Savitribai Phule Pune University

Modern Education Society's Wadia College of Engineering, Pune

19, Bund Garden, V.K. Joag Path, Pune – 411001.

ACCREDITED BY NBA AND NAAC WITH 'A++' GRADE

DEPARTMENT OF COMPUTER ENGINEERING

A

SEMINAR REPORT

on

DATA REDACTION USING NAMED ENTITY RECOGNITION

T.E. (Computer)

SUBMITTED BY

EKTA CHAUGULE

GUIDED BY

PROF. A. A. SHAHBAD

(Academic Year: 2024-25)

Savitribai Phule Pune University

Modern Education Society's Wadia College of Engineering, Pune

19, Bund Garden, V.K. Joag Path, Pune – 411001.

ACCREDITED BY NBA AND NAAC WITH 'A++' GRADE

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that

Miss. Ekta Chaugule

has been completed Seminar entitled

DATA REDACTION USING NAMED ENTITY RECOGNITION

As a partial fulfillment of the Third Year of Bachelor degree in "Computer Engineering" as prescribed by the Savitribai Phule Pune University in TE COMP Second Shift the Semester - I of academic year 2024-25.

Guide Name	Dr. N. F. Shaikh
PROF. A.A.SHAHBAD	**HEAD OF DEPARTMENT**

Place: Pune

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude towards my seminar guide **Prof. A. A. Shahbad** for her support, continuous guidance and being so understanding and helpful throughout the seminar.

I furthermore thank Computer Department HOD **Dr. N. F. Shaikh** to encourage me to go ahead and for continuous guidance.

I would like to thank all those, who have directly or indirectly helped me for the completion of the work during this seminar.

Miss. Ekta Chaugule

ABSTRACT

Redaction of personal, confidential and sensitive information from documents is becoming increasingly important for individuals and organizations. In past years, there have been many well-publicized cases of data leaks from various popular companies. When the data contains sensitive information, these leaks pose a serious threat. To protect and conceal sensitive information, many companies have policies and laws about processing and sanitizing sensitive information in business documents.

The traditional approach of manually finding and matching millions of words and then redacting is slow and error- prone. This paper examines different models to automate the identification and redaction of personal and sensitive information contained within the documents using named entity recognition. Sensitive entities example person's name, bank account details or Aadhaar numbers targeted for redaction, are recognized based on the file's content, providing users with an interactive approach to redact the documents by changing selected sensitive terms.

Keywords: Data privacy, Document redaction, Data sanitization, sensitive information, Named Entity recognition.

Contents

List of Figures

1 Introduction **1**
1.1 Motivation....2

2 Literature Survey **4**

3 Objectives **5**

4 Methodology **6**
4.1 How to Determine Sensitive Words? 6
4.2 Handling Varieties of Documents....8
4.3 Auto-Redactor....8
4.4 Redaction of User Selection 8
4.5 Retraining Models....10
4.6 Different Models Tested for Training Approaches 10
4.7 Redaction 10

5 Applications....11

6 Advantages & Limitations **13**
Advantages 13
Limitations....14

7 Conclusion **15**

8 REFERENCES **16**

List of Figures

1.1 The Project Overview……..1

4.1 Syntax Analysis……...7

4.2 Recognized Entities……..7

4.3 Selection of type of Sensitive Entity…….....9

4.4 Output File……...9

Chapter 1

1. Introduction

As more information gets stored and shared digitally, it's important to keep sensitive data, like names, addresses, and credit card numbers, safe from unauthorized access.

Dataredaction is the process of hiding or removing such confidential information from documents. Traditionally, redaction has been done manually, which can be time-consuming and lead to mistakes.

Named Entity Recognition (NER), a tool from the field of Natural Language Processing (NLP), helps automate this process.NER identifies key pieces of information, like names, dates, or locations, making it easier and faster to redact sensitive data accurately. In this presentation, we'll discuss how NER works, why it's useful for data redaction, and how it can help protect important information in a more efficient way.

