.83	N/A	0.24	0.86	0.00	0.78	0.00	0.03	0.03	0.00	0.81	0.00	0.06	0.06	
BC	0.93	0.83	0.94	0.68	0.00	0.91	0.72	0.92	0.41	0.00	0.92	0.77	0.93	0.51	0.00	
DWE	0.94	0.82	0.94	0.69	0.77	0.92	0.75	0.92	0.81	0.09	0.93	0.79	0.93	0.74	0.16	
NG	0.94	0.83	0.94	0.94	0.80	0.93	0.75	0.94	0.93	0.70	0.94	0.79	0.94	0.93	0.75	
PS	0.95	0.83	0.94	0.92	0.88	0.93	0.72	0.94	0.90	0.62	0.94	0.77	0.94	0.91	0.73	CT
RI	0.94	0.82	0.94	0.54	0.22	0.93	0.71	0.93	0.37	0.01	0.93	0.76	0.93	0.44	0.01	
S	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
SP	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
WE	0.94	0.83	0.94	0.94	0.71	0.93	0.75	0.94	0.93	0.41	0.93	0.79	0.94	0.93	0.52	
WS	0.94	0.83	0.94	0.93	0.99	0.93	0.71	0.94	0.89	0.68	0.94	0.77	0.94	0.91	0.81	
WRE	0.00	0.83	0.00	0.56	0.01	0.00	0.70	0.00	0.09	0.00	0.00	0.76	0.00	0.15	0.00	
BC	0.93	0.89	0.93	0.84	0.16	0.90	0.87	0.90	0.54	0.00	0.91	0.88	0.92	0.66	0.00	
DWE	0.93	0.90	0.93	0.82	0.96	0.92	0.87	0.91	0.84	0.02	0.93	0.88	0.92	0.83	0.04	
NG	0.94	0.90	0.93	0.95	0.88	0.93	0.88	0.93	0.95	0.80	0.93	0.89	0.93	0.95	0.84	
PS	0.93	0.90	0.93	0.94	0.95	0.93	0.86	0.93	0.94	0.78	0.93	0.88	0.93	0.94	0.86	
RI	0.93	0.90	0.93	0.71	0.12	0.93	0.86	0.91	0.48	0.00	0.93	0.88	0.92	0.57	0.00	XR
S	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
SP	N/A	N/A	N/A	N/A	N/A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
WE	0.94	0.90	0.93	0.95	0.82	0.93	0.88	0.93	0.95	0.49	0.93	0.90	0.93	0.95	0.61	
WS	0.94	0.90	0.93	0.94	0.95	0.93	0.87	0.93	0.94	0.80	0.93	0.89	0.93	0.94	0.87	
WRE	0.58	0.89	0.58	0.37	0.05	0.77	0.85	0.78	0.12	0.00	0.67	0.87	0.66	0.18	0.00	

Figure 2. Precision, Recall, and F_1-score achieved by all the features for all the PHI items in all of the reports in the form of a heatmap.

We repeated similar observations on each report-metric heatmap. The set of the best-performing features among all of the reports was consistent with the findings from IR heatmaps. While identifying the best features for a type of reports, we excluded any such feature that yielded the best value for one PHI item but performed poorly for multiple items, which was the case with discrete word embedding in the IR-Recall heatmap. We encountered similar cases in other heatmaps as well. For example, brown clustering had the highest value of precision for DATE in MA reports. Since it failed badly in identifying NAME and ID with a very poor value of all the metrics, it was discarded from the best feature set.

The performance of these models and their comparison are presented in Table 2. It can be seen from Table 2 that there is no significant difference between the performance of the best 4 features in combination and all 10 features in combination. For example, in IR reports, M10 had a precision value of 0.71 for LOCATION and M4 had a value of 0.72. Although M4 had a negligibly larger value in this case, there were multiple cases where M10 had larger values as can be seen with the recall value of DATE in IR reports. M10 had a recall value of 0.87 whereas M4 had a value of 0.86. Moreover, in XR reports, both the models had exactly the same values of metrics for each PHI item.

Table 2. Precision (P), Recall (R), and F_1-score (F1) values for all the reports achieved by the models with the best 4 features (M4) and all 10 features (M10).

