Analyzing Privacy Policies in Mobile Health Apps

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*—Mobile health applications often collect sensitive personal data, yet their privacy policies vary widely in structure, clarity, and transparency. This study presents a scalable approach for analyzing healthcare app privacy policies using knowledge graph extraction, unsupervised clustering, and hierarchy-aware interpretation. By structuring disclosures as actor–relation–data triples, we clustered 107 app policies and identified three distinct patterns: structured identifiers with unclear purpose, generalized containers linked to services, and opaque behavioral tracking with low transparency. To improve semantic interpretation, we introduced a parent entity layer that preserved the hierarchical nature of policy terms without altering model inputs. Our findings reveal gaps in disclosure precision and purpose justification, especially for behavioral and tracking data. The methodology offers a flexible foundation for future applications in privacy scoring, compliance monitoring, and cross-domain policy analysis.**

I. Introduction

A. Background and Motivation

Privacy policies are foundational to protecting user data, especially in healthcare, where sensitive medical information is collected, stored, and shared. The rise of mobile health (mHealth) applications has intensified this need, as millions of users now depend on these platforms for health tracking and digital consultations. However, studies show that health app privacy policies often lack transparency, use vague language, and fail to disclose critical data-sharing practices (Hakiem et al., 2024(1);Benjumea et al., 2020)(2)

Despite frameworks such as HIPAA in the U.S. and GDPR in the EU, which require clear communication of data practices, many policies remain inaccessible to users due to legal jargon or a lack of specificity (Javed & Sajid, 2024)(3). Automated tools like PoliGraph and PolicyLint have been developed to detect inconsistencies and contradictions in these documents, leveraging NLP and knowledge graphs for large-scale analysis (Cui et al., 2023(4); Andow et al., 2019)(5). However, most existing work focuses on readability or supervised classification, offering limited insight into the structural patterns and latent groupings of privacy policy behavior.

To address this, our research uses unsupervised learning and exploratory data analysis on structured privacy policy representations. By leveraging knowledge graphs and clustering techniques, we aim to identify hidden patterns in data collection and sharing disclosures. This work contributes to the growing need for dynamic, scalable methods that enhance transparency and support regulatory compliance in digital health applications.

B. Problem Definition

Although privacy policies are designed to inform users about how their data is handled, their actual content often fails to meet this goal. Prior studies reveal that many mHealth app policies obscure key information through vague or overly complex language, limiting user understanding and regulatory accountability (Javed & Sajid, 2024(3); Farooq et al., 2020)(6). Automated methods like PolicyGPT and PoliGraph have advanced the analysis of policy content (Tang et al., 2023; Cui et al., 2023)(4), yet they largely rely on supervised models or rule-based extraction, which require labeled data and constant maintenance.

What remains underexplored is the use of unsupervised techniques to reveal latent structure in privacy policy disclosures — particularly in how health apps differ in the types of data they collect and the purposes they assign to that data. Current tools are insufficient for identifying thematic groupings or detecting vague or redundant practices across policies in a scalable, adaptable manner.

This creates a pressing need for methods that move beyond static compliance checking to offer pattern discovery and policy landscape insights, especially in high-risk domains like healthcare.

C. Objectives and Contributions

This research aims to improve the analysis of healthcare privacy policies by exploring scalable, automated techniques that go beyond traditional readability or compliance checks. Specifically, the study focuses on unsupervised learning and structured policy representation to uncover patterns in how mHealth apps disclose their data practices. The key objectives of this study are:

- O1: To investigate how healthcare apps vary in the types of personal data they claim to collect and share.
- O2: To identify the most frequently stated purposes associated with data collection and sharing in health app privacy policies.
- O3: To apply clustering and visualization techniques that group policies based on shared disclosure patterns.

The novel contributions of this work include:

1) A knowledge graph–based pipeline for extracting structured data from healthcare app privacy policies using PoliGraph-er.
2) The application of unsupervised machine learning (K-Means) and dimensionality reduction (UMAP, t-SNE) to uncover thematic clusters among policies.
3) An exploratory analysis that highlights vague or unspecified disclosures, advancing the understanding of policy transparency gaps in the mHealth space.

By combining data wrangling, NLP-driven policy parsing, and clustering, this study provides actionable insights into the structural variability of privacy policies in healthcare, informing future efforts in transparency scoring and regulatory enforcement.

D. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews prior research on privacy policy analysis, with a focus on automated tools, readability challenges, and the regulatory context in healthcare. Section 3 outlines the methodology, including data acquisition, preprocessing, feature engineering, and clustering techniques. Section 4 presents the key results from exploratory data analysis and unsupervised modeling. Section 5 discusses the implications of these findings, their strengths and limitations, and comparisons with prior work. Finally, Section 6 concludes the paper and outlines directions for future research.

