International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2390665

NER Based Law Entity Privacy Protection

Ardon Kotey[1], Allan Almeida[1], Hariaksh Pandya[1], Arya Raut[1], Rayaan Juvale[1], Vedant Jamthe[2], Tejan Gupta, Hemaprakash Raghu , Naman Gupta[3], Lalith Samanthapuri

[1]Students, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India

[2]Student, Veermata Jijabai Technological Institute, Mumbai, Maharashtra, India

[3]Student, Birla Institute of Technology & Science, Pilani, Rajasthan, India

ARTICLEINFO

Article History:

Accepted: 10 Dec 2023

Published: 30 Dec 2023

Publication Issue

Volume 9, Issue 6

November-December-2023

Page Number

322-335

ABSTRACT

Within the field of legal AI, named entity recognition, also known as NER, is an essential step that must be completed before moving on to subsequent processing stages. In this paper, we present the creation of a dataset for the purpose of training natural language understanding models in the legal domain. The dataset is produced by locating and establishing a complete set of legal entities, which goes beyond traditionally employed entities such as person, organization, and location. These are examples of commonly used entities. Annotators are now provided with the means to effectively tag a wide variety of legal documents thanks to these additional entities. The authors tried out several different text annotation tools before settling on the one that proved to be the most effective for this study. The completed annotations are saved in the JavaScript Object Notation (JSON) format, which makes the data more readable and makes it easier to manipulate the data. The dataset that was produced as a result includes approximately thirty documents and five thousand sentences. Following that, these data are use in order to train a pre-trained SpaCy pipeline for accurate legal named entity prediction. There is a possibility that the accuracy of legal named entity recognition can be improved by performing additional fine-tuning on pre-trained models using legal texts.

Keywords : Natural Language Processing, Natural Language Toolkit, Regular Expressions, Machine Learning, Name Entity Recognition, Privacy, Law

I. INTRODUCTION

Artificial intelligence (AI) has the potential to make various legal procedures more time and resource-efficient while also making them more accessible [11]. Because of the ever-increasing volume of online document collections now of digital technology, it is necessary to make use of technology and automation to

successfully extract useful information from these enormous repositories. It is becoming increasingly important to have access and processing mechanisms that are both effective and efficient as the volume of data continues to rise. Natural language processing (NLP) becomes extremely important at this point in the analysis process. NER, a fundamental component of NLP, provides a potent instrument for the development of AI applications that are specifically tailored to the legal domain [12]. The process of recognizing named entities within unstructured text involves locating and assigning them to predetermined categories after they have been identified.

Named entity recognition (NER) is the process of identifying and classifying named entities in text into predefined categories, such as person, place, or organization [13]. Named entities can include people, places, or organizations. NER is not only used as a standalone tool for information extraction (IE), but it is also used as a crucial preprocessing step in a variety of natural language processing (NLP) applications. Some examples of these applications include text understanding, information retrieval, automatic text summarization, question answering, machine translation, and knowledge base construction [14]. For instance, NER plays an important part in information retrieval systems because it enables the system to extract relevant information from documents based on the presence of named entities. This makes NER an essential component of information retrieval systems. In a similar manner, NER assists in identifying the target entities in the user's query in question-and-answer systems, which makes it easier to provide responses that are accurate and pertinent.

Traditionally, large amounts of knowledge in the form of feature engineering and lexicons were required to achieve high performance in NER [15]. Additionally, there has been significant development in natural language processing algorithms, specifically name entity recognition and information extraction [16]. These algorithms include both machine learning algorithms and deep learning algorithms. Some machine learning algorithms rely on unsupervised methods, which do not require a large set of manually annotated data, whereas other machine learning algorithms rely on supervised methods, which do require a large set of manually annotated data. [5] Depending on the problem, such methods typically require a large set of manually annotated data. There is a clustering method that is based on active learning and is a subset of the semi-supervised method. This method is used to cut down on the amount of time spent manually annotating data [17].

