WHITE PAPER

BENCHMARKING PII IDENTIFICATION IN UNSTRUCTURED TEXT: A QUANTITATIVE COMPARISON

JAN 2024

By DataXpert, in collaboration with
UT Dallas Erik Jonsson School of Engineering and Computer Science Graduate School
(Shivashankar Chandrashekaraiah)

UT DALLAS

Dataxpert
RAPID INSIGHTS. RAPID GROWTH

ABSTRACT

This white paper presents a quantitative benchmark analysis of three PII identification solutions: Microsoft Presidio (open-source), AWS Comprehend (cloud service), and Protecto (specialized software). We assessed each model's effectiveness in precision and recall across diverse PII categories.

Our results demonstrate Protecto's consistent delivery of superior precision and a minimal false positive rate, signifying its proficiency in accurate PII identification while minimizing errors.

While other models exhibit competitive performance in specific areas, Protecto's overall accuracy positions it as a compelling choice for organizations seeking a dependable and comprehensive PII identification solution.

QUANTITATIVE BENCHMARK STUDY FOR PERSONALLY IDENTIFIABLE INFORMATION (PII) IDENTIFICATION

GROWING DATA RISKS

In a world increasingly interwoven with artificial intelligence, the need to protect individual privacy and data has never been more paramount. As vast amounts of information flow through complex systems, safeguarding Personally Identifiable Information (PII) becomes a critical line of defense against unintended exposure and potential harm. The accuracy and efficiency of identifying this sensitive data significantly impact the effectiveness of such protection, making the choice of the right tools crucial.

This whitepaper delves into this critical domain by presenting a comprehensive benchmarking study of three prominent PII identification solutions: Microsoft Presidio, AWS Comprehend, and Protecto. Ultimately, our findings aim to empower organizations with the knowledge to choose the most reliable and robust solution for their specific needs.

PROTECTING PRIVACY: UNDERSTANDING PII

In today's digital world, where information forms the backbone of every interaction, safeguarding Personally Identifiable Information (PII) is crucial. PII refers to any data that can be used to identify an individual, like their social security number, full name, or email address.

The increased reliance on technology has led to a significant rise in PII shared with various organizations. Businesses collect customer data to understand their markets, while individuals readily share information like phone numbers and addresses for online services and purchases.

PII can be categorized into two types:

- **Direct identifiers**: These are unique to an individual and readily identify them, like passport numbers or driver's licenses.
- **Indirect identifiers**: These are broader personal details, like race or birthdate, not unique on their own, but when combined, can be used to identify someone. Studies show that combining seemingly innocuous details like gender, ZIP code, and birthdate can potentially identify up to 87% of U.S. citizens, highlighting the importance of cautious PII handling as highlighted in a Dataprivacylab study.

Increasing need for accurate PII identification in various sectors, the growing reliance on digital platforms has amplified the significance of accurate PII identification.

- **Financial institutions**: Financial institutions require precise PII detection to combat identity theft, detect fraudulent activities, and adhere to regulatory compliance.
- **Healthcare sector**: The healthcare sector relies on accurate PII identification to safeguard patient records, maintain confidentiality, and ensure compliance with privacy regulations.
- **Online businesses and customer service**: PII detection benefits online businesses and customer service entities by enhancing user experience, providing personalized services, and ensuring secure transactions.
- **Government agencies**: Government agencies utilize PII identification to enable citizen services, aid law enforcement, and prevent unauthorized access to sensitive information.
- **The healthcare**

CHALLENGES AND LIMITATIONS OF PII IDENTIFICATION

- **Unstructured Data**: Unstructured data, such as free-form text or multimedia content, poses challenges for conventional PII detection algorithms. Advanced techniques and natural language processing are required to effectively identify and protect PII within unstructured data.
- **Evolving Formats**: The constant evolution of data formats and communication channels introduces complexities in PII identification. Adaptable solutions are needed to recognize PII in new and emerging formats, ensuring comprehensive coverage and protection.
- **Privacy Regulations**: Adhering to privacy regulations adds a layer of complexity. Striking a balance between accurate PII identification and compliance with diverse and evolving privacy laws requires careful consideration and robust solutions.
- **Cross-Border Challenges**: Global operations introduce challenges related to varying data protection laws across jurisdictions. PII detection solutions must navigate these complexities to ensure compliance on a global scale.
- **Balancing Accuracy and User Privacy**: Achieving accurate PII identification while respecting user privacy is a delicate balance. Stricter identification measures may enhance security but can raise concerns about surveillance and invasive data practices.