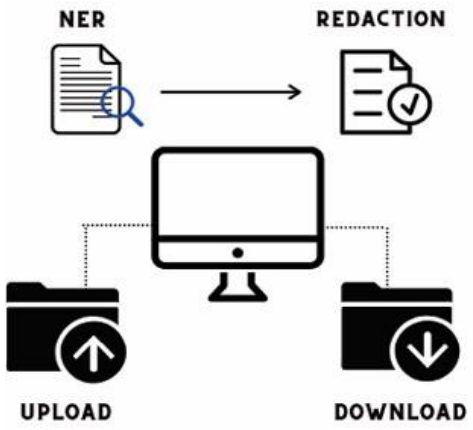

Figure 1.1 Project Overview

Motivation

1. **Growing Data Privacy Concerns:**

   The rapid increase in digital data has led to significant challenges in maintaining user privacy. Data breaches and unauthorized sharing of sensitive information can result in severe consequences, including financial loss, reputational damage, and loss of trust. Addressing these issues requires innovative solutions to ensure data privacy without hindering data usability.

2. **Compilance with Regulations:**

   Regulatory frameworks like GDPR (General Data Protection Regulation), HIPAA (Health Insurance Portability and Accountability Act), and CCPA (California Consumer Privacy Act) impose strict requirements for data protection. Organizations must adhere to these regulations by implementing systems that can detect and redact sensitive information efficiently, making automated redaction solutions critical.

3. **Significance of Named Entity Recognition (NER):**

   Named Entity Recognition is a cutting-edge Natural Language Processing (NLP) technique that identifies and classifies entities such as names, locations, and dates in text. By leveraging NER, it is possible to systematically detect sensitive information that needs to be redacted, enhancing both accuracy and efficiency in data privacy applications.

4. **Technological Relevance:**

   The integration of NER with data redaction demonstrates the practical application of theoretical concepts in NLP. This fusion represents a significant step toward solving real-world challenges, showcasing how advanced technologies can provide innovative solutions to pressing problems.

5. **Social Responsibility:**

   Data privacy is not just a technical challenge but also a social responsibility. Ensuring the secure handling of sensitive information fosters trust between users and organizations. This project aligns with the ethical goal of protecting individual privacy and contributes to building a safer digital ecosystem.

6. **Industry Demand**:

   With the increasing reliance on digital data, industries such as healthcare, finance, and e-commerce face mounting pressure to manage and protect sensitive information. Automated data redaction systems that utilize NER are becoming an essential part of data management strategies, making this project relevant and highly valuable for future career opportunities.

Chapter 2

Literature Survey

Study	Key Contributions	Challenges
Johnston et al., 2015	Manual redaction is prone to errors and in- efficient for large datasets.	Human errors, missed sensitive information, time-consuming.
Smith & Jones, 2018	Introduced automated redaction using NLP, leveraging NER for identifying sensitive data.	Requires fine- tuning for domain-specific data; difficulty handling noisy text.
Collobert et al., 2011	Used machine learning models for NER to automatically detect entities like names, locations, and dates.	Limited to structured data ; early models struggled with contextual ambiguity.
Devlin et al., 2019	BERT-based NER models significantly improved the accuracy of entity recognition in various contexts.	Transformer models require large datasets and computational resources for training and fine-tuning.
Gupta et al., 2020	Address identity recognition in unstructured and noisy text.	Difficulty in detecting out-of-vocabulary terms and handling diverse entity types.
Yao & Shi, 2021	Explored the balance between over-redaction and under-redaction in auto-mated systems.	Risk of over-redaction removing important non-sensitive content, impacting data us- ability.
Lee et al., 2020	Developed domain-specific NER models for healthcare and financial sectors, improving precision.	Specialized models require domain-specific training and do not generalize well to other areas.
Abadi et al., 2016	Introduced privacy - preserving techniques in machine learning to protect sensitive data during processing.	Ensuring data privacy without compromising NER accuracy remains a challenge.

