Privacy Policy Analysis in Healthcare

Dev Ambrose Kommu Christopher
Department of Computer Science
Georgia Southern University
Statesboro, GA
dk09461@georgiasouthern.edu

Jessica Paul
Department of Computer Science
Georgia Southern University
Statesboro, GA
jp27926@georgiasouthern.edu

Sushma Rachel Palle
Department of Computer Science
Georgia Southern University
Statesboro, GA
sp24400@georgiasouthern.edu

***Abstract*—Privacy policies serve as critical instruments for informing users about data collection, storage, and sharing practices. In healthcare, where sensitive personal information is frequently processed, ensuring transparency and regulatory compliance is essential. This systematic literature review examines advancements in privacy policy analysis, focusing on automated techniques, readability assessments, and regulatory alignment. The study synthesizes findings from multiple research efforts, including the application of NLP and knowledge graphs for detecting inconsistencies, evaluating policy readability, and assessing compliance with regulations such as HIPAA and GDPR. Key findings highlight the limitations of existing methodologies, including the lack of standardized readability metrics, the static nature of compliance assessments, and the underdevelopment of integrated frameworks combining security, privacy, and transparency. This review underscores the need for dynamic, adaptive, and standardized approaches to privacy policy analysis, advocating for AI-driven tools to enhance policy comprehension and regulatory adherence in healthcare applications.**

I. Introduction

Privacy policies are essential for protecting personal data, particularly in healthcare, where sensitive patient information is collected, processed, and often shared across digital platforms. With the rapid advancement of digital health technologies which includes mobile health (mHealth) applications, concerns over privacy, security, and regulatory compliance have intensified. As healthcare applications require access to sensitive data, ensuring transparency in privacy policies is critical for maintaining user trust and regulatory adherence.

Privacy policies are intended to inform users about how their data is collected, stored, and shared; however, many policies remain difficult to comprehend, inconsistent in structure, and insufficiently transparent. Readability issues and lack of standardization in policy documentation create challenges in user comprehension, potentially leading to uninformed consent and privacy risks. Furthermore, knowledge graphs and machine learning techniques are being explored to enhance policy analysis by detecting contradictions and mapping compliance with regulations such as HIPAA and GDPR.

This review systematically examines existing research on privacy policy analysis, identifying key themes, methodologies, and findings. By addressing gaps in transparency, readability, and regulatory adherence, this study aims to contribute to improving privacy policy assessment and the development of more effective privacy frameworks in healthcare applications.

II. Methodology

A structured approach was used to select relevant literature. The process involved defining search terms such as "Privacy policy analysis," "Healthcare privacy policies," "Data security in mobile health applications," and "Knowledge graphs in privacy analysis." Searches were conducted using databases like ScienceDirect, PubMed, IEEE Xplore, USENIX Security Symposium, and Google Scholar. Inclusion criteria focused on peer-reviewed journal articles and conference proceedings published between 2020 and 2024, while studies lacking empirical analysis or written in non-English languages were excluded.

III. Literature Review

Recent advancements in privacy policy analysis leverage NLP and machine learning to enhance transparency and compliance evaluation. Tang et al. (2023)(1) introduced PolicyGPT, a Large Language Model (LLM)-based framework designed to categorize privacy policy contents with high accuracy. The study focused on addressing inconsistencies in policy interpretation by training PolicyGPT on diverse datasets containing privacy policies from different sectors. The model demonstrated significant improvements over traditional machine learning classifiers by achieving a higher precision in identifying key policy clauses and regulatory obligations. However, its reliance on pre-trained models posed limitations in interpreting context-specific policies, especially for domain-specific regulatory requirements.

Cui et al. (2023)(2) developed PoliGraph, an innovative framework that integrates knowledge graphs and NLP techniques to detect contradictions and inconsistencies within privacy policies. The authors proposed a structured representation of privacy terms using graph-based relationships, which allows automated assessments of policy coherence. This approach significantly improved the traceability of data-sharing practices across different policy versions. A major contribution of this study was its ability to map policies to legal frameworks like GDPR and HIPAA, facilitating regulatory compliance monitoring. However, PoliGraph's dependency on predefined ontologies limited its adaptability to new policy structures, necessitating continuous updates to knowledge bases for optimal performance. Andow et al. (2019)(3) did a very similar research with focus on using PolyLint to investigate internal privacy policy contradictions on Google Play. They used

PolicyLint to analyze the policies of 11,430 apps and found that 14.2% of these policies contain contradictions that may be indicative of misleading statements. They hypothesized that in doing so, PolicyLint significantly advanced automated analysis of privacy policies. Unlike PoliGraph, their limitation was that they realized extracting structured information from unstructured natural language text continues to be an open and active area of NLP research, and there were no perfect techniques for this challenging problem. PolicyLint was thus limited by the current state of NLP techniques, such as the limitations of NLP parsers and named-entity recognition. Its performance also depended on its verb lists and policy statement patterns, which may have been incomplete despite their best efforts, reducing overall recall.

Del Alamo et al. (2022)(4) conducted a systematic mapping study that classified privacy policy research into key themes, emphasizing the necessity for automated methods to improve the contextualization of extracted information. Their study aimed to bridge the gap between traditional manual privacy policy assessments and modern computational techniques by analyzing over 150 research papers. They identified that while there are several approaches to the analysis of privacy policies, there is a lack of standardized methodologies to effectively evaluate policies. Their work contributed to developing a taxonomy of privacy research, but the study's limitation was its inability to validate the effectiveness of these taxonomies in real-world applications.