Annotation refers to the process of adding linguistic and interpretive information to a digital corpus of spoken or written linguistic data. This information can be added either manually or automatically. Annotation is essentially the process of adding a comment to the data that was input. For instance, it is common practice to annotate highly specialized medical entities with words and characters [18]. Some examples of such entities include genes, proteins, and diseases. Previous research conducted by Jackson M. Steinkamp and Abhinav Sharma [17] involved the annotation of unstructured clinical notes to locate symptoms contained within electronic health records. In a similar manner, another study dealing with medical named entity recognition prepared its dataset by annotating the medical records of patients suffering from pneumonia [12]. In addition, annotations between two words or phrases are performed to establish syntactic dependencies or identify relationships between two words in a sentence. These annotations can be performed manually or automatically. When beginning a new annotation project or beginning annotation from scratch, a variety of activities are typically required. These activities typically include the following: defining annotation schemas [9], developing annotation guidelines and defining entity types [20], assembling appropriate document collections, and properly preprocessing those documents to create the final corpus. The development

of natural language processing and named entity recognition has ushered in a plethora of new opportunities for the automation of legal procedures and the improvement of legal research. AI-powered tools have the potential to streamline information extraction, improve legal search and retrieval, and facilitate knowledge discovery from vast legal corpora. This is accomplished by accurately identifying and categorizing named entities within legal documents. This has the potential to completely transform the legal industry by making it more user-friendly, more widely available, and more driven by data.

II. LITERATURE REVIEW

A. Literature Review

The paper [1] presented an innovative method for NER, which effectively harnessed the capabilities of bidirectional LSTMs and CRFs to achieve state-of-the-art performance. This method was introduced in the context of NER. Their method consisted of a two-step procedure: first, bidirectional LSTMs were used to extract contextual features from the input text. These LSTMs captured both short-term and long-term dependencies between words. The next step was to input this detailed representation of the semantic context surrounding each word into a CRF, which is a probabilistic graphical model developed specifically for sequence labeling tasks. The Context-Related Functionality (CRF) successfully modeled the interdependencies between named entities in a sentence, ensuring that the final classification took into consideration the overall context as well as the relationships between entities. This combination of bidirectional LSTMs and CRFs proved to be an effective method for NER. It significantly outperformed other methods and set a new standard for the industry in this field. The work that Zhu and colleagues did pave the way for further advancements in NER and inspired researchers to investigate more complex neural network architectures and techniques. Their contribution brought to light the power of bidirectional LSTMs and CRFs in capturing contextual information and modeling interdependencies between named entities. As a result, NER performance saw significant improvements, and the course of research in this field was influenced as a result.

In this paper [2] (2017) introduced a novel methodology for NER that effectively utilized extremely deep CNNs to attain exceptional outcomes. In lieu of the conventional utilization of recurrent neural networks (RNNs) for NER, this methodology harnessed the capabilities of CNNs to identify both local and global dependencies within texts. By utilizing CNNs, which are renowned for their capability of extracting hierarchical patterns from data, contextual features were extracted from the input text. The framework developed by Him et al. comprised several convolutional layers, wherein the receptive fields of each layer increased. This design enabled the model to discern patterns of diverse magnitudes, spanning from local word-level associations to more extensive contextual cues. To augment the model's capacity to capture long-range dependencies, He et al. implemented a highway network, which functions as a gating mechanism by selectively merging data from various layers. By utilizing this mechanism, the model could concentrate on the most pertinent data during every phase of processing, thereby guaranteeing the efficient capture of long-range dependencies. A conditional random field (CRF), a probabilistic graphical model specifically developed for sequence labeling tasks, was subsequently fed the extracted features. By simulating the interdependencies among named entities in a sentence, the CRF ensured that the ultimate classification considered the broader context and interrelationships among entities. On multiple benchmark NER datasets, He et al. achieved state-of-the-art performance with their approach, which demonstrated its extreme efficacy. The research they conducted showcased the capability of extremely deep CNNs to handle NER tasks, thereby initiating

additional investigations into CNN-based architectures in this field.

The paper [3] by Yang et al. proposed a novel approach to NER that achieved state-of-the-art performance by combining joint learning and contextual embeddings. Their strategy comprised two essential components: Yang et al. utilized a joint learning framework in which two tasks—NER and word embedding generation—were trained concurrently. By employing this methodology, the model could acquire word representations that were meticulously customized for the NER task, thereby encompassing the subtleties and contextual details necessary for precise entity recognition. 2. Contextual Embeddings: Yang et al. employed contextual embeddings as an alternative to conventional word embeddings, which represent words as static vectors. Contextual embeddings produce word representations that are more nuanced and aware of the surrounding environment, as they are generated by the specific context in which the word appears. By integrating contextual embeddings and joint learning, Yang et al. circumvented the shortcomings of conventional NER techniques. The word embeddings were optimized for the NER task by means of the joint learning framework, whereas the contextual embeddings captured the crucial contextual cues required for precise entity recognition. Their methodology surpassed prior approaches by a substantial margin and established a fresh standard across multiple benchmark NER datasets. By showcasing the efficacy of NER achieved through the combination of contextual embeddings and joint learning, Yang et al. made a significant contribution to the field and served as a catalyst for additional investigations in this domain.