II. Literature Review

A. Theoretical Background

Privacy policy analysis sits at the intersection of natural language processing, regulatory compliance, and data ethics. In healthcare, this analysis becomes even more critical due to the sensitivity of personal health data and the potential consequences of its misuse. Regulatory frameworks like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) form the legal foundation for privacy disclosures, requiring organizations to clearly state their data collection, sharing, and protection practices (Hakiem et al., 2024(1); Benjumea et al., 2020)(2).

From a technical perspective, early privacy policy analysis relied heavily on manual reviews, which proved inefficient and inconsistent at scale. To address this, researchers introduced automated methods using natural language processing (NLP) and machine learning to parse policy language, classify clauses, and assess readability (Farooq et al., 2020)(6). Advanced models such as PolicyGPT (Tang et al., 2023)(7) and PoliGraph (Cui et al., 2023)(4) use large language models and knowledge graphs, respectively, to capture the semantics and structure of policies. These tools enable systematic extraction of policy entities (e.g., actors, data types, purposes) and support mapping policy statements to legal obligations.

Clustering and unsupervised learning further expand the analytical toolkit, offering ways to identify hidden patterns without requiring annotated datasets. Techniques like TF-IDF, K-Means, UMAP, and t-SNE help group similar disclosures and visualize relationships between policies. Combined with graph-based representations, these methods support scalable, interpretable privacy assessments — particularly useful in domains like mHealth where policy language is often inconsistent or vague.

B. Related Work

Recent work in privacy policy analysis has focused on enhancing automation, interpretability, and compliance verification. One key direction is the development of automated policy parsers using NLP. For instance, PolicyGPT (Tang et al., 2023)(7) applies large language models to categorize privacy statements, improving clause classification and regulatory alignment. However, its reliance on pre-trained models limits flexibility for domain-specific contexts like healthcare.

PoliGraph (Cui et al., 2023)(4) advances this by constructing knowledge graphs from HTML policy text, capturing the relationships between actors, data types, and purposes. Its extraction accuracy significantly outperforms traditional methods, though it depends on predefined ontologies that require frequent updates. PolicyLint (Andow et al., 2019)(5) focuses on detecting internal contradictions within policies and revealed that over 14% of app policies contain conflicting statements.

Beyond structural parsing, several studies have investigated policy readability and transparency. Farooq et al. (2020)(6) found that the average reading level of free healthcare app policies exceeded the 9th-grade level, hindering user comprehension.. Javed and Sajid (2024)(3) further emphasized that inconsistent formats and complex legal terminology pose barriers to informed consent.

Some efforts have attempted to quantify compliance with legal standards. Hakiem et al. (2024)(1) reviewed over 50 mHealth apps and reported widespread non-compliance with core GDPR and HIPAA requirements, especially around data retention and third-party sharing. Similarly, Benjumea et al. (2020)(2) applied a GDPR-based checklist to cancer app policies, finding that many lacked key disclosures like data controller identity or retention duration.

While these studies contribute valuable methods, they primarily use rule-based or supervised techniques, often requiring annotated training data or manual feature design. Unsupervised approaches, particularly clustering and pattern discovery from structured representations, remain underexplored.

C. Research Gaps

Despite progress in privacy policy analysis, several limitations persist in current approaches: Overreliance on Supervised Models: Tools like PolicyGPT and PolicyLint depend on annotated datasets or predefined rule sets, making them difficult to adapt to new domains without significant retraining or customization (Tang et al., 2023(7); Andow et al., 2019)(5).

Limited Use of Unsupervised Learning: Few studies have applied unsupervised machine learning techniques, such as clustering, to identify structural similarities or thematic groupings in privacy policies. This gap limits the discovery of latent patterns that could inform transparency scoring or regulation enforcement.

Insufficient Pattern Analysis in mHealth Policies: While many works evaluate policy readability or compliance, they do not explore how different apps group together based on the types of data collected or purposes disclosed — a critical aspect in understanding privacy risk at scale (Hakiem et al., 2024(1); Javed and Sajid, 2024)(3).

Vagueness and Lack of Specificity: Many mHealth privacy policies contain ambiguous phrases like "personal information" or "unspecified data," which weakens transparency and undermines user autonomy (Farooq et al., 2020(6); Benjumea et al., 2020)(2).

These gaps highlight the need for scalable, domain-aware approaches that can detect disclosure patterns across a broad set of healthcare applications — ideally without requiring manually labeled data.

D. Positioning

This study addresses the limitations in prior research by introducing an unsupervised, data-driven approach to analyzing healthcare app privacy policies. Unlike existing tools that rely on manual rule sets or supervised classifiers, our method uses structured knowledge graph representations combined with clustering techniques to identify thematic groupings in policy disclosures.