Hakiem et al. (2024)(5) focused on the transparency, data management, and disclosure practices of privacy policies in mobile health applications. Their study analyzed privacy policies from over 50 mHealth platforms and found that a significant number lacked clear information regarding data-sharing practices. Their method involved a combination of NLP techniques and manual evaluation to assess policy compliance with legal frameworks. They concluded that most mHealth applications do not fully adhere to privacy regulations, particularly concerning data retention and third-party sharing disclosures. However, their study did not provide a scalable solution for automated policy evaluation.

Javed and Sajid (2024)(6) conducted a comprehensive literature review on privacy policy usability and compliance challenges in healthcare applications. Their work highlighted the inconsistencies in privacy policy structures across different healthcare services, making it difficult for users to understand their rights and data management practices. They identified that many policies use complex legal jargon, making them inaccessible to the general public. Their study contributed to the growing call for simplified privacy policy formats but was limited by its reliance on secondary data sources without empirical testing of proposed solutions.

For COVID-19 contact-tracing applications, Bardus et al. (2022)(7) examined privacy policies and found a direct correlation between policy transparency and user trust. They emphasized that clearer privacy disclosures result in increased adoption rates of digital health technologies. Javed and Sajid (2024)(6) extended this analysis by evaluating the readability of privacy policies and their impact on user comprehension, underscoring the importance of accessible policy language.

Another critical theme in privacy policy analysis is readability. Farooq et al. (2020)(8) assessed the readability of privacy policies in free healthcare apps, utilizing Flesch-Kincaid Reading Grade Level, SMOG, and Gunning-Fog indices. Their findings indicate that the average reading grade level of these policies is 9.5, slightly above the recommended standard for general readability. This aligns with Bardus et al. (2022)(7), who found similar readability challenges in COVID-19 contact-tracing apps, reinforcing the need for clearer, more accessible privacy policies in healthcare applications. These studies collectively highlight the need for privacy policies that cater to a diverse, global user base, ensuring compliance with readability and transparency standards.

Adhikari et al. (2023)(9) conducted a very innovative research on how to automatically extract privacy-specific artifacts from the policies, predominantly by using natural language classification tools. Their research focused on identifying the gap in classifying policies at a segment level, and provide an alternate definition of segment classification using sentence classification. The authors trained and evaluated sentence classifiers for privacy policies using BERT and XLNet and after doing that, their results showed that using sentence classifiers demonstrated improvements in prediction quality of existing models and , surpassed the current baselines for classification models, without requiring additional parameter and model tuning like some traditional methods require. It is great as a novel research and has contributed as a new method to privacy policy analysis that can be built on.

While these studies contribute to improved privacy policy assessment, notable gaps persist. The absence of standardized readability metrics impedes cross-platform evaluations, and real-world applications of automated privacy analysis tools remain limited. Existing approaches primarily focus on static compliance checks rather than dynamic, context-aware evaluations. Moreover, frameworks integrating security, privacy, and transparency remain underdeveloped.

IV. Synthesis and Conclusion

Automated privacy policy analysis tools, such as PolicyGPT, PolicyLint and PoliGraph, demonstrate superior accuracy in policy categorization and compliance evaluation compared to manual reviews. However, transparency in Health applications remains a critical concern, as many platforms fail to provide clear and accessible privacy disclosures. The correlation between privacy transparency and user adoption in contact-tracing applications highlights the necessity of clear, well-structured policies.

To address these gaps, our research project explores three key questions: (1) How can we measure the readability of privacy policies using various metrics? (2) How can knowledge graphs be used to map and detect contradictions within privacy policies? (3) How well do healthcare privacy policies align with HIPAA and GDPR regulations? By integrating NLP and knowledge graphs, our project aims to improve privacy policy

assessment, ensuring regulatory compliance and enhancing user comprehension. Future research should focus on refining automated frameworks to evaluate privacy policies dynamically, providing adaptive regulatory compliance and stronger data protection measures.

REFERENCES

[1] C. Tang, Z. Liu, C. Ma, Z. Wu, Y. Li, W. Liu, D. Zhu, Q. Li, X. Li, T. Liu, and L. Fan, "PolicyGPT: Automated analysis of privacy policies with large language models."

[2] H. Cui, R. Trimananda, A. Markopoulou, and S. Jordan, "PoliGraPh: Automated privacy policy analysis using knowledge graphs,"

[3] B. Andow and S. Y. Mahmud, "PolicyLint: Investigating internal privacy policy contradictions on google play,"

[4] J. M. Del Alamo, D. S. Guaman, B. García, and A. Diez, "A systematic mapping study on automated analysis of privacy policies," vol. 104, no. 9, pp. 2053–2076.

[5] N. Hakiem, S. H. Afrizal, Y. Setiadi, H. S. Albab, M. Riasetiawan, and S. Zulhuda, "Security and privacy policy assessment in mobile health applications: A literature review," vol. 14, no. 2.

[6] Y. Javed and A. Sajid, "A systematic review of privacy policy literature," vol. 57, no. 2, pp. 1–43.

[7] M. Bardus, M. A. Daccache, N. Maalouf, R. A. Sarih, and I. H. Elhajj, "Data management and privacy policy of COVID-19 contact-tracing apps: Systematic review and content analysis," vol. 10, no. 7, p. e35195. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[8] E. Farooq, M. A. Nawaz Ui Ghani, Z. Naseer, and S. Iqbal, "Privacy policies' readability analysis of contemporary free healthcare apps," in *2020 14th International Conference on Open Source Systems and Technologies (ICOSST)*, pp. 1–7, IEEE.

[9] A. Adhikari, S. Das, and R. Dewri, "Privacy policy analysis with sentence classification," in *2022 19th Annual International Conference on Privacy, Security & Trust (PST)*, pp. 1–10.