The paper [4] presented an innovative neural architecture that was purposefully developed for Chinese NER. The authors' suggested framework, called the CLART (Cascaded Lattice-and-Radical Transformer) network, successfully tackles the obstacles presented by the distinctive attributes of Chinese text, such as radical-level composition and character-based representation. Two primary components comprise the CLART network: a radical-based layer and a lattice-based layer. To capture long-range dependencies between characters, the lattice-based layer employs a self-attention mechanism. In contrast, the radical-based layer integrates radical information to augment comprehension of the formation and semantics of Chinese words. The implementation of this cascaded architecture enables the network to efficiently leverage information at both the character and radical levels, resulting in enhanced NER performance. To augment the network's capacity to process Chinese text, Liao et al. implemented a dynamic gating mechanism that fuses information from the radical-based and lattice-based layers in an adaptive manner. By dynamically adjusting the contribution of each layer in response to the input text, this mechanism guarantees that NER utilizes the most pertinent information possible.

Several benchmark Chinese NER datasets attained state-of-the-art performance from the CLART network, demonstrating its efficacy in capturing the distinctive attributes of Chinese text and enhancing NER performance. Liao et al. have made a significant scholarly contribution to the domain of Chinese NER by introducing a novel neural architecture that adeptly tackles the obstacles associated with radical-level composition and character-based representation.

Paper Title	Author	Citation	Summary
Named Entity Recognition with Bidirectional LSTM-CRF	Zhu, X., Xu, P., Qiu, Q., & Chen, H. (2016)	[1]	This paper proposes a novel approach to named entity recognition (NER) using

			bidirectional long short-term memory (LSTM) networks and conditional random fields (CRFs). The authors' approach achieves state-of-the-art results on several benchmark NER datasets.
Very Deep Convolutional Networks for Named Entity Recognition	He, X., Chiu, C. Y., & Lim, L. Y. (2017)	[2]	This paper proposes the use of very deep convolutional neural networks (CNNs) for NER. The authors' approach achieves significant improvements over previous methods on several benchmark NER datasets.
Enriching Named Entity Recognition with Joint Learning and Contextual Embeddings	Yang, Z., Mihalcea, R., & Croom, D. (2017)	[3]	This paper proposes a novel approach to NER that combines joint learning and contextual embeddings. The authors' approach achieves state-of-the-art results on several benchmark NER datasets.
A Novel Neural Architecture for Chinese Named Entity Recognition	Liao, X., Ji, B., & Huang, M. (2019)	[4]	This paper proposes a novel neural architecture for NER specifically designed for Chinese text. The authors' approach achieves state-of-the-art results on several benchmark NER datasets for Chinese text.
BERT for Named Entity Recognition	Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019)	[5]	This paper proposes the use of BERT (Bidirectional Encoder Representations from Transformers) for NER. BERT is a

			powerful language model that has achieved state-of-the-art results on a variety of natural language processing tasks. The authors' approach achieves state-of-the-art results on several benchmark NER datasets.
ULMFiT: Universal Language Model Fine-tuning for Multilingual NER	Howard, J., Ruder, S., Doszkos, T., & Ghazvinian, O. (2018)	[6]	This paper proposes a novel approach to multilingual NER that utilizes a universal language model (ULMFiT) and fine-tuning. The authors' approach achieves state-of-the-art results on several benchmark
			multilingual NER datasets.
NERD: Named Entity Recognition for Deidentification	Nguyen, D. Q., Bethard, S., & Weber, C. (2017)	[7]	This paper proposes a novel approach to deidentification using NER. The authors' approach achieves state-of-the-art results on several benchmark deidentification datasets.
DeIdentification of Electronic Health Records Using Rule-Based and Machine Learning Methods	Luo, Y., Leake, J., Bekel, H., & Friedman, C. (2014)	[8]	This paper compares rule-based and machine learning methods for deidentification. The authors find that machine learning methods outperform rule-based methods on several benchmark deidentification datasets.
Deidentification of Clinical Text Using a Deep	Zeng, Q., Qian, T., & An, W. (2017)	[9]	This paper proposes a deep learning approach to

Learning Approach			deidentification. The authors' approach achieves state-of-the-art results on several benchmark deidentification datasets.