Building on tools like PoliGraph (Cui et al., 2023)(4), we extract key relationships—such as actor, data type, and purpose—from privacy policies and convert them into structured datasets. By applying K-Means clustering and dimensionality reduction methods like UMAP and t-SNE, we reveal hidden patterns in how health apps describe data collection and sharing. This approach provides a scalable framework for comparing privacy practices across applications without requiring prior labeling or regulatory rule encoding.

Furthermore, our exploratory analysis surfaces vague or ambiguous disclosures (e.g., "unspecified data"), highlighting policy inconsistencies that could impact user trust and legal compliance. In doing so, our work complements supervised policy analysis efforts (e.g., PolicyGPT, PolicyLint) by offering interpretability, pattern discovery, and generalizability—especially in domains with sensitive data like mHealth.

III. Methodology

A. Data Acquisition and Description

To analyze privacy policy disclosures in healthcare applications, we manually curated a dataset of health-related mobile apps available on major app marketplaces. Using the keyword "health", we identified 127 apps, from which we attempted to extract their privacy policies. The overall pipeline is illustrated in Figure 1, which shows the transition from raw policy acquisition through to clustering and final interpretation using semantic groupings.

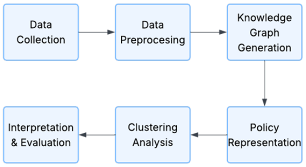


Fig. 1. Overview of the privacy policy analysis pipeline, including data collection, preprocessing, knowledge graph generation, clustering, and post-analysis interpretation.

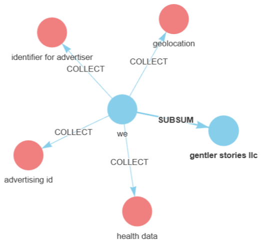


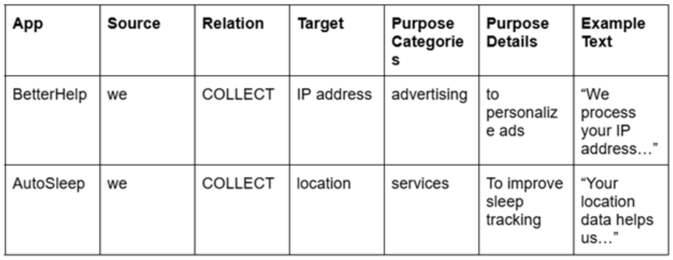
Fig. 2. Example of a Knowledge Graph

We used the PoliGraph-er tool (Cui et al., 2023) to scrape, parse, and structure the privacy policies. This tool transforms unstructured HTML text into knowledge graph representations, capturing key entities such as:

- Source (e.g., "we", "third parties")
- Relation (e.g., COLLECT, SUBSUME)
- Target (e.g., IP address, email, geolocation)
- Purpose (e.g., analytics, advertising, services)

Out of the 127 apps:

- 100 privacy policies were successfully processed and structured.
- 27 apps were excluded due to missing policies or scraping failures.



App	Source	Relation	Target	Purpose Categorie s	Purpose Details	Example Text
BetterHelp	we	COLLECT	IP address	advertising	to personaliz e ads	"We process your IP address…"
AutoSleep	we	COLLECT	location	services	To improve sleep tracking	"Your location data helps us…"

Fig. 3. Structured CSV

The final dataset consisted of several structured policy statements, stored as rows in a unified CSV file. Each row represents one data practice, encoded as a subject–relation–object triple with an associated purpose and example text from the policy. This structured format enabled direct application of data wrangling, exploratory analysis, and machine learning techniques. By focusing on healthcare apps, the dataset emphasizes a high-risk domain where user data sensitivity is paramount. The extracted policy elements serve as a foundation for identifying common disclosure patterns and variations across the mHealth privacy landscape.

B. Feature Engineering

To prepare the structured privacy policy data for unsupervised learning, we performed a series of feature engineering steps that transformed textual and categorical fields into numerical representations suitable for clustering. The key features extracted from the dataset included:

- **Target** – the type of data being collected or referenced (e.g., email, IP address, geolocation)
- **Purpose Category** – the stated purpose for collecting the data (e.g., services, advertising, analytics)
- **Relation** – the action taken on the data (e.g., COLLECT, SUBSUME)
- **Source** – the entity performing the action (e.g., app provider, third party)

The following transformations were applied:

- **TF-IDF Vectorization:** The Target field, which contains textual data, was converted into a high-dimensional numerical feature space using Term Frequency–Inverse Document Frequency (TF-IDF). This captured the importance of each data type across the corpus of policy statements.
- **One-Hot Encoding:** Categorical fields — Purpose Category, Relation, and Source — were one-hot encoded to retain the discrete nature of these attributes while making them compatible with clustering algorithms.
- **Cleaning and Normalization:** Missing values were removed, and multi-word phrases (e.g., "personal information") were replaced with underscores to ensure consistency during vectorization.