Report Type		LOCATION		DATE		HOSPITAL		NAME		ID	
		M10	**M4**	**M10**	**M4**	**M10**	**M4**	**M10**	**M4**	**M10**	**M4**
IR	P	0.71	0.72	0.88	0.88	0.71	0.71	0.84	0.84	0.73	0.73
	R	0.68	0.69	0.87	0.86	0.70	0.71	0.78	0.78	0.60	0.61
	F1	0.70	0.70	0.87	0.87	0.71	0.71	0.80	0.80	0.66	0.66
MA	P	0.67	0.69	0.87	0.87	0.66	0.68	0.86	0.87	0.91	0.94
	R	0.60	0.64	0.82	0.82	0.63	0.66	0.81	0.81	0.91	0.91
	F1	0.63	0.67	0.84	0.85	0.64	0.67	0.83	0.84	0.91	0.92
MRI	P	0.82	0.83	0.88	0.89	0.83	0.83	0.91	0.91	0.86	0.86
	R	0.82	0.83	0.85	0.85	0.82	0.82	0.87	0.87	0.84	0.84
	F1	0.82	0.83	0,87	0.87	0.82	0.82	0.89	0.89	0.85	0.85
NM	P	0.82	0.81	0.90	0.90	0.83	0.82	0.91	0.92	0.88	0.89
	R	0.80	0.80	0.85	0.85	0.83	0.82	0.81	081	0.88	0.87
	F1	0.81	0.80	0.88	0.88	0.83	0.82	0.86	0.86	0.88	0.88
US	P	0.78	0.78	0.90	0.91	0.79	0.80	0.97	0.97	0.86	0.86
	R	0.77	0.77	0.87	0.87	0.78	0.78	0.95	0.95	0.86	0.85
	F1	0.77	0.78	0.89	0.89	0.78	0.79	0.96	0.96	0.86	0.86
CT	P	0.94	0.94	0.82	0.83	0.94	0.94	0.94	0.94	0.80	0.80
	R	0.93	0.93	0.75	0.75	0.94	0.94	0.93	0.93	0.71	0.71
	F1	0.94	0.94	0.79	0.79	0.94	0.94	0.94	0.94	0.75	0.75
XR	P	0.94	0.94	0.90	0.90	0.93	0.93	0.95	0.95	0.88	0.88
	R	0.93	0.93	0.88	0.88	0.93	0.93	0.95	0.95	0.80	0.80
	F1	0.93	0.93	0.89	0.89	0.93	0.93	0.95	0.95	0.84	0.84
ALL	P	0.90	0.90	0.88	0.88	0.90	0.90	0.93	0.93	0.86	0.87
	R	0.89	0.90	0.86	0.87	0.89	0.90	0.91	0.92	0.85	0.86
	F1	0.89	0.89	0.87	0.87	0.88	0.90	0.92	0.91	0.85	0.86

Figure 3 shows how the performance of the models varied with the number of training data. Here, only F_1-score was considered to evaluate the performance of a model. As expected, in most of the cases, the performance reached a saturation level after a number of training data points. Ideally, it was expected that the performance would initially be at a lower level, increase as the number of training data increased, and eventually reach a saturation level. However, there were cases where the performance deteriorated as the training data increased. For example, in MA, the F_1-score for LOCATION and HOSPITAL increased from 20 to 50 data but started to decrease after 50 data points.

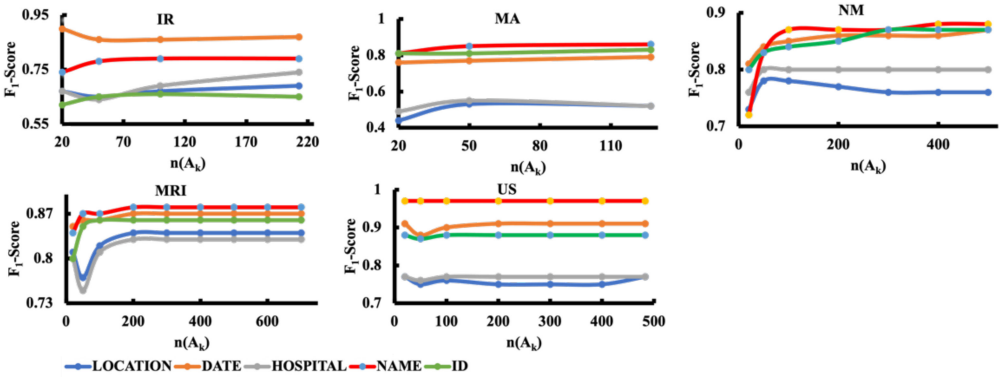

Figure 3. Variation of F_1-score with the size of the training data, $n(A_k)$.

4. Discussion

In this study, we investigated the utility of multiple features on a significantly large volume of clinical corpora. The dataset was enriched not only in terms of the total number of clinical reports but also in terms of variation. There were 7 types of radiology reports, and each type of report had a different structure of narrative from the other. Such a diverse dataset was vital for this study. If any feature performs well in one type of reports but fails in other types, it cannot be recognized as one of the best-performing features in a NER system. The best-performing features identified by our analysis performed consistently better than other features across all types of reports. Moreover, the large volume of the dataset made sure that the findings of our analysis are not subject to a small number of structured texts from where the feature learning was straightforward from the machine learning algorithm's point of view. This was the largest dataset among the existing clinical de-identification-related studies to the best of our knowledge.