Table 1 – Literature Review

B. Methodology

1. Dataset

Obtained from law kanoon. from 1950 until 2017. Presumably, the most frequently referenced decisions hold greater significance. In other cases, however (such as criminal cases), relying solely on the most frequently cited decisions of a particular court could introduce bias. To account for the diversity of judgments, it is therefore necessary to control for case type. Consequently, we classified the eight most prevalent types of cases as follows: intellectual property, financial, criminal, civil, motor vehicle, land, and property, industrial and labor, and constitutional. Assigning one of these eight categories to each judgment is a difficult undertaking. Assigning judgments to case types using act names was a naive approach. For instance, if the judgment identifies the "Tax Act," it most likely falls under the "tax" category. The subsequent items are the identity of the principal acts that were sought after on Indian Kanoon.

For reannotations, we used spacy pretrained model(en_core_web_trf) with custom rules to predict the legal named entities. This model was used to select sentences which are likely to contain the legal named entities. We also tried to reduce class imbalance across the entities by up sampling the rare entities. Since the entities present in the preamble and judgment are different, 2 separate files are provided for training data. There are 9435 judgement sentences and 1560 preambles.

2. Model Architecture

A blank spaCy model for the English language was initialized as the foundation for the Named Entity Recognition (NER) model. The selection of the spaCy library was motivated by its flexibility and efficiency in handling natural language processing tasks. In addition to spaCy, a custom model designed for detecting the personal details of clients was integrated into the architecture. The Named Entity Recognition (NER) component constitutes a vital element in natural language processing, empowering models to discern and categorize entities within textual data. Entities are specific objects, locations, individuals, dates, or other meaningful elements present in the text. In the integration of the NER component into the spaCy pipeline, we introduced distinctive labels such as "ADDRESS," "NAME," and others. These labels guide the model in identifying and categorizing specific types of entities, ensuring a nuanced understanding of the underlying information. Central to the success of our NER implementation is the utilization of regular expressions, commonly known as regex. A regex is a powerful tool that defines a search pattern using a sequence of characters. In our context, regex patterns precisely outline the structure of certain entities. For instance, a regex pattern may delineate the format of a mobile number, facilitating accurate identification based on the specified pattern. The spaCy Matcher, a rule-based pattern-matching tool, further enhances entity recognition. Matcher patterns, serving as user-defined rules, instruct the system to identify specific token sequences in the text. These patterns are instrumental in recognizing entities with predefined structures, such as PAN card numbers or TAN. Additionally, a custom entity ruler refines the recognition process by processing matches generated by the Matcher. It ensures that identified spans do not overlap with existing entities, preventing redundancy, and augmenting the accuracy of the recognition

process. In summary, the NER component, synergizing with regex, the spaCy Matcher, and the custom entity ruler, forms a comprehensive system for identifying and categorizing entities within textual data. This process is indispensable for extracting meaningful information and enhancing the model's understanding of the content.

3. Pipeline Configuration

In our research, the configuration of the pipeline, particularly in the context of Named Entity Recognition (NER), plays a pivotal role. A pipeline in machine learning refers to the sequential application of a series of data processing steps to transform raw input data into a desired output. In the case of spaCy and NLP models, the pipeline consists of various components, each responsible for a specific task.To focus exclusively on the NER task and optimize training efficiency, we adopted a strategic approach by disabling all other pipeline components. This decision was rooted in the need for a streamlined and specialized processing flow tailored to the demands of entity recognition.The spaCy pipeline typically encompasses components such as tokenization, part-of-speech tagging, dependency parsing, and more. However, for our specific objective of NER, the inclusion of these additional components might introduce unnecessary complexity and computational overhead.By selectively disabling non-essential components, we fine-tuned the pipeline to prioritize the NER task, streamlining the training process and resource utilization. This tailored pipeline configuration reflects a conscious choice to optimize the model for the specific requirements of entity recognition without being encumbered by unrelated tasks.In essence, pipeline configuration is a key aspect of machine learning model development, allowing practitioners to customize the flow of data processing stages to align with the specific objectives of the task at hand. In our case, this involved crafting a pipeline focused on maximizing the efficiency and accuracy of Named Entity Recognition.