This engineered feature set captures both semantic and structural properties of privacy policies. It enables unsupervised learning methods to group similar disclosures and detect patterns in how different apps document their data collection practices.

C. Model Selection

To uncover patterns in privacy policy disclosures, we selected unsupervised machine learning models that support thematic clustering and high-dimensional data visualization. The primary model used was:

K-Means Clustering: A centroid-based clustering algorithm chosen for its efficiency and interpretability in grouping similar observations. K-Means is well-suited for structured, vectorized feature sets like those derived from our privacy policy data.

To support visualization and enhance interpretability, we also applied:

- **Truncated Singular Value Decomposition (SVD)**: Used for initial dimensionality reduction, especially on sparse TF-IDF vectors, to compress the feature space while preserving variance.
- **Uniform Manifold Approximation and Projection (UMAP)** and **t-distributed Stochastic Neighbor Embedding (t-SNE)**: Applied to the reduced feature space to produce 2D projections that visually represent the similarity structure among policy statements.

These models were selected for the following reasons:

- **K-Means** effectively captures cluster centroids, making it easy to interpret representative disclosure patterns.
- **UMAP** and **t-SNE** preserve both local and global relationships in high-dimensional data, providing intuitive visualizations that help explain clustering behavior.
- **SVD** enables efficient preprocessing of large TF-IDF matrices, reducing noise before visualization.

This combination of methods allows for both quantitative grouping of policies and qualitative insight through visualization, supporting a well-rounded analysis of mHealth privacy disclosures.

D. Experimental Setup

The experimental workflow was structured to ensure reproducibility and consistent evaluation across all stages of analysis. The setup included the following key components:

1) Data Splitting: As an unsupervised learning task, no traditional train-test split was required. Instead, the full dataset was used for clustering and visualization.

2) Preprocessing Pipeline: **Data Cleaning**: Rows with missing or null values in key fields (e.g., Target, Purpose Category) were removed. **Normalization**: Multi-word phrases were standardized by replacing spaces with underscores (e.g., "personal information" → "personal_information"). **Feature Transformation**:

- One-hot encoding for Purpose Category, Relation, and Source.
- TF-IDF vectorization for Target field

3) ***Clustering Procedure:***

- K-Means was applied to the combined feature matrix.
- The elbow method was used to determine the optimal number of clusters, revealing that **k = 3** offered the best balance between compactness and separation.

4) Dimensionality Reduction and Visualization :

- SVD reduced the TF-IDF feature space to 50 components.
- UMAP and t-SNE were then applied to the reduced data to create 2D cluster plots.
- Visualizations were implemented using Matplotlib.

5) Tools and Libraries:

- Python 3.10 with libraries including scikit-learn, pandas, numpy, matplotlib, seaborn, umap-learn, and yaml for graph parsing.
- Experiments were conducted in Jupyter Notebook for modular development and result tracking.

This experiment design ensured that all steps, from data ingestion to clustering, followed a consistent and reproducible methodology aligned with the project's objectives.

E. Evaluation Metrics

Since this study employs unsupervised learning, traditional accuracy-based evaluation is not applicable. Instead, the quality and interpretability of the clustering results were assessed using the following metrics and methods:

1) Elbow Method: To determine the optimal number of clusters for K-means, we applied the elbow method, which plots the within-cluster sum of squares (inertia) against the number of clusters. The ”elbow point” indicated where additional clusters yield diminishing improvements in compactness. Based on this method, K=3 was selected as optimal.

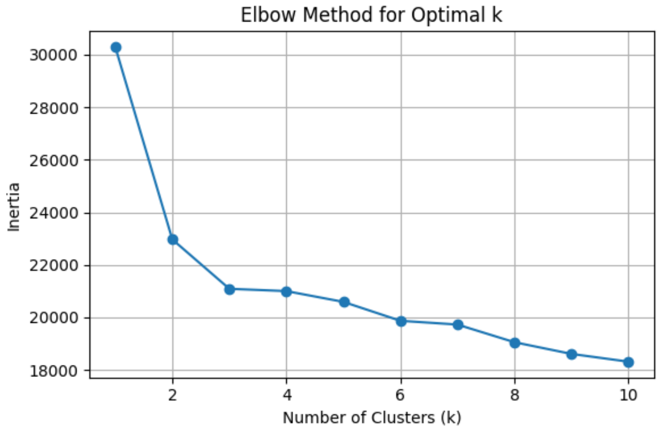


Fig. 4. Elbow Method for Optimal K

2) Silhouette Score: We computed the Silhouette Coefficient, which measures how similar each data point is to its own cluster versus other clusters. Values closer to 1.0 indicate well-separated, dense clusters. This score was used to confirm cluster cohesion and separation.

3) Cluster Interpretability: Interpretation was supported by:

- Top keywords (from TF-IDF features) in each cluster.
- Dominant purpose categories and relations (from one-hot encoded attributes).
- Visual inspection of UMAP and t-SNE plots to assess how well data points grouped based on policy semantics.