QUANTITATIVE BENCHMARKING OF PII IDENTIFICATION SYSTEMS

There are many PII identification systems in the market, such as AWS Comprehend, AWS Macie, Google DLP, and Microsoft Presidio. However, there is no single benchmarking study that compares all of these models. In this whitepaper, we compare the prominent PII frameworks against Protecto to see how they fare against each other.

PII categories:

- Name of the person (both first name, middle name, and last name)
- Email address
- IP address
- Address
- Credit card number
- US SSN
- Organization name
- Phone number (US and UK region)

Dataset size:

We benchmarked these systems using a custom dataset created by us that is of 3,000 samples.

Systems used for comparison:

- Cloud offering of AWS Comprehend
- Microsoft Presidio (en_core_web_sm, evaluator version 0.1.1, analyzer version 2.2.351)
- Cloud version of Protecto

DATA PREPROCESSING FOR PII IDENTIFICATION BENCHMARKING

This section outlines the data preprocessing methodology used in our benchmarking study of PII identification systems. This data preprocessing approach aimed to create a balanced and realistic dataset for benchmarking PII identification systems. By incorporating synthetic data generation, specific entity tagging, and false positive mitigation, this methodology provides a robust foundation for fair and informative model evaluation.

Synthetic Data:

We utilized Faker, a Python library for generating realistic synthetic data, to create approximately 3,000 samples containing the target PII entities (name, email, address, etc.).

Template-Based Generation:

To ensure data relevance, we provided Faker with expected template formats and data sources for sampling, mimicking real-world data structures.

Tokenization and Span Generation:

- **Tokenization:** The generated text samples were segmented into individual tokens (words or sub-word units).
- **Span Identification:** For each PII entity in a sample, we identified the specific character positions (start and end) within the text, creating "spans." This information allows models to understand the location and boundaries of each PII instance.

Labeling:

- **BILOU Labeling:** We adopted the BILOU (Begin, Inside, Last, Outside, Unit) scheme to tag each token within a span. This standard format helps models understand the structure and continuity of PII entities within a sentence.
- **Model-Specific Adjustments:** When a model uses different label names for specific classes, we adjusted the dataset to match its requirements.
- **False Positive Mitigation:** To simulate real-world scenarios and evaluate false positive detection, we tagged as mask “O” if it is an invalid PII instance. For example, a credit card number that fails the Luhn algorithm was labeled as a mask instead of "credit card.“

Metrics

The metrics used in this benchmarking are:

- **Precision**: Precision is a metric used in the context of classification or information retrieval to measure the accuracy of positive predictions. It is calculated as the ratio of true positive predictions to the sum of true positive and false positive predictions.
- **Recall**: Recall measures the ability of a classification model to capture and correctly identify all relevant instances in a dataset. It is calculated as the ratio of true positives to the sum of true positives and false negatives.
- **F1 score**: It is a metric that combines both precision and recall into a single value, providing a balanced assessment of a model's performance. It is the harmonic mean of precision and recall.

BENCHMARKING RESULTS

This section analyzes the performance of the evaluated PII identification systems using various metrics. Each metric provides a specific insight into the strengths and weaknesses of each model.

Precision:

- ↗ Protecto consistently boasts higher precision scores across all eight PII categories compared to AWS Comprehend and Microsoft Presidio (MS Presidio).
- ↗ Lower precision in Comprehend and MS Presidio indicates a tendency to misclassify non-PII entities as actual PII (false positives), particularly for phone numbers, US SSNs, and credit card numbers.
- ↗ Protecto excels in accurately identifying false positives, demonstrating its superior ability to distinguish genuine PII from lookalike data.
- ↗ Figures 1 and 2 visually depict these precision results for further reference.