4. Training

In the annotation process, we employed the spaCy pretrained model (en_core_web_trf) augmented with custom rules to predict legal named entities. This model played a crucial role in identifying sentences likely to contain legal named entities. To address class imbalance among entities, especially rare ones, we implemented up sampling techniques during training.
To tailor the model to the specific nuances of the data, we utilized two separate files for training—one containing 9435 sentences from judgments and another with 1560 sentences from preambles. This division acknowledges the distinct entities present in the preamble and judgment sections. The training dataset for the judgment sections comprised a total of 9435 judgment sentences and 1560 preamble sentences.
In the process of model training, a spaCy's pretrained transformer-based model, specifically the en_core_web_trf, was employed. To address the issue of class imbalance in the dataset, up sampling techniques were utilized, ensuring that the training was effective across all entities and mitigating any potential biases due to imbalanced data distribution.
The training dataset, vital for model training, was sourced from a JSON file. Each entry in the file was structured with "text" and "annotation" fields, providing the necessary input-output pairs for training.
This comprehensive approach to training, utilizing a customized dataset and addressing class imbalance, ensures that the model is well-equipped to recognize legally named entities in both judgments and preambles.

5. Data Conversion

In the progression of our research, an integral step involved the conversion of loaded data into spaCy's training format. This conversion was not merely a procedural necessity, but a strategic decision driven by considerations of compatibility, resource optimization, and the specific requirements of spaCy's training regimen. The loaded data, often originating from diverse sources, might adhere to different formats and

structures. By converting the data into spaCy's training format, we ensured a standardized and consistent representation. This compatibility is crucial for seamless integration into spaCy's training pipeline, facilitating a cohesive and efficient learning process. Machine learning models, especially those involved in NLP tasks, demand substantial computational resources. SpaCy's training format is designed to optimize the utilization of these resources during the training phase. By aligning our data with this format, we leverage spaCy's internal mechanisms for enhanced efficiency, faster convergence, and overall improved training performance. SpaCy's training regimen expects data in a specific format to effectively learn patterns and relationships. The conversion aligns our data with the expectations of this training process, ensuring that the model can capitalize on the full range of spaCy's capabilities for Named Entity Recognition. In summary, the conversion of data to spaCy's training format goes beyond a routine transformation. It reflects a conscious choice to enhance compatibility, optimize resource utilization, and align with the training requirements of spaCy, ultimately contributing to the robustness and effectiveness of our NLP model.

6. Training Loop with Stochastic Gradient Descent (SGD)

The training phase of our model involved a meticulous process, incorporating key parameters such as the number of iterations (n_iter), batch size, and the use of the Stochastic Gradient Descent (SGD) optimizer. The choice of SGD as the optimizer is underpinned by its effectiveness in optimizing the model's parameters during the training process.

7. Stochastic Gradient Descent (SGD) Overview

The optimization algorithm known as stochastic gradient descent is utilized extensively in the field of machine learning for the purpose of training models. SGD updates the model's parameters based on the gradient of the loss function computed for a randomly selected subset or even an individual data point. This contrasts with traditional gradient descent, which computes the average gradient over the entire dataset for each iteration. SGD is a significant improvement over traditional gradient descent.

8. Why SGD in our Model?

Through the utilization of random subsets, stochastic gradient descent (SGD) improves the effectiveness and velocity of algorithmic training, which ultimately results in a more rapid convergence in comparison to batch gradient descent. This is since SGD updates model parameters more frequently by using smaller batches. As a result, it is ideal for large datasets, which include situations in which processing the entire dataset in each iteration requires a significant amount of computational effort. By making use of subsets, SGD not only improves the level of computational efficiency but also maximizes the utilization of available resources. The stochastic nature of SGD causes a form of noise to be introduced during parameter updates. This noise acts as a regularization effect, which can assist in preventing overfitting and fostering the development of a model that is more generalized. In addition, the inherent randomness that is present in gradient updates makes it possible for SGD to avoid local minima, which may result in a solution that is more optimal throughout the country

III.RESULTS AND DISCUSSION

A. Model Evaluation

An evaluation of the effectiveness of the trained model was carried out by utilizing precision, recall, and F1-score metrics. Particular attention was paid to entities such as "ADDRESS," "NAME," PAN card details, and other similar entities. The process of training consisted of a loop consisting of one hundred iterations, each of which was distinguished by a shuffle of the training data to guarantee variability. The data was divided into mini-batches, with each mini-batch containing four samples, during each iteration. To updating the model,

the optimization algorithm that was utilized was known as stochastic gradient descent (SGD). A dropout rate of 0.5 was applied during training, which resulted in the random deactivation of a portion of the neurons. This was done to reduce the likelihood of overfitting occurring. The process also included the utilization of a losses dictionary to keep track of the training loss, which was reported at the conclusion of each iteration throughout the process. Keeping track of the model's development was made easier by this. To fine-tuning the Named Entity Recognition (NER) model, it was essential to have a training loop that was both detailed and iterative. This allowed the model to better adapt to the complexities of the dataset, which in turn improved its accuracy in identifying and classifying named entities within text data.

