World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Efficient compliance with GDPR through automating privacy policy captions in web and mobile application

Trudy-Ann Campbell [1], Samson Eromonsei [1] and Olusegun Afolabi [2]

[1] *School of Engineering Prairie View, A and M University Prairie View, Texas USA.*
[2] *Department of Information Systems and Business Analysis, Aston Business School, Aston University, Birmingham, UK.*

World Journal of Advanced Engineering Technology and Sciences, 2024, 12(02), 446–467

Publication history: Received on 15 June 2024; revised on 26 July 2024; accepted on 29 July 2024

Article DOI: https://doi.org/10.30574/wjaets.2024.12.2.0317

Abstract

Ensuring compliance with the General Data Protection Regulation (GDPR) presents significant challenges for organizations, especially those developing web and mobile applications. This study investigates the use of automation to enhance GDPR compliance by generating privacy policy captions through static code analysis and deep learning models. Privacy policy captions offer concise, user-friendly summaries of data processing practices, improving transparency and user trust. The research combines qualitative and quantitative methodologies, including static code analysis of application source codes and the application of neural machine translation models to generate privacy policy captions. Findings indicate that automation can effectively produce accurate, consistent, and comprehensible privacy policy captions that align with GDPR requirements. However, limitations such as tool capabilities, dataset diversity, and user testing scale highlight areas for future research. This study provides practical guidelines for implementing automated privacy policy captions, emphasizing the importance of continuous monitoring and updates to maintain compliance. By leveraging automation, organizations can enhance their data protection practices, build user trust, and achieve efficient GDPR compliance.

Keywords: GDPR compliance; Privacy policy automation;Static code analysis; Neural machine translation; Data protection

1. Introduction

1.1. Background of GDPR and its Importance

The General Data Protection Regulation (GDPR), enacted by the European Union (EU) in May 2018, represents a significant overhaul of data protection laws aimed at safeguarding individuals' personal data and ensuring privacy rights across member states. The GDPR was introduced to address the growing concerns over data privacy in the digital age, where personal data is often collected, processed, and stored by organizations without adequate transparency or user consent (Voigt & von dem Bussche, 2017).

One of the core motivations behind the GDPR is to empower individuals with greater control over their personal data, ensuring that they are informed about how their data is used and providing them with the rights to access, rectify, and erase their data (Albrecht, 2016). The regulation applies to all organizations operating within the EU, as well as those outside the EU that offer goods or services to EU citizens, making it one of the most comprehensive data protection frameworks globally (Voigt & von dem Bussche, 2017).

* Corresponding author: Trudy-Ann Campbell

The GDPR's importance is underscored by the substantial fines and penalties for non-compliance, which can reach up to €20 million or 4% of an organization's global annual turnover, whichever is higher (Albrecht, 2016). This stringent enforcement mechanism has prompted organizations to re-evaluate their data protection practices and implement robust measures to ensure compliance. As a result, the GDPR has become a benchmark for data privacy regulations worldwide, influencing similar legislation in other regions (White & Case, 2018).

1.2. Challenges in GDPR Compliance for Web and Mobile Applications

Compliance with GDPR presents several challenges for organizations, particularly those developing and managing web and mobile applications. One of the primary challenges is the requirement for explicit user consent for data collection and processing activities. Applications must ensure that consent is obtained through clear and unambiguous mechanisms, often necessitating significant changes to existing user interfaces and workflows (Tankard, 2016). This involves not only creating consent forms that are easily understandable but also ensuring that users can withdraw their consent as easily as it was given (Tikkinen-Piri, Rohunen, & Markkula, 2018).

Another significant challenge is the implementation of data protection by design and by default, which mandates that data protection measures are integrated into the development processes of web and mobile applications from the outset. This requires developers to adopt privacy-preserving technologies and practices, such as data minimization, pseudonymization, and encryption, which can be technically complex and resource-intensive (Tankard, 2016). Additionally, maintaining comprehensive records of processing activities and ensuring compliance with data subject rights, such as the right to access and the right to be forgotten, adds further layers of complexity (Albrecht, 2016).