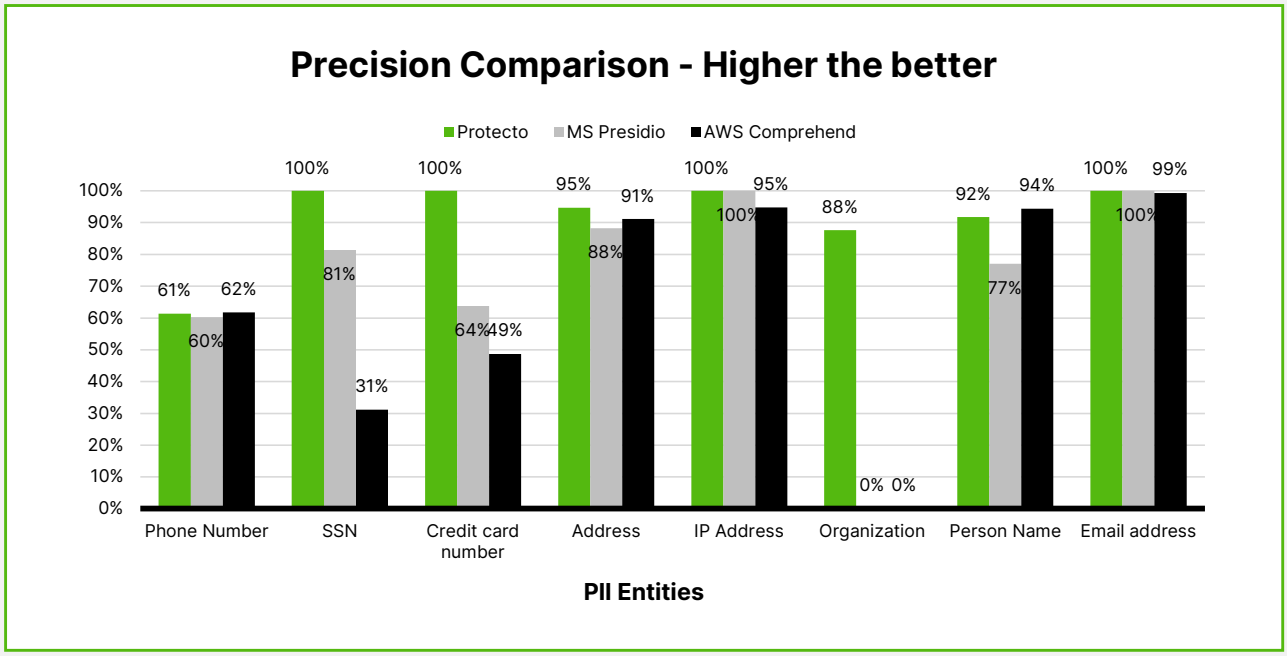
Recall:

- ↗ Recall scores across all PII systems are largely comparable, with some minor variations.
- ↗ AWS Comprehend exhibits higher recall for phone numbers and addresses compared to other models.
- ↗ Protecto leads in recall for IP addresses, person names, and organization names.
- ↗ All models achieve 100% recall for several PII classes.

F1 Score:

- ↗ As the F1 score harmonizes precision and recall, Protecto demonstrates higher F1 scores for most PII categories, including SSNs, credit card numbers, IP addresses, organizations, people, and email addresses.
- ↗ Comprehend holds the lead in F1 scores for phone numbers and addresses.