2.1. Literature Survey Table

Chapter 3

Objectives

1. **Growing Sensitive Data:**

   The digital explosion has led to a massive increase in sensitive information (e.g. ,personal, financial, medical) that requires protection.

2. **Regulatory Compliance:**

   Stricter privacy regulations (GDPR, HIPAA, CCPA) mandate the secure handling of personal data, compelling organizations to ensure compliance to avoid legal penalties.

3. **Inefficiency of Manual Processes:**

   Traditional manual redaction methods are slow and error-prone, risking exposure of sensitive information and compromising data usability.

4. **Advancements in NLP:**

   Modern techniques like Named Entity Recognition (NER) leverage machine learning to automate and enhance the accuracy of data redaction.

5. **Enhanced Security :**

   Automating redaction with NER improves data security, reduces the risk of breaches, and builds trust with clients by ensuring consistent privacy protection.

Chapter 4

Methodology

The idea of automatically finding sensitive content from data is achieved through Named Entity Recognition (NER). NER is a method in natural language processing that tags words based on the context in a sentence. For example, in " Neha works in India," "Neha" is a Person entity and "India" is a Place (GPE). These tags are crucial for identifying sensitive content in documents. Personally identifiable entities include names, places, and numbers like bank account numbers, Social Security Numbers, Aadhaar numbers, and dates of birth.

Users can select specific entities to redact from the data, such as names and Aadhaar numbers. These entities are identified by NER using SpaCy, retrained on a custom dataset. After identifying these entities, they are passed as input parameters to a Python script that redacts the identified words from the document. The Python script uses a "py-redact" module that searches for the identified words using regex and performs the redaction.
The project is split into the following components:

- **Automatic Redactor:** Takes a user's text, identifies sensitive entities, and presents a redacted copy.
- **User Choice Redaction:** Finds sensitive terms in a user's content and allows them to choose which ones to redact.

4.1 How to Determine Sensitive Words?

NER is the process of discovering and categorizing key information (entities) in text, also known as entity chunking , extraction, or identification. An entity can be any phrase or collection of phrases that refers to the same thing. When recognized, it is classified into a predetermined category. The main go a list to find and classify named items into specified categories , such as people's names , organization names , locations , events , time expressions , quantities , monetary values , percentages , etc.

The core of any NER model consists of a two-step approach:

- **Detection of Named Entity:** Recognizes an entity in a word or a string of words. For example, “Lake Charlotte in Canada” denotes two entities: Lake Charlotte and Canada. The inside- outside-beginning method is often used to express the start and end of entities.
- **Categorization of Named Entity :** Involves creating entity categories, such as Person or Location.

This project focuses on the ability to recognize sensitive words in documents using Natural Language Processing (NLP), a sub-field of Artificial Intelligence that enables computers to interpret, analyze , and understand human languages. NER categorizes named entities into specified categories, including people’s names , organizations, monetary values, locations, and dates . The project leverages SpaCy, an open-source Python toolkit for advanced NLP.

```
Noun phrases: ['The service', 'a network', '650 India Post Payments Bank', '1.46
lakh postmen', 'Gramin Dak Sevaks', 'GDS', 'The mobile update service', 'UIDAI',
'the ubiquitous and accessible network', 'post offices', 'postmen', 'GDS', 'IPPB
s vision', 'the underserved and unbanked areas', 'the digital divide', 'IPPB Man
aging Director', 'CEO', 'J Venkatramu', 'a statement', 'Tuesday']
Verbs: ['be', 'help', 'actualize', 'serve', 'bridge', 'IPPB', 'say']
```

Figure 4.1: Syntax Analysis

```
650 India Post Payments Bank ORG
1.46 CARDINAL
Gramin Dak Sevaks ORG
GDS ORG
GDS ORG
J Venkatramu PERSON
Tuesday DATE
```

Figure 4.2: Recognized Entities

4.2 Handling Varieties of Documents

Redaction occurs across many different sectors, necessitating a project that accommodates various document types. For example, the wording of a legal document differs from that of a hospital report. AI models, built by exposing a system to numerous examples, generalize across languages. This project uses the standard SpaCy model retrained on a custom dataset.