4) Visualization Validation: Visual plots generated using UMAP and t-SNE provided qualitative validation by displaying how distinctly clusters formed. These helped identify whether policies with vague disclosures or similar purposes grouped naturally.

These metrics ensured both technical soundness and semantic validity of the clusters, aligning the outcomes with the study’s goal of revealing hidden disclosure patterns in healthcare privacy policies.

IV. Results

A. Main Results

The analysis of 107 healthcare app privacy policies revealed several key trends in data practices, supported by visual and statistical findings.

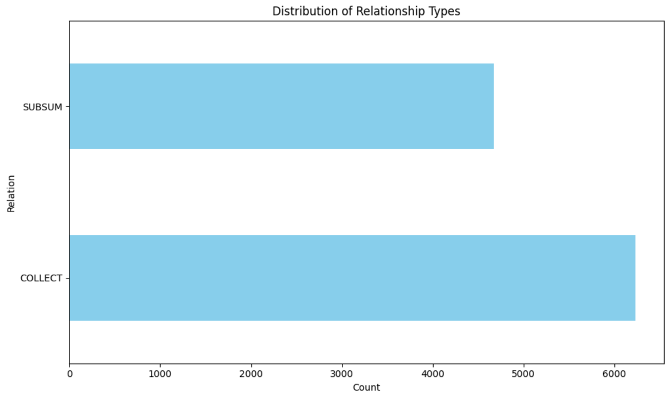


Fig. 5. Distribution of Relationship Types

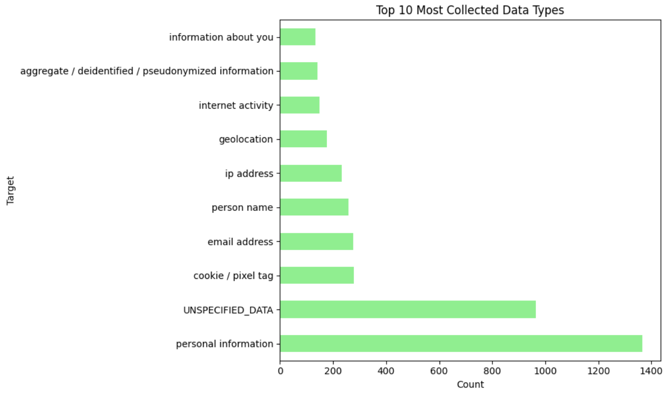


Fig. 6. Top 10 Most Collected Data Types

1) Relationship Types and Data Focus: As shown in Figure 5, the most common relationship type was COLLECT, indicating that most policies disclosed active data gathering by the app or third parties. SUBSUM relations, used to represent data aggregation or inclusion, were also frequent but less dominant.

Figure 6 highlights the Top 10 most collected data types, with “personal information” and “UNSPECIFIED_DATA” leading the list. This reflects a trend toward vague or generalized disclosures, limiting transparency and user understanding.

2) Purpose of Data Use: According to Figure 7 , nearly 50% of policy statements lacked a clearly defined purpose (unspecified). The most common explicit purposes were services (15.4%), analytics (10.2%), and advertising (8.9%). This confirms earlier findings that many apps collect data for non-essential or monetization-driven goals.

3) Language Patterns: The word cloud in Figure 8 visualizes recurring language in the policy statements. Common terms like “use,” “collect,” “email address,” “third party,” and “service provider” reinforce the dominance of generalized disclosures and third-party data involvement.

4) Clustering Results: Clustering was performed on the structured privacy policy records using a feature set consisting