Figure 1: Precision comparison



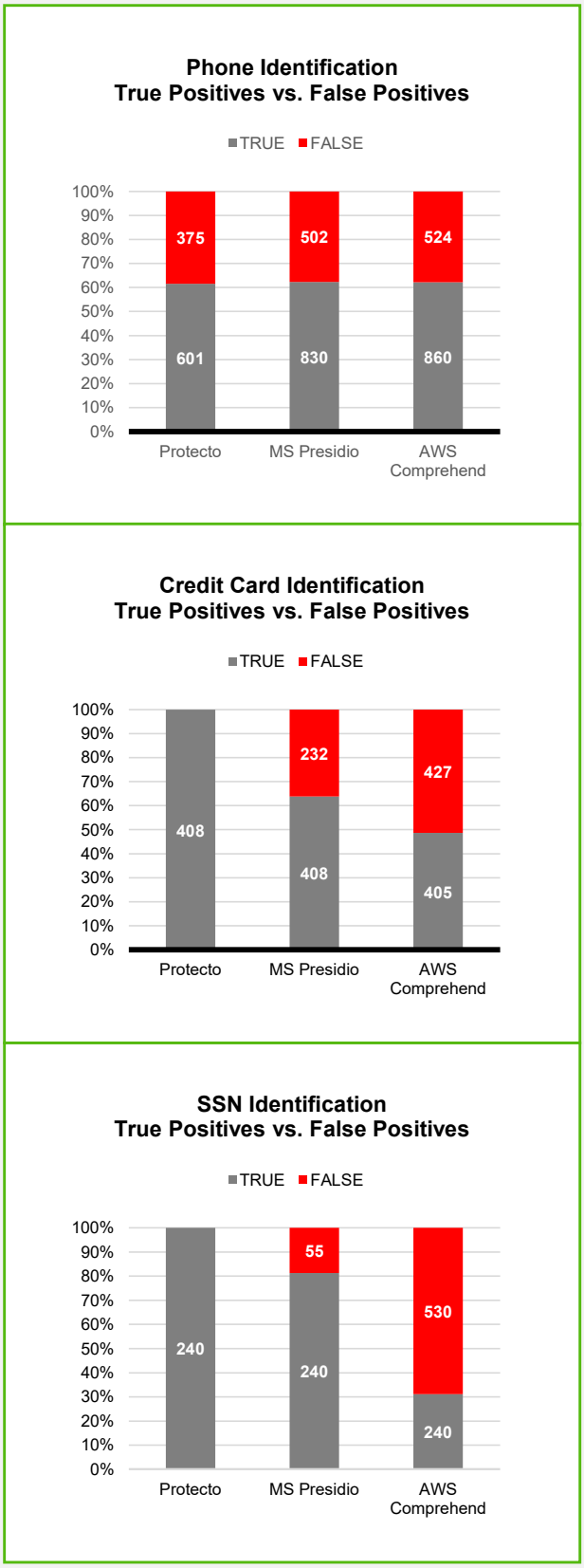
Interpretation:

The results highlight Protecto's overall superiority in accurately identifying PII entities (high precision) while minimizing false positives. Its recall performance excels for certain PII categories, with other models showing strengths in specific areas. These findings suggest that Protecto offers a reliable and balanced solution for various PII identification tasks.

Why False Positives are bad?

- **Impact on Decision Making:** Can lead to wrong decisions, like removing important data due to misclassification.
- **Resource Allocation:** Waste resources investigating and correcting false alarms.
- **Trust in the Model:** Frequent false positives erode trust in the system and its results.
- **Cost Implications:** Can cause financial losses due to unnecessary investigations or declines.
- **Model Evaluation:** Affect metrics like precision, hindering accurate performance evaluation.

Figure 2:
False Positives vs True Positives (Lower False Positives is Better)



CONCLUSION: CHOOSING THE RIGHT PII IDENTIFICATION SYSTEM

- **Protecto** stands out for its accuracy, particularly in handling sensitive data like credit card numbers and SSNs. Its ease of use and bulk processing capabilities make it a strong contender for organizations seeking a comprehensive and user-friendly solution. However, its coverage for phone numbers and addresses is limited.
- **AWS Comprehend** excels in detecting phone numbers, addresses, and person names. However, be mindful of its high false positive rate for sensitive data, which can hinder precision. Additionally, it lacks support for organization entities and has text size limitations.
- **Microsoft Presidio** offers flexibility with its multiple underlying NLP models and boasts impressive speed. However, its false positive rate and lower performance in specific areas (names, organization entities, postal codes) raise concerns. Additionally, it lacks bulk processing capabilities, and its metric scores are lower compared to the other models.