The objective of this study was not to build a NER model with high values for performance metrics. Rather, the objective was to find the best features that can be used to build a high-performance model. The features play a significant role in a model's performance, but those are not the only factors that dictate a model's performance. The performance depends on the machine learning algorithm itself and its parameter estimation techniques. Since we used CLAMP to build the models, we restricted ourselves within the default setup of the tool and analyzed the models' performances using different features. The M4 model had precision, recall, and F_1-score of 0.78, 0.77, and 0.78, respectively (Table 2) for LOCATION in ultrasound reports which are not satisfactory values of performance metrics. However, such values do not prevent us from meeting the objective since we did not aim at a high-performance model. The best four features yielded the highest values of performance metrics than other features for each PHI no matter how high or low those values were, and thus the objective of identifying the best features was met.

We excluded the CT and XR reports from the second stage of analysis, where we observed how the performance of a model varies with the number of training data. The number of reports in each of these two types was too high to conduct a multifold training process with a variable number of training data. Out of the five types of reports included in this stage of analysis, both IR and MA had a small number of reports. On the other hand, NM, MRI, and US had a significantly larger number of reports, as can be seen from Table 1. Figure 3 shows that most of the PHIs reached a saturation level in NM, MRI, and US. This can be attributed to the large number of reports.

5. Conclusions

In this study, we investigated the efficiency of different features in a clinical de-identification NER model. A gold-standard dataset was created by manually annotating the PHI items in seven types of radiology reports. We built multiple CRF-based models by using 10 feature extraction modules in the CLAMP toolkit both individually and as combinations. By evaluating the performance of the features, we concluded that the best-

performing features were n-gram, prefix-suffix, word embedding, and word shape. These four features performed the best regardless of the type of reports involved. Moreover, these features outperformed the others even when all types of reports were used to build a single model. The dataset used in this study was rich in quantity and variation which made sure that the conclusion is not drawn by observing the performance on a small number of structured clinical texts. Since clinical reports form the foundation of numerous clinical research studies, demand for a reliable de-identification model is increasing day by day. This study contributes in identifying the best possible features for a machine-learning-based de-identification model out of a wide array of features at one's disposal.

Author Contributions: A.S.M.M., R.D.H., V.S. and M.B. conceptualized the project. M.K.Z.R. and V.M. setup the technology for data annotation. M.K.Z.R. and V.M. set up the technology for data annotation. The data was annotated by P.A.T., T.V.P.K., Y.J., N.S. and T.P. developed the de-identification models, evaluated their performances, and prepared the initial draft of the manuscript. H.I. enhanced the manuscript with her critical review. A.S.M.M. and V.M. managed the project. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by Roche Diagnostics Information Solutions.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of the University of Missouri (protocol code is "IRB #2013602 MU" and date of approval is 23 January 2019).

Informed Consent Statement: Patient consent was waived due to retrospective review of records that exist at that time and it is not possible to contact all patients and request their consent specifically for this project, especially since they have already been notified that their data might be used for secondary purposes.

Data Availability Statement: The data analyzed in this study is subject to the HIPAA Compliance. Requests to access these datasets should be directed to the corresponding author.

Conflicts of Interest: The study received funding from Roche Diagnostic Information Solutions. The funder designed the study and contributed to manuscript preparation. All authors declare no other competing interests.