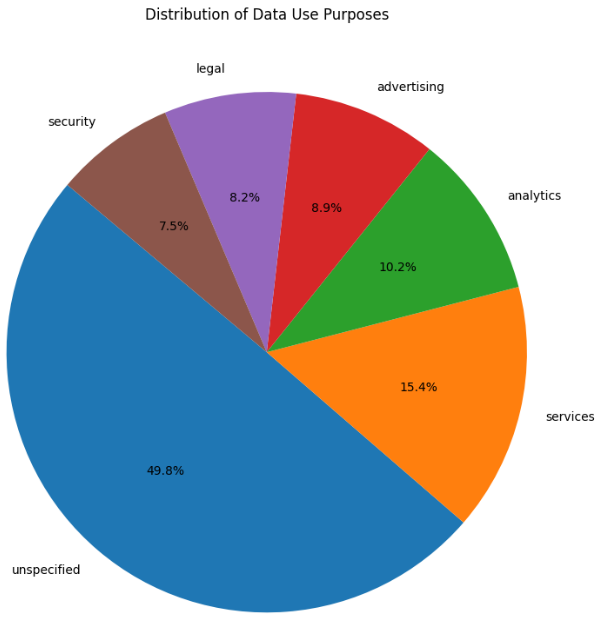


Fig. 7. Distribution of Data Use Purposes

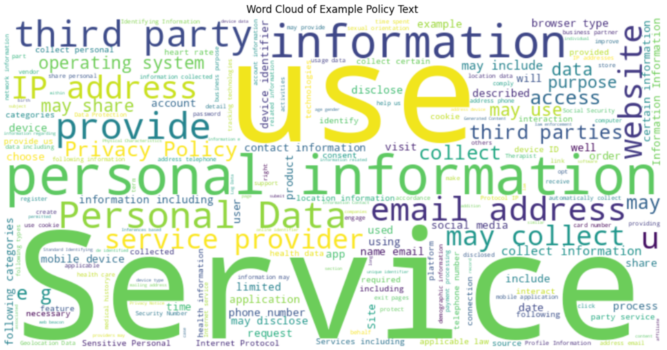


Fig. 8. Word Cloud of Example Policy Text

of Target, Purpose Categories, Relation, and Source. After preprocessing and vectorization, K-Means clustering with k = 3 produced three distinct clusters, evaluated both quantitatively and semantically.

5) Top-Level Trends:

- The most common relation was COLLECT, followed by SUBSUM, reflecting direct versus hierarchical disclosures.
- Frequent data types included personal_information, unspecified_data, cookie_, and email_address.
- Nearly 50% of disclosures had vague or undefined purposes (tagged as unspecified).

6) Cluster 0: Hierarchically Structured Identifiers:

- **Top Targets:** email_address, person_name, ip_address, postal_address, phone_number
- **Top Relation:** SUBSUM

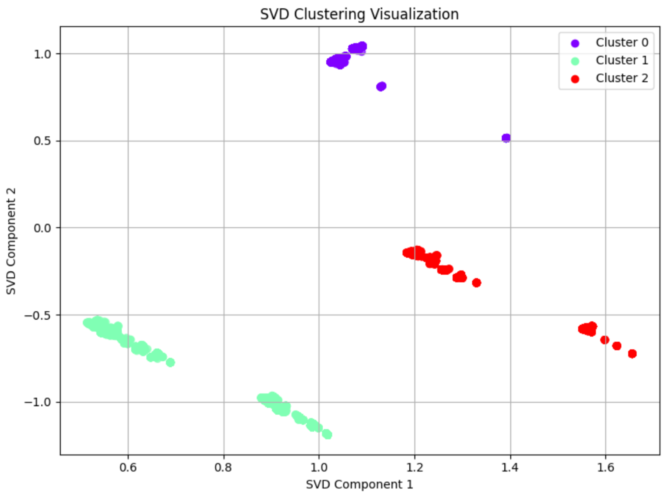


Fig. 9. SVD Clustering

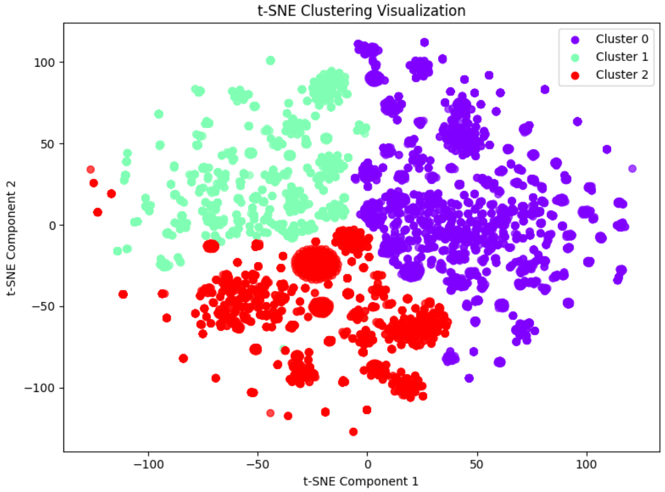


Fig. 10. t-SNE Clustering

- **Dominant Purpose:** personal_information
- **Dominant Parent Entities:** tracking_data, general_data.
- **Interpretation:** This cluster groups disclosures that are components of a broader concept like personal_information. These are not direct collections, but elements subsumed within a hierarchical structure, as seen in object-oriented models. Their presence supports transparency in structure, but often lacks accompanying purpose explanation, which may reduce clarity.