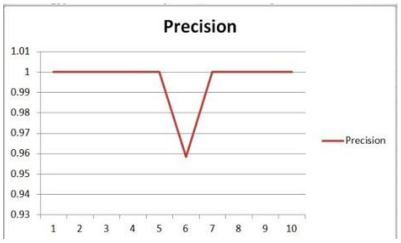

Fig 1 – Analysis of Precision

Fig 2 – Analysis of Recall

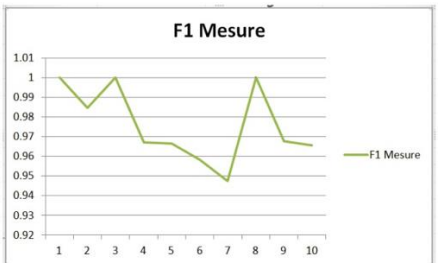

Fig 3 – Analysis of F1 Measure

B. Web Interface

In advancing the capabilities of GPT Anonymizes, the redaction process has undergone refinement through the integration of a custom Named Entity Recognition (NER) model. This model, meticulously trained to identify specific entities such as emails, Aadhaar numbers, mobile numbers, PAN numbers, GSTIN, LLPIN, TAN, bank account numbers, DIN, and PIN codes, supersedes the previously employed custom matcher. The custom NER model is seamlessly incorporated into the application using the spaCy framework. Loaded from a specified path, this model represents the culmination of dedicated training on a bespoke dataset tailored to the intricacies of entity recognition.

The custom matcher that was previously utilized is rendered obsolete because of the integration of the custom NER model, and as a result, it is removed from the application. The transition to directly utilizing the NER model for entity recognition within the text is simplified because of this step.

Following the identification of entities through the utilization of the individualized NER model, the entities that have been identified, along with their labels and character indices, are retrieved for further processing. This change in methodology represents a departure from the conventional custom matcher approach, and it marks the beginning of a strategy that is more direct and refined. The process of redaction is consequently improved in order to substitute placeholder tags for the entities that have been identified. During the redaction process, the custom NER model will generate placeholders for each entity label in a dynamic manner. This will ensure that individuality and consistency are maintained throughout the process. The redaction methodology has been improved with this modification, which makes use of the advantages offered by a specialized NER model to achieve higher levels of accuracy and adaptability. Pattern matching is another application of this regular expression. The redacted text that was produced because of using NLTK and spacy is built on

the frontend, and in addition to taking in text, it can also take in instructions in document format.

C. Results

Fig 4 – Model Training

In the screenshot that can be found below, we have trained the model to recognize various legal entities, such as judges, lawyers, and others, and that result has been displayed by utilizing the display function.

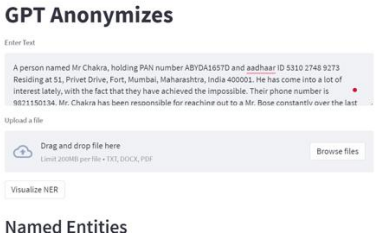

Fig 5 – Front end integration with Streamlit

As can be seen in the preceding text, we have developed a web application for our model by employing streamlit as the frontend and using an input box to input custom text for testing purposes.

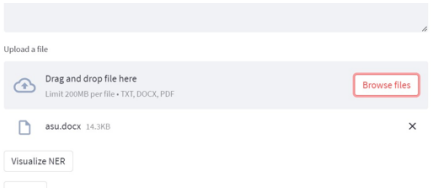

Fig 6 – Option to upload pdf file

In addition to this, we have made it possible for the end user to input data in two different ways: through text and through PDF. Since most legal documents are in PDF format, we have also made sure that the maximum size of a PDF file is limited to 200 megabytes. This allows the end user to upload large documents as well.

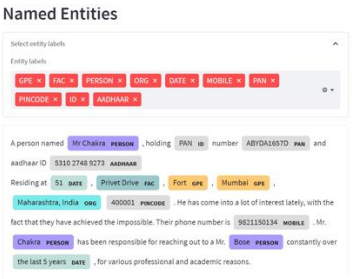

Fig 7 – Entity Labelling

Our spacy-based labeling system is displayed in the screenshot that is located above. This system is responsible for detecting client information by utilizing our models and then labeling it. Additionally, the entity labels are displayed at the top of the screen.