Context dependence raises false positive concerns

AWS Comprehend and Microsoft Presidio models can be manipulated by providing misleading context around PII data, leading to inaccurate classifications and increased false positives.

For instance, stating "My MAC address is <CREDIT_CARD_NUMBER>" might trick these models into identifying the credit card number as a MAC address. This vulnerability allows sensitive PII to be disguised and bypass detection, posing a significant security risk.

Protecto stands out for its resilience to such context-based manipulation. It consistently and accurately identifies PII entities, regardless of misleading surroundings. This makes it a more reliable choice.

PROTECTO

Pros	Cons
✓ With a seamless user experience, Protecto eliminates the need for any setup, requiring only a single API call for PII detection. It simplifies the decision-making process by removing any confusion associated with selecting the appropriate model.	x Restricted to detecting US and UK phone numbers.
✓ Excels in the detection of credit card numbers and SSNs, outperforming other models.	x Identifies location and street addresses correctly but doesn't identify zip codes in addresses.
✓ Offers support for async/bulk mode, allowing for efficient processing of multiple requests.	

AWS COMPREHEND

Pros	Cons
✓ Excels in detecting Phone numbers, Address, and Person names.	x High False positives for sensitive data such as credit card, phone numbers, and SSN can lead to a decrease in precision, as demonstrated by its context-aware detection.
✓ Supports detecting PII fields such as passwords, usernames, and AWS credentials.	x Lacks support for detecting Organization names which could be critical for enterprises.
	x Some additional drawbacks include the absence of an async/bulk mode and a text size limit of 100 KB.

MICROSOFT PRESIDIO

Pros	Cons
✓ Presidio provides the flexibility to use multiple models from Spacy, HuggingFace as the underlying NLP model for named entity recognition, but this raises confusion when choosing the right model as Presidio doesn't provide any specific recommendation or numbers to choose the right model for a certain task.	x Presidio exhibits certain drawbacks in its performance. One notable concern is the occurrence of numerous False Positives in credit card and phone number data samples, resulting in a decrease in precision.
✓ Delivers speed as it outpaces all three models due to the absence of external API calls.	x Presidio faces challenges in named entity recognition, as it performs worse than the other models in detecting names. Additionally, it lacks support for organization entities and fails to identify postal/zip codes in addresses.
✓ Doesn't impose any text size, providing flexibility in handling various document sizes.	

APPENDIX

Appendix – Data

Link to **Input Datasets**	Link to **Output Results**

Appendix – Product Details

Protecto

Protecto.ai distinguishes itself as the leading solution for enterprises leveraging Large Language Models (LLMs) by securing sensitive data without sacrificing data utility. Unlike competing solutions that face challenges with consistent tokenization, resulting in information loss, Protecto guarantees that LLMs deliver coherent and logical responses—even when processing tokenized personal and sensitive data. This unparalleled capability ensures adherence to data privacy laws while preserving the efficiency and precision of LLM applications, positioning it as an essential asset for businesses focused on data security and privacy in their AI initiatives. Visit their developer documentation for more information.

AWS Comprehend: Cloud-Based

AWS Comprehend is a cloud-based service offering a range of natural language processing capabilities, including PII identification. It boasts scalability and ease of use, allowing users to integrate PII detection directly into their applications through APIs. Comprehend offers pre-trained models for various PII types and languages, making it a convenient option for those seeking a quick and efficient solution.

Microsoft Presidio: Open-Source

Microsoft Presidio is an open-source library designed for identifying and anonymizing sensitive data in text. It offers a modular approach, allowing users to customize the detection of various PII types based on their specific needs. Presidio's strengths lie in its flexibility and transparency, making it well-suited for developers seeking fine-grained control over PII identification logic. However, its open-source nature means setting up and maintaining the infrastructure falls on the user, and its accuracy might not match that of commercially developed solutions.