*7) **Cluster 1: Functional Data Collection with Vague Containers**:*

- **Top Targets**: personal_information, unspecified_data, _pixel_tag, cookie_, information_about_you
- **Top Relation**: COLLECT
- **Dominant Purpose**: services
- **Dominant Parent Entities**: tracking_data, general_data

Interpretation: Policies in this cluster reference broad containers like personal_information and information_about_you,

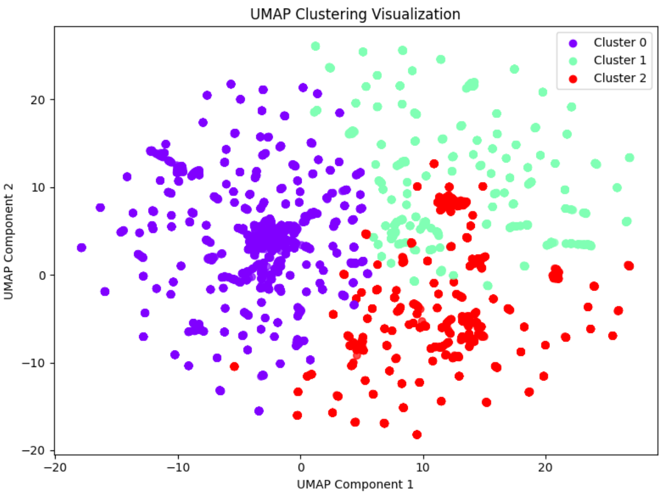


Fig. 11. UMAP Clustering

often collected for platform functionality. Although they state a purpose (services), the use of ambiguous or high-level target terms may hinder informed consent.

*8) **Cluster 2: Behavioral and Device-Linked Tracking**:*

- **Top Targets**: unspecified_data, internet_activity, information_about_you, geolocation
- **Top Relation**: COLLECT
- **Dominant Purpose**: unspecified
- **Dominant Parent Entities**: behavioral_data, location_data, ambiguous_data
- **Interpretation**: This cluster reflects behavioral surveillance — collecting sensitive data like user activity and location with little clarity on intent. The dominance of unspecified purposes raises concerns over transparency, consent, and compliance.

*9) **Evaluation Scores**:* The Silhouette scores showed strong separation for **SVD (0.8008)**, while **t-SNE (0.3464)** and **UMAP (0.2937)** offered moderate clustering validation. Trustworthiness scores for all visualizations were high (all above **0.95**), confirming that dimensionality reduction retained the structure of the original space.

These metrics confirm that the clustering is both structurally coherent and visually faithful to the high-dimensional space.

B. Qualitative Analysis

The three discovered clusters not only differed in technical features like data type and relation but also revealed structural and semantic themes within privacy policy disclosures. These themes become especially apparent when we interpret clusters using a Parent Entity mapping that reflects hierarchical relationships (e.g., email_address as part of personal_information).

*a) **Cluster 0: Hierarchical Disclosure of Identifiable Data**:* This cluster primarily featured targets such as email_address, person_name, and ip_address connected via the SUBSUM relation. These are not independent data items, but components of a higher-level concept — personal_information.

The structure of these disclosures aligns with a hierarchical data model, where the policy lists parts of an object without necessarily describing the object itself. For example, a policy might say it collects email_address and IP address, which are both attributes of a user's identity. However, the lack of accompanying purposes or collection justification limits transparency. This cluster reflects technical completeness but semantic ambiguity, as the hierarchy is clear, but the reasoning behind data usage is not.

*b) **Cluster 1: Generalized Containers with Functional Purpose**:* Cluster 1 includes high-level targets like unspecified_data, personal_information, cookie_, and information_about_you. These are linked with the COLLECT relation and commonly paired with the purpose services.

While these policies name what they collect, the terms are often overly broad. For instance, information_about_you is semantically rich, but without decomposition into specific fields, it tells users very little. These disclosures fulfill the formal requirement of a privacy policy but risk obfuscating the exact nature of the data being collected. However, the presence of a functional purpose (services) gives them more clarity than Cluster 0.

*c) **Cluster 2: Behavioral Surveillance with Low Transparency**:* This cluster contains terms like internet_activity, geolocation, and unspecified_data—targets often associated with user tracking and behavioral profiling. These are collected (COLLECT relation) without a clearly stated purpose (unspecified is dominant).

Semantically, these disclosures are the most concerning. The data types are sensitive, but the lack of justification or context raises red flags for both regulatory compliance and ethical standards. Moreover, grouping under parent entities like behavioral_data and location_data reveals that this cluster focuses on tracking users, often silently or ambiguously.

*d) **Hierarchy-Driven Insights**:* By analyzing Parent_Entity associations post-clustering, we preserved the semantic intent of the original data. This allowed us to:

- See how clusters group disclosures by function (services) vs. structure (subsumed elements).
- Reveal which clusters rely on broad, vague containers.
- Clarify that terms like personal_information are containers, not endpoints, aligning with object-oriented logic.

V. Discussion

A. Interpretation of Results

TABLE I
Top Parent Entities per Cluster

Cluster	Top Parent Entities
0	*personal_information*
1	*tracking_data, general_data*
2	*behavioral_data, location_data, ambiguous_data*

The clustering and qualitative analysis revealed three distinct strategies in how health app privacy policies present data collection practices. Cluster 0 grouped hierarchically structured identifiers, typically disclosed using the SUBSUM relation. This suggests that these policies recognize data structure but often fail to provide usage context, weakening user comprehension.