References

1. Neamatullah, I.; Douglass, M.M.; Lehman, L.W.H.; Reisner, A.; Villarroel, M.; Long, W.J.; Szolovits, P.; Moody, G.B.; Mark, R.G.; Clifford, G.D. Automated De-Identification of Free-Text Medical Records. *BMC Med. Inform. Decis. Mak.* **2008**, *8*, 32. [CrossRef] [PubMed]
2. Department of Health and Human Services Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule; 2003; ISBN 2800228032. Available online: https://privacyruleandresearch.nih.gov/ (accessed on 7 April 2022).
3. Xia, H.; Rao, R. The Method of Medical Named Entity Recognition Based on Semantic Model and Improved SVM-KNN Algorithm. In Proceedings of the 7th International Conference on Semantics, Knowledge, and Grids, SKG 2011, Beijing, China, 24–26 October 2011.
4. Zhu, F.; Patumcharoenpol, P.; Zhang, C.; Yang, Y.; Chan, J.; Meechai, A.; Vongsangnak, W.; Shen, B. Biomedical Text Mining and Its Applications in Cancer Research. *J. Biomed. Inform.* **2013**, *46*, 200–211. [CrossRef] [PubMed]
5. Dehghan, A.; Keane, J.A.; Nenadic, G. Challenges in Clinical Named Entity Recognition for Decision Support. In Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013, Manchester, UK, 13–16 October 2013.
6. Saha, S.; Ekbal, A. Combining Multiple Classifiers Using Vote Based Classifier Ensemble Technique for Named Entity Recognition. *Data Knowl. Eng.* **2013**, *85*, 15–39. [CrossRef]
7. Nadeau, D.; Sekine, S. A Survey of Named Entity Recognition and Classification. Lingvisticae InvestigationesLingvisticæ InvestigationesLingvisticæ Investigationes. *Int. J. Linguist. Lang. Resour.* **2007**, *30*, 3–26. [CrossRef]
8. Goyal, A.; Gupta, V.; Kumar, M. Recent Named Entity Recognition and Classification Techniques: A Systematic Review. *Comput. Sci. Rev.* **2018**, *29*, 21–43. [CrossRef]
9. Grouin, C.; Zweigenbaum, P. Automatic De-Identification of French Clinical Records: Comparison of Rule-Based and Machine-Learning Approaches. In *Studies in Health Technology and Informatics*; IOS Press: Amsterdam, The Netherlands, 2013.
10. Jaćimović, J.; Krstev, C.; Jelovac, D. A Rule-Based System for Automatic de-Identification of Medical Narrative Texts. *Informatica* **2015**, *39*, 43–51. [CrossRef]
11. Shaalan, K. Rule-Based Approach in Arabic Natural Language Processing. *Int. J. Inf. Commun. Technol.* **2010**, *3*, 11–19.
12. Sil, A.; Yates, A. Re-Ranking for Joint Named-Entity Recognition and Linking. In Proceedings of the International Conference on Information and Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013.

13. Yoshida, K.; Tsujii, J. Reranking for Biomedical Named-Entity Recognition. In Proceedings of the ACL 2007-Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, 2007, Prague, Czech Republic, 29 June 2007.
14. Jiang, M.; Chen, Y.; Liu, M.; Rosenbloom, S.T.; Mani, S.; Denny, J.C.; Xu, H. A Study of Machine-Learning-Based Approaches to Extract Clinical Entities and Their Assertions from Discharge Summaries. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 601–606. [CrossRef] [PubMed]
15. Tang, B.; Cao, H.; Wang, X.; Chen, Q.; Xu, H. Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks. *Biomed. Res. Int.* **2014**, *2014*, 240403. [CrossRef] [PubMed]
16. Li, M.; Carrell, D.; Aberdeen, J.; Hirschman, L.; Malin, B.A. De-Identification of Clinical Narratives through Writing Complexity Measures. *Int. J. Med. Inform.* **2014**, *83*, 750–767. [CrossRef] [PubMed]
17. Tsochantaridis, I.; Joachims, T.; Hofmann, T.; Altun, Y. Large Margin Methods for Structured and Interdependent Output Variables. *J. Mach. Learn. Res.* **2005**, *6*, 1453–1484.
18. Lafferty, J.; Andrew, M.; Fernando, C.N.P. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the ICML '01: Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001. [CrossRef]
19. Wu, Y.; Jiang, M.; Xu, J.; Zhi, D.; Xu, H. Clinical Named Entity Recognition Using Deep Learning Models. *AMIA Annu. Symp. Proc.* **2017**, *2017*, 1812–1819. [PubMed]
20. Aberdeen, J.; Bayer, S.; Yeniterzi, R.; Wellner, B.; Clark, C.; Hanauer, D.; Malin, B.; Hirschman, L. The MITRE Identification Scrubber Toolkit: Design, Training, and Assessment. *Int. J. Med. Inform.* **2010**, *79*, 849–859. [CrossRef] [PubMed]
21. Soysal, E.; Wang, J.; Jiang, M.; Wu, Y.; Pakhomov, S.; Liu, H.; Xu, H. CLAMP—A Toolkit for Efficiently Building Customized Clinical Natural Language Processing Pipelines. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 331–336. [CrossRef] [PubMed]
22. Turian, J.; Ratinov, L.; Bengio, Y. Word Representations: A Simple and General Method for Semi-Supervised Learning. In Proceedings of the ACL 2010-48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, 11–16 July 2010.
23. Chaudhuri, B.B.; Bhattacharya, S. An Experiment on Automatic Detection of Named Entities in Bangla. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages, Hyderabad, India, 12 January 2008; pp. 75–82.
24. Sandin, F.; Emruli, B.; Sahlgren, M. Random Indexing of Multidimensional Data. *Knowl. Inf. Syst.* **2017**, *52*, 267–290. [CrossRef]
25. Zhang, X.; Zhao, J.; Lecun, Y. Character-Level Convolutional Networks for Text Classification. *Adv. Neural Inf. Processing Syst.* **2015**, *28*, 649–657.