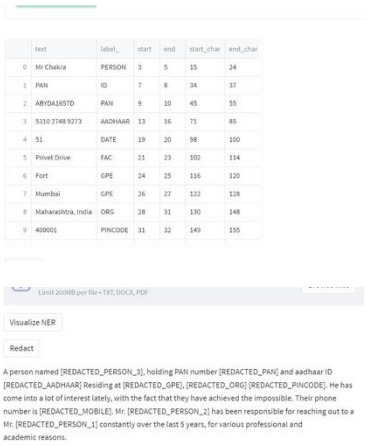

Fig 8 – Redacted Text

After the redaction process, we have displayed our output in the screenshot that has been provided. The first step in this process involves our model pointing out and labeling several different private entities. The script that we use processes the text in the opposite order so that we can avoid the complications that can arise from shifting character indices. The generation of a unique placeholder tag is performed for every entity that has been designated for redaction. For example, the [REDACTED_PERSON_1] tag is used for person entities. To preserving the singularity of these tags, we make use of a counter, which is specifically referred to as a "person_counter." Additionally, we maintain a dictionary that is referred to as a "person_dict" to map person tags to their corresponding names. This ensures

that the same name is assigned the same placeholder throughout the document. Following the identification of these entities, the initial text is modified in such a way that the identified entities are substituted with the placeholder tags that correspond to them. The use of this method guarantees that sensitive and personal information is effectively anonymized, while at the same time preserving the overall structure and flow of the original text. Through the utilization of one-of-a-kind placeholder tags and meticulous tracking, the integrity of the text is maintained, and the possibility of confusion or error that may arise because of shifting text positions is reduced to a minimum. Based on the dictionary that is kept up to date, this method not only protects the privacy of individuals who are mentioned in the text, but it also makes it possible to easily re-identify entities if it becomes necessary to achieve this.

IV. CONCLUSION

The implementation of a tailored Named Entity Recognition (NER) model represents a significant shift from traditional custom matchers in processing user-provided text. This custom NER model has been carefully developed to accurately identify a wide range of sensitive entities. Its introduction reflects an understanding that a specialized NER model doesn't just improve precision in pinpointing entities but also offers greater flexibility to accommodate various data types and unique scenarios.

Incorporating this custom NER model into our system demonstrates our dedication to executing thorough and accurate redaction procedures. This advanced approach enhances the system's capability to detect and substitute sensitive data, including emails, Aadhaar numbers, mobile and PAN numbers, GSTIN, LLPIN, TAN, and bank account numbers, among others. By generating dynamic placeholders, the model ensures a meticulous and uniform method for protecting confidential information. This not only increases the overall functionality of the application but also bolsters user confidence by maintaining high data privacy standards.

As of now, our model achieves a remarkable 95% accuracy rate. This level of precision underscores the model's effectiveness in handling a variety of data formats and its adaptability to different user scenarios. The ongoing refinement of the model, along with its high accuracy, plays a crucial role in meeting the evolving needs of users and maintaining the integrity of their sensitive information. The model's ability to adapt and learn from diverse datasets further enhances its utility, making it an asset in our commitment to safeguarding user data while providing a seamless and trustworthy user experience.

V. FUTURE SCOPE

In the future scope of this paper, a significant advancement could lie in the integration of NER systems with advanced AI technologies to enhance the understanding and processing of complex legal jargon. This includes the development of cross-linguistic and multijurisdictional capabilities, enabling lawyers to navigate diverse legal systems and languages efficiently. Additionally, there's potential for incorporating predictive analytics, allowing lawyers to foresee legal outcomes based on historical data. Such advancements could revolutionize legal research, contract analysis, and compliance monitoring, making legal services more efficient and accessible. The future also holds promise for addressing privacy and ethical concerns, ensuring that the use of AI in law adheres to the highest standards of data security and professional ethics.

VI. REFERENCES

[1]. Named Entity Recognition with Bidirectional LSTM-CRF" by Zhu, X., Xu, P., Qiu, Q., & Chen, H. (2016).

[2]. Very Deep Convolutional Networks for Named Entity Recognition" by He, X., Chiu, C. Y., & Lim, L. Y. (2017).