Cluster 1, focused on broad container terms like personal_information and unspecified_data, provided slightly more transparency by linking disclosures to services. However, the use of high-level terms without detailed breakdowns still limits interpretability.

Cluster 2 demonstrated opaque behavioral tracking, collecting sensitive information such as internet_activity and geolocation with little explanation of purpose. This pattern raises the most serious concerns regarding transparency, trust, and regulatory compliance.

B. Comparison with Prior Work

Our findings align with those of Hakiem et al. (2024) and Benjumea et al. (2020), who found that many mHealth policies lack detail and fail to meet GDPR standards. Similarly to PolicyLint (Andow et al., 2019), our results show that ambiguity and inconsistency persist in many privacy policies. However, unlike most previous studies, we leveraged unsupervised clustering and hierarchical interpretation to uncover latent patterns without relying on labeled data or rule-based scoring systems.

C. Strengths

Our approach respects the semantic hierarchy of data disclosures, aligning with real-world policy structures.

We applied a flexible interpretation layer (Parent_Entity) without altering model inputs, preserving clustering integrity.

The pipeline is scalable and reproducible, enabling application to larger or more diverse policy datasets

D. Limitations

The Parent_Entity mapping relies on rule-based heuristics, which may not capture all edge cases or domain-specific terms.

Some policy statements (e.g., unspecified_data) remain difficult to resolve, even after interpretation.

Clustering quality is partly dependent on the vectorization of text fields, which may miss deeper semantic relationships.

E. Implications and Recommendations

Our results suggest that many healthcare privacy policies use broad or vague language to obscure data practices. Even when technically compliant, these policies may fail to support informed consent or data autonomy. We recommend that app developers:

- Break down container terms into specific, user-understandable data types.
- Pair every data disclosure with a clearly stated purpose.
- Avoid overuse of ambiguous terms like “information about you” without clarification.

For researchers, future work could explore embedding ontology-aware models or LLMs to dynamically interpret policy structure and intent.

VI. Conclusion

A. Summary of Findings

This study investigated how mobile health applications disclose data collection and usage in their privacy policies. Using a combination of knowledge graph extraction, unsupervised clustering, and hierarchical interpretation, we analyzed 100 structured privacy policy records.

The analysis revealed three distinct clusters:

- **Cluster 0** grouped subcomponents of personal information (e.g., email, IP address) using SUBSUM, reflecting technical structure but often lacking explanatory context.
- **Cluster 1** used broad container terms like unspecified_data or personal_information, paired with functional purposes such as services. While slightly more transparent, these disclosures often remained abstract.
- **Cluster 2** involved behavioral tracking and sensitive data (e.g., geolocation, internet activity), but with minimal purpose justification — a pattern of low transparency and high risk.

By introducing a Parent Entity interpretation layer, we preserved and respected the hierarchical structure of the original data while offering clearer semantic insight. Our findings highlight the variability, vagueness, and structural inconsistencies across mHealth privacy policies, reinforcing concerns raised in prior literature and offering a novel method for analyzing disclosure patterns at scale.

B. Significance

This study demonstrates the value of combining unsupervised learning with semantic interpretation to assess privacy policies in the healthcare domain. While many prior approaches rely on supervised models or predefined rule sets, our method uses clustering to uncover natural patterns in how apps disclose data practices—without requiring labeled data.

A key innovation of our work is the integration of a Parent Entity interpretation layer, which allows us to respect the hierarchical nature of data disclosures (e.g., treating email_address as part of personal_information). This helped us interpret the structure and intent of disclosures more accurately, without modifying the clustering process.

By revealing how some policies are structurally complete but semantically vague, and others are explicitly behavioral but lacking in purpose, our results highlight important privacy risks. The framework is adaptable, transparent, and well-suited for large-scale analysis—making it a useful tool for researchers, developers, and regulators who seek to improve policy clarity and user trust.

C. Future Work

Future research can build on this work in several directions such as :

Privacy Transparency Scoring: Building on this structure, a scoring system could be developed to rank apps based on the clarity, completeness, and specificity of their privacy disclosures, providing users with an interpretable risk signal.

Larger and Multi-Domain Datasets: Expanding this analysis to non-healthcare domains (e.g., finance, education) or across app platforms could validate the model's generalizability and reveal sector-specific disclosure patterns.

References

[1] 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.

[2] J. Benjumea, J. Ropero, O. Rivera-Romero, E. Dorronzoro-Zubiete, and A. Carrasco, "Privacy assessment in mobile health apps: Scoping review," vol. 8, no. 7, p. e18868.

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

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

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

[6] 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.

[7] 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."

Appendix

https://github.com/DiabeticDonut/Privacy-Policy-Analysis---CSCI-7090-Team5