SafePost:
AI-Powered Post PII Detection

Abstract:

This paper addresses concerns on privacy risks associated with posts on social media platforms, focusing on the detection of personally identifiable information (PII) exposure using AI-driven techniques. PII is information that can be used to distinguish or trace an individual's identity. As social media usage continues to proliferate, individuals are increasingly disclosing sensitive information and PII, often without fully understanding the associated risks. This widespread oversharing raise significant privacy concerns, including heightened susceptibility to identity theft, privacy breaches, and other malicious activities. The proposed solution leverages AI to detect PII in posts. In our project, users upload images to the SafePost website, which are processed by the server for text extraction using AWS Textract. The extracted text is then analyzed by three AI models—Gemini, AWS Comprehend, and Presidio—each assigned a respective weight of 0.2, 0.4, and 0.4. By applying a weighted threshold of 0.7, the system determines whether PII is present before alerting the user to potential privacy risks, thereby helping them safeguard their information prior to publishing. SafePost aims to enhance user privacy by proactively detecting and alerting users about the sharing of PII, offering a functionality that can be integrated into today's social media platforms for enhanced privacy protection.

Written by:

Hayat Alzahrani
King Abdulaziz University
Hhaz1122@hotmail.com

Rahaf Algarni
King Abdulaziz University
rahaf.bg1@gmail.com

Retaj Alghamdi
King Saud University
gretaj4221@gmail.com

Layan Alaboudi
Taibah University
faisal233ksa@gmail.com

Date:
August 29, 2024

1. Introduction

In today's digital era, user privacy has become a critical issue as more individuals engage with social media platforms, often disclosing vast amounts of personal information. With every post, sensitive details such as personally identifiable information (PII) can be unintentionally exposed. These digital breadcrumbs provide attackers with an easy opportunity to collect information for various cyberattacks, ranging from identity theft and privacy breaches to more sophisticated forms of exploitation like phishing, social engineering, and financial fraud.

Despite these significant risks, many social media platforms still do not offer comprehensive measures to protect users from oversharing their personal information. There is a distinct lack of real-time privacy analysis tools that assess the content for PII before it is shared. This gap exposes users to preventable risks by allowing potentially harmful posts to be published without thorough scrutiny.

Our work addresses this critical gap by integrating a real-time privacy analysis system that detects the presence of PII in social media posts before they permanently propagate on the internet. This system alerts users through simple notification prompts, encouraging them to review the content before posting.

In this project, we focused on improving the accuracy of PII detection by employing a combination of advanced AI tools: AWS Textract for text extraction, and three models for PII detection—Gemini, AWS Comprehend, and Presidio. Initially, we evaluated various solutions for extracting text and identifying personally identifiable information across different contexts. We selected AWS Textract for its robust text extraction capabilities, while Gemini, AWS Comprehend, and Presidio were chosen based on their individual strengths in detecting PII. Each model was integrated into our workflow with weighted contributions—0.2 for Gemini, 0.4 for AWS Comprehend, and 0.4 for Presidio. This multi-model approach allowed us to leverage their strengths and apply a weighted threshold to enhance the accuracy of PII detection. We aim to integrate this tool within today’s most popular social media platforms, which serve as the primary channels through which individuals widely share their personal information.

2. Literature Review

Several studies have highlighted the significant privacy risks posed by the exposure of PII across various online platforms, where cybercriminals can exploit such information for malicious purposes. The study "Identifying, Collecting, and Monitoring Personally Identifiable Information: From the Dark Web to the Surface Web" sheds light on the critical issue of PII exposure and its potential exploitation by cybercriminals across various online platforms [1]. This research focuses on systematically identifying and collecting PII records from both the dark web and surface web to assess privacy risks and safeguard at-risk individuals. While this approach is highly valuable for identifying exposed PII after data breaches, it largely deals with the aftermath of privacy violations, aiming to protect victims post-incident. The study emphasizes the importance of timely PII collection and monitoring, particularly from covert dark web sources, yet acknowledges the challenges in establishing comprehensive monitoring systems. By contrast, our project seeks to prevent PII from being exposed in the first place by integrating automated PII detection systems within social media platforms. Through real-time analysis, users are alerted before their posts are published, helping them avoid inadvertently sharing PII. This proactive approach bridges a critical gap by protecting privacy at the source, thus reducing the likelihood of PII being exposed and exploited across various online platforms.

3. Methodology

3.1 Text Extraction from Posts

PII in social media posts typically appears as text, such as phone numbers, identity numbers, or personal emails. To effectively detect PII, we need to focus on extracting all visible text within the image. While text extraction can be performed using various OCR algorithms or easily integrated tools, we opted for Amazon Textract to enhance accuracy. Amazon Textract is a machine learning service that automatically extracts text, handwriting, layout elements, and data from scanned documents.[2] Compared to standard OCR tools, Textract provided us with more accurate results, including successful extraction of handwritten text.