4.3 Auto-Redactor

The auto- redact or is a crucial component of the project. When a user uploads a file, they must choose its form (e.g., legal, insurance, general) to determine which model to use . After submission, various processes occur to make predictions about the words or phrases in the document. The auto-redactor aims to maintain the document's meaning while removing confidential details. Sensitive terms are identified, and if a term is classified , it is replaced with a label (e.g. , ' PERSON ') to ensure the content remains intact while removing confidential information.

4.4 Redaction of User Selection

Not all sensitive terms in a document need to be removed. For instance, if a company has specific client information, that data must be deleted, but information a boutthe company may be retained. The User Selection Redaction feature allows users to preview the sensitive words found in a document. When uploading a document, users can choose to preview the sensitive words identified by the model. They can then select which terms to remove, and only the chosen terms are redacted in the final document.

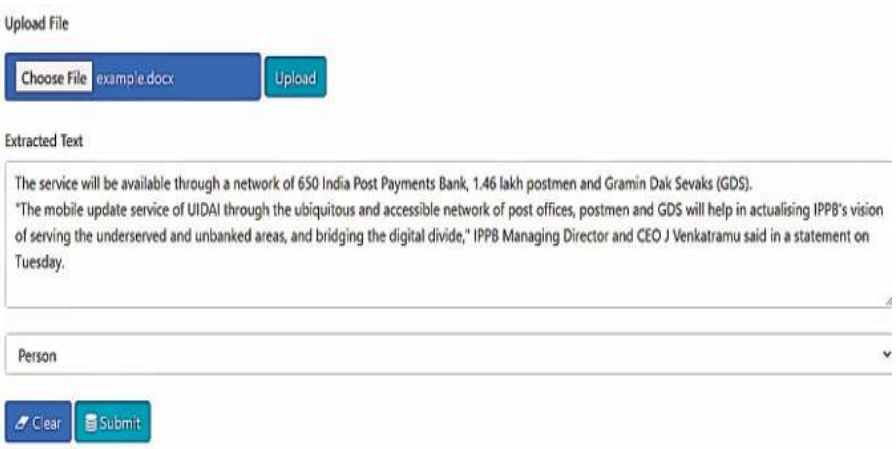

Figure.4.3: Selection of type of Sensitive Entity

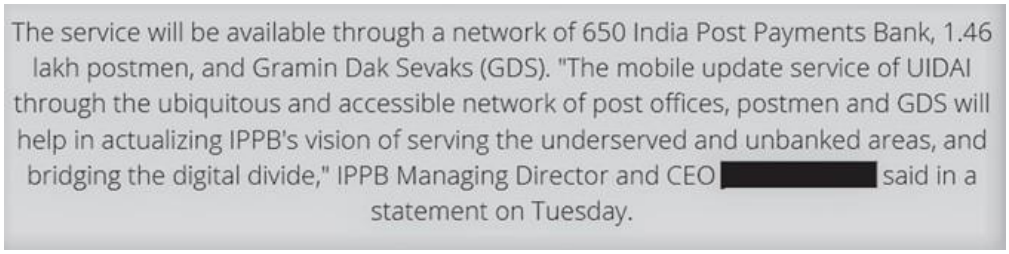

Figure 4.4: Output File

4.5 Retraining Models

Retraining models improves their algorithms to generalize across a more significant number of documents. The auto - redactor's models develop predictions based on supporting documents, identifying sensitive terms and retraining with more data. To retrain, the entire file is required, including the beginning and end indexes of relevant words. These indexes and labels create training data , combined with the original file. The model predicts named entities, and if predictions are incorrect, the weight is adjusted for better accuracy. The updated model is then saved.