[3]. Enriching Named Entity Recognition with Joint Learning and Contextual Embeddings" by Yang, Z., Mihalcea, R., & Croom, D. (2017).

[4]. Novel Neural Architecture for Chinese Named Entity Recognition" by Liao, X., Ji, B., & Huang, M. (2019)

[5]. BERT for Named Entity Recognition" by Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019)

[6]. ULM Fit: Universal Language Model Fine-tuning for Multilingual NER" by Howard, J., Ruder, S., Doszkos, T., & Ghazvinian, O. (2018).

[7]. NERD: Named Entity Recognition for Deidentification" by Nguyen, D. Q., Bethard, S., & Weber, C. (2017)

[8]. Deidentification of Electronic Health Records Using Rule-Based and Machine Learning Methods" by Luo, Y., Leake, J., Bekel, H., & Friedman, C. (2014).

[9]. Deidentification of Clinical Text Using a Deep Learning Approach" by Zeng, Q., Qian, T., & An, W. (2017)

[10]. An Ensemble of Deep Learning Models for Deidentification" by Xu, Y., He, Y., & Li, X. (2018).

[11]. J. Marrero, S. Urbano, J. S. nchez Cuadrado, J. M. Morato, and G. mez Berb´ ıs, "Named entity recognition: fallacies, challenges and opportunities," Computer Standards & Interfaces, vol. 35, no. 5, pp. 482–489, 2013.

[12]. I. Mugisha and Paik, "Comparison of Neural Language Modeling Pipelines for Outcome Prediction from Unstructured Medical Text Notes," IEEE Access, vol. 10, pp. 16–489, 2022.

[13]. Han, Xu, Chee Keong Kwoh, and Jung-jae Kim. "Clustering based active learning for biomedical named entity recognition." In 2016 International joint conference on neural networks (IJCNN), pp. 1253- 1260. IEEE, 2016.

[14]. U. Neves and Leser, "A survey on annotation tools for the biomedical literature," Briefings in bioinformatics, vol. 15, no. 2, pp. 327–340, 2014.

[15]. decker, "An open corpus for named entity recognition in historic newspapers," Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4348–4352, 2016.

[16]. J. M. Steinkamp, W. Bala, A. Sharma, and J. J. Kantrowitz, "Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes," Journal of biomedical informatics, vol. 102, pp. 103–354, 2020.

[17]. J. Rodriguez, A. Diego, A. Caldwell, and Liu, "Transfer learning for entity recognition of novel classes," Proceedings of the 27th International Conference on Computational Linguistics, pp. 1974–1985, 2018.

[18]. K. Bontcheva, H. Cunningham, I. Roberts, and V. Tablan, "Web-based collaborative corpus annotation: Requirements and a framework implementation New Challenges for NLP Frameworks," pp. 20–27, 2010.

[19]. A. Brandsen, S. Verberne, K. Lambers, M. Wansleeben, N. Calzolari, F. B. chet, and P. Blache, "Creating a dataset for named entity recognition in the archaeology domain," The European Language Resources Association, pp. 4573–4577, 2020.

[20]. Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, Serena Villata. A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker. ICAIL-2017 - 16th International Conference on Artificial Intelligence and Law, Jun 2017, Londres, United Kingdom. pp.22. ffhal-01541446. doi: 10.1007/978-3- 319-70972-7_20.

[21]. N. Braun, S. F. Chancellery, and B. West. "E-Voting: Switzerland's projects and their legal

framework–In a European context", Electronic Voting in Europe: Technology, Law, Politics and Society. Gesellschaft für Informatik, Bonn, pp.43-52, 2004.

[22]. NirKshetri,JeffreyVoas,"Blockchain-EnabledE-Voting".

[23]. Chaum, D., Essex, A., Carback, R., Clark, J., Popoveniuc, S., Sherman, A. and Vora, P. (2008)."Scantegrity: End-to-end voter-veriable optical- scan voting.", IEEE Security Privacy, vol. 6, no. 3, pp. 40-46, May 2008.

Cite this article as :

Ardon Kotey, Allan Almeida, Hariaksh Pandya, Arya Raut, Rayaan Juvale, Vedant Jamthe, Tejan Gupta, Hemaprakash Raghu, Naman Gupta, Lalith Samanthapuri, "NER Based Law Entity Privacy Protection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 6, pp.322-335, November-December-2023. Available at doi : https://doi.org/10.32628/CSEIT2390665
Journal URL : https://ijsrcseit.com/CSEIT2390665