3.2 Using AI for PII Detection: Gemini, Amazon Comprehend, Microsoft Presedio

After extracting the text, we need to determine whether it contains PII based on predefined criteria. In our context, we have defined PII to include data such as phone numbers, emails, and other personal information, while excluding first and last names in social media contexts. Users are alerted if such personal information is detected.

To make accurate determinations, we employ three different models: Gemini, Amazon Comprehend, and Microsoft Presidio.

1. Gemini is a Large Language Model (LLM) developed by Google DeepMind. LLMs are a subset of Natural Language Processing (NLP) models, trained on vast amounts of text data to understand and generate human language effectively.[3]
2. Amazon Comprehend is an NLP service that uses machine learning to extract key insights from text, including the identification of entities and PII.[4]
3. Microsoft Presidio is a specialized tool designed for the rapid identification and anonymization of private entities (PII), such as credit card numbers, names, locations, social security numbers, and other sensitive data.[5]

These models operate in parallel to evaluate the sensitivity of the extracted information. Each model independently returns a binary (true/false) determination regarding the presence of PII. For Gemini, we provide a prompt asking, 'Does the following text contain any personally identifiable information (PII), such as emails, phone numbers, etc.? considers names are not PII' and interpret its response accordingly.

While Gemini offers broad language understanding capabilities, Amazon Comprehend and Microsoft Presidio are more specialized in detecting PII, resulting in higher accuracy for these specific tasks. This complementary use of models allows for a robust and reliable PII detection system.

3.3 Computing Weighted Threshold

Based on our evaluation, Amazon Comprehend and Microsoft Presidio consistently demonstrated higher accuracy in identifying PII compared to Gemini. Consequently, we have assigned a weighted threshold of 0.4 each to Comprehend and Presidio, while Gemini has been assigned a weight of 0.2. The detection process operates as follows: when a model identifies PII in the extracted text, its corresponding weight is added to the cumulative score. If a model does not detect PII, its weight is excluded from the total.
We empirically determined that a threshold of 0.7 for the cumulative weight offers an optimal classification. This threshold was established after testing on a diverse set of images, followed by extensive manual verification of the results. When the cumulative weight meets or exceeds 0.7, the

system classifies the text as containing PII ('true'). Conversely, if the cumulative weight falls below this threshold, the system classifies the text as not containing PII ('false').

3.4 SafePost Dummy Social Media Website

We developed SafePost as a dummy social media website to simulate a real social media experience, incorporating our PII detection tool before publishing content. This platform allows users to post content just as they would on a typical social media site. Prior to publishing, our PII detection tool scans the post for any PII and alerts users to potential privacy risks. This simulation provides a hands-on demonstration of how the tool functions in a real-world scenario, ensuring that users are aware of privacy issues before their posts go live.

4. Results

The integration of multiple AI models with a weighted threshold effectively improved the accuracy of PII detection, and successfully prompted alerts for posts containing PII.

As Figure 1 shows, a post uploaded to our social media platform contains PII-phone number. Through the utilization of SafePost's AI models—Gemini (0.2), AWS Comprehend (0.4), and Presidio (0.4)—the system effectively identified the phone number as a type of PII. This depiction emphasizes the system's proficiency in detecting and flagging potential privacy concerns linked to sensitive data like phone numbers before content dissemination on social media platforms.

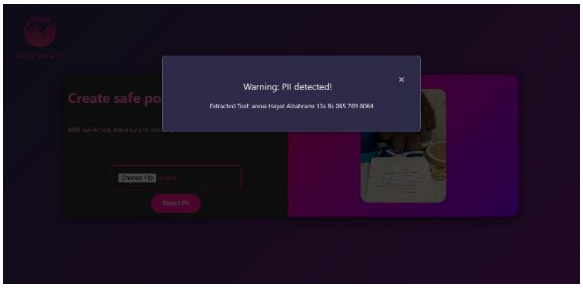


Figure1 :PII Detection Alert Message

Moreover, as shown on Figure 2, an analysis of the interface of our dummy social media website, where a user uploaded an image bearing text, revealed that following scrutiny by SafePost's AI models—Gemini (0.2), AWS Comprehend (0.4), and Presidio (0.4)—no PII was detected.