4.6 Different Models Tested for Training Approaches

- **Transfer Learning on SpaCy:**

  NER identifies and classifies entities in text into predefined classes like 'person' or 'location.'Existing SpaCy models can be trained on custom data to improve accuracy for specific document contexts.

- **Using CRF and BiLSTM:**

  BILSTM , a neural network model based on RNN, combines forward and backward LSTM. It retains relevant information in the data context and captures dependencies of each word before and after.

- **Using PyTorch and BERT:**

  BERT performs well in various NLP tasks due to Semi-Supervised Learning. It utilizes a transformer architecture with an encoder stack and self-attention, generating hidden-size output vectors for various tasks, such as classification and translation.

4.7 Redaction

The input to the Python program is the file needing redaction and the sensitive words identified from the NER module. The script reads the document line by line and searches forwords to redact. Using regex, it compares each word with the input, and if a match is found, it blackouts the word while preserving the document's structure.

Chapter 5

Applications

1. **Healthcare**

   - **Patient Privacy:**

     Automatically redact sensitive patient information (like names, social security numbers, and medical records) from documents to comply with HIPAA regulations.

   - **Clinical Trials:**

     Redact identifying details in clinical trial reports to ensure participant confidentiality while sharing data for research purposes.

2. **Finance**

   - **Fraud Prevention:**

     Redact sensitive financial data from documents and reports to prevent identity theft and fraud.

   - **Regulatory Compliance:**

     Automatically redact personally identifiable information (PII) from documents submitted to regulatory bodies.

3. **Legal**

   - **Legal Document Management:**

     Redact names, addresses, and other sensitive information from legal filings, contracts, and court documents.

   - **Discovery Process:**

     Streamline the discovery process by automatically identifying and redacting sensitive information from large volumes of documents.

CHAPTER 5. APPLICATIONS

4. Government

- **Public Records:**

  Redact sensitive information from public records before they are released to the public, ensuring citizen privacy.

- **Intelligence Reports:**

  Automatically redact names and locations in intelligence reports to protect sources and methods.

5. Human Resources

- **Resume Screening:**

  Redact personal information from resumes before sharing them with hiring managers to reduce bias in the hiring process.

- **Employee Records:**

  Automatically redact sensitive employee information from HR documents when sharing them with external parties.

Chapter 6

Advantages & Disadvantages

Advantages

1. **Automated Process:**

   NER automates the identification of sensitive information, reducing the time and effort required for manual redaction.

2. **Increased Accuracy:**

   Advanced NER models can achieve high accuracy in recognizing entities, minimizing the risk of human error in identifying sensitive data.

3. **Scalability:**

   NER systems can handle large volumes of documents efficiently, making it suitable for organizations dealing with significant amounts of data.

4. **Customization:**

   NER models can be retrained on custom datasets to improve performance in specific domains, allowing for more effective detection of relevant entities.

5. **Consistency:**

   Automated systems provide consistent results across different documents, ensuring a uniform approach to data redaction.

6. **Compliance:**

   Using NER helps organizations comply with data protection regulations (e.g., GDPR, HIPAA) by ensuring that sensitive information is properly redacted before sharing.

7. **Resource Efficiency:**

   By automating the redaction process, organizations can allocate human resources to more strategic tasks rather than manual data protection.

Disadvantages

1. **Initial Setup Complexity:**

   Setting up an NER system can be complex , requiring time, resources, and expertise to train the models effectively.

2. **False Positives and Negatives:**

   NER may misidentify entities, resulting in false positives (incorrectly tagging non-sensitive information) or false negatives (failing to tag sensitive information).

3. **Context Sensitivity:**

   NER may struggle with context-specific terms or jargon, leading to challenges in accurately recognizing sensitive information in specialized fields.

4. **Language Limitations:**

   Many NER models are trained primarily on English data, which may limit their effectiveness in recognizing entities in other languages or dialects.

5. **Maintenance:**

   NER models require regular updates and retraining to remain effective as language and terminology evolve, necessitating ongoing maintenance.

6. **Cost:**

   Implementing and maintaining an NER system can incur costs related to software, hardware, and personnel, especially for smaller organization.

Chapter 7

Conclusion

In conclusion, Named Entity Recognition (NER) plays a vital role in the automated redaction of sensitive information across various sectors. By leveraging advanced natural language processing techniques, NER enhances the efficiency, accuracy, and consistency of data redaction processes , making it a valuable tool for organizations striving to protect sensitive information. Despite its advantages, challenges such as initial setup complexity, potential misidentifications, and the need for continuous maintenance must be addressed. Overall, NER represents a significant advancement in safeguarding personal and sensitive data, ensuring compliance with data protection regulations while streamlining workflows in diverse fields.

Ultimately, the integration of NER into data redaction processes represents a significant advancement in the protection of personal and sensitive data. By embracing this technology, organizations cannot only safeguard their data but also foster trust with their clients and stakeholders, reinforcing their commitment to data privacy and security. As we move forward, the continued evolution of NER and its applications will play a crucial role in shaping the future of data management, ensuring that sensitive information remains protected in an increasingly digital world.

Chapter 8

REFERENCES

[1] Grechanik Mark, McMillan Collin, Dasgupta Tathagata, Poshyvanyk Denys, Gethers Malcom,"Redacting Sensitive Information in Software Artifacts," Association for Computing Machinery ICPC 2014.

[2] Shashi Prakash Tripathi, Harshita Rai,"SimNER-An Accurate and Faster Algorithm for Named Entity Recognition, " Second International Conference on Advances in Computing, Control and Communication Technology (IAC3T), 2018.

[3] Jinhua Ma, Xinyi Huang, Yi Mu, Robert H. Deng, "Authenticated Data Redaction with Accountability and Transparency: Significance of Term Relationships on Anonymization," IEEE Transactions on Dependable and Secure Computing, 2019.

[4] EricBier, Richard Chow, Philippe Golle, Tracy Holloway King, Jessica Staddon," The Rules of Redaction- Identity, Protect, Review (and Repeat)," The IEEE Computer and Reliability Societies, 2019.

[5] Du Yanrui, Zhao Weixiang, "Named Entity Recognition Method with Word Position," 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI), 2020, pp. 154-159, doi: 10.1109/IWECAI50956.2020.00038.

[6] spaCy.[GitHub]. [2021.4] (https://github.com/explosion/spaCy)

CHAPTER 8. REFERENCES

[7] Zhiheng Huang, Wei Xu, Kai Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," arXiv 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, " BERT :Pre training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805,May2019.

[9] P. Sun, X. Yang, X. Zhao and Z. Wang, "An Overview of Named Entity Recognition," 2018 International Conference on Asian Language Processing (IALP), Bandung, Indonesia.

[10] Named Entity Recognition using SpaCy and Tensor Flow. (https://aihub.cloud.google.com/u/0/p/products)

[11] B. Ertopc ¸ uetal .,"A new approach for named entity recognition, " International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 2017, pp. 474-479.

[12] Emma Strubell, Patrick Verga, David Belanger, Andrew McCallum, "Fast and Accurate Entity Recognition with Iterated Dilated Convolutions," arXiv:1702.02098v3, July 2017.

[13] David Sánchez, Montserrat Batet, " C-sanitized: A Privacy Model for Document Redaction and Sanitization," unpublished.

[14] K. Elissa, "Named Entity Recognition Method with Word Position," 2020 International Work- shop on Electronic Communication and Artificial Intelligence (IWECAI), IEEE.