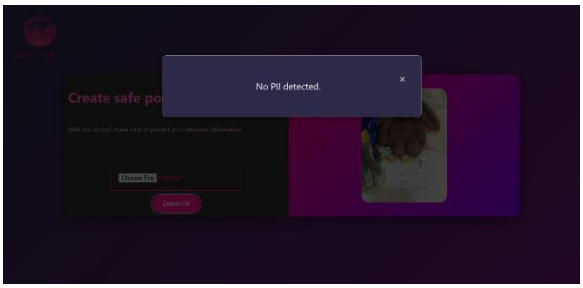
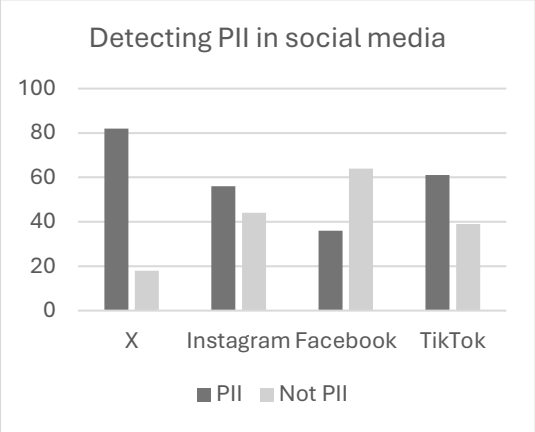


Figure 2 : No PII in the visible text

To further illustrate our findings as shown in Figure 3, we collected 400 random posts from widely used social media platforms—TikTok, Instagram, X, Facebook. A bar chart was utilized to display these findings. Each bar represents the prevalence of PII within posts from each application, providing an overview of the presence of PII in social media content. It is important to note that this short study does not claim to rank platforms in terms of PII exposure but rather showcases the existence of PII instances shared across the internet. Additionally, we used our dummy social media platform to detect PII in these posts, which promptly notified users of any detected PII.



5. Discussion

In the early stages of our project, we relied solely on the Gemini model for detecting PII. However, we quickly observed that Gemini's performance was insufficient for reliable PII detection. This led us to integrate additional models—AWS Comprehend and Microsoft Presidio—into our system. The inclusion of these specialized models significantly enhanced detection accuracy. By assigning different weights to each model, we optimized the system's performance through iterative adjustments. Our experiments with various weight distributions ultimately yielded the best results, balancing the strengths and weaknesses of each model.

Despite these improvements, our approach has limitations. A key limitation is its reliance on external APIs for text extraction and PII detection. While these APIs provide robust capabilities, they may introduce dependencies that could affect performance and flexibility. Ideally, developing an integrated solution that processes images without relying on third-party APIs could offer more control and potentially better performance. Looking forward, we aim to develop our own custom PII detection model specifically trained on user-generated content, which will allow us to refine the detection process and improve system efficiency. Additionally, our current system does not address PII detection in videos, which is an important area for future development. As video content becomes increasingly prevalent on social media platforms, incorporating video analysis capabilities will be crucial for comprehensive privacy protection.

6. Conclusion

In conclusion, the effective deployment of AI models in detecting and flagging PII within image-based content highlights the value of preemptive strategies for privacy protection. The ultimate goal of this tool is to raise awareness and provide users with a moment of reflection, helping to prevent the "fast click" mentality that often leads to the unintentional sharing of personal information. By proactively safeguarding privacy, our approach aims to empower individuals to make informed decisions, ultimately reducing the risk of identity theft, privacy breaches, and other cyberattacks that prey on overshared information in the digital space. Nevertheless, this work underscores the pressing need for robust privacy safeguards and user awareness initiatives to better protect personal information on social media platforms.

7. Acknowledgements

This project would not have been possible without the unwavering support and encouragement from the KAUST, KAUST Academy, the National Cybersecurity Authority (NCA), and the University of Glasgow. We would like to express our deepest gratitude to our mentors Mudassar Aslam and Ali Hassan for their invaluable guidance and to our peers for their collaboration and insights. Special thanks to the entire program team for providing us with the resources, knowledge and platform to explore and develop innovative solutions in the field of privacy and cybersecurity. Moreover, this project was inspired by the privacy topic studied at the University of Glasgow, which highlighted the pressing need for effective privacy protection mechanisms.

8. References

[1] Y. Liu, M. Ebrahimi, J. Xin, F. Y. Lin, N. Zhang, W. Li, Z. Ahmad-Post, J. L. Hu, and H. Chen, "Identifying, Collecting, and Monitoring Personally Identifiable Information: From the Dark Web to the Surface Web," 2021.

[2] Amazon Web Services, "Amazon Textract." Accessed on: August 31, 2024. [Online]. Available: https://aws.amazon.com/textract/

[3] DeepMind, "Gemini: Flash." Accessed on: August 31, 2024. [Online]. Available: https://deepmind.google/technologies/gemini/flash/

[4] Amazon Web Services, "Amazon Comprehend Documentation." Accessed on: August 31, 2024. [Online]. Available: https://docs.aws.amazon.com/comprehend/

[5] Microsoft, "Presidio." Accessed on: August 31, 2024. [Online]. Available: https://microsoft.github.io/presidio/