Furthermore, the dynamic nature of web and mobile applications, which often involve continuous updates and feature enhancements, poses ongoing compliance challenges. Each update or new feature must be assessed for GDPR compliance, requiring robust governance and monitoring frameworks. Organizations must also be vigilant about third-party integrations and data sharing practices, as they are equally responsible for ensuring that any third parties they collaborate with are GDPR compliant (Voigt & von dem Bussche, 2017). This necessitates thorough vetting and regular audits of third-party services, adding to the operational burden.

1.3. Overview of Privacy Policy Caption

Privacy policy captions are concise summaries or snippets that provide users with essential information about an application's data collection and processing practices. These captions play a crucial role in enhancing transparency and ensuring that users are adequately informed about how their personal data will be used. The GDPR emphasizes the importance of transparency and mandates that privacy policies must be easily accessible, clear, and understandable to users, thereby making privacy policy captions an effective tool for compliance (Voigt & von dem Bussche, 2017).

One of the main purposes of privacy policy captions is to simplify complex legal jargon, making it easier for users to comprehend the core aspects of data processing activities. This is particularly important for web and mobile applications, where users may be deterred by lengthy and complicated privacy statements. Captions can highlight key points such as the types of data collected, the purposes of data processing, data sharing practices, and users' rights under the GDPR (Albrecht, 2016). By presenting this information in a clear and concise manner, privacy policy captions help in building user trust and promoting informed consent (Tikkinen-Piri, Rohunen, & Markkula, 2018).

Moreover, the use of privacy policy captions aligns with the GDPR's requirement for data protection by design and by default. Integrating these captions into the user interface of web and mobile applications ensures that privacy information is presented at relevant points of interaction, such as during account creation or when requesting permissions for data access. This proactive approach not only facilitates compliance but also enhances the overall user experience by making privacy practices more transparent and accessible (Tankard, 2016). Effective privacy policy captions can thus serve as a bridge between legal requirements and user engagement, ensuring that both compliance and user needs are adequately addressed.

1.4. Problem Statement

The technology industry today is plagued by a lack of properly formatted privacy policies for individual apps, most especially with software industries that are continuously rolling out new products. There a massive inconsistency in the posted privacy notices of what the flow graph/call graph in software codes does compared to the PII data collected. A lot of these software applications are becoming over-privileged applications due to nonproper privacy notice disclosure.

1.5. Overview of Privacy Policy Captions

Privacy policy captions are concise summaries or snippets that provide users with essential information about an application's data collection and processing practices. These captions play a crucial role in enhancing transparency and ensuring that users are adequately informed about how their personal data will be used. The GDPR emphasizes the importance of transparency and mandates that privacy policies must be easily accessible, clear, and understandable to users, thereby making privacy policy captions an effective tool for compliance (Voigt & von dem Bussche, 2017).

One of the main purposes of privacy policy captions is to simplify complex legal jargon, making it easier for users to comprehend the core aspects of data processing activities. This is particularly important for web and mobile applications, where users may be deterred by lengthy and complicated privacy statements. Captions can highlight key points such as the types of data collected, the purposes of data processing, data sharing practices, and users' rights under the GDPR (Albrecht, 2016). By presenting this information in a clear and concise manner, privacy policy captions help in building user trust and promoting informed consent (Tikkinen-Piri, Rohunen, & Markkula, 2018).

Moreover, the use of privacy policy captions aligns with the GDPR's requirement for data protection by design and by default. Integrating these captions into the user interface of web and mobile applications ensures that privacy information is presented at relevant points of interaction, such as during account creation or when requesting permissions for data access. This proactive approach not only facilitates compliance but also enhances the overall user experience by making privacy practices more transparent and accessible (Tankard, 2016). Effective privacy policy captions can thus serve as a bridge between legal requirements and user engagement, ensuring that both compliance and user needs are adequately addressed.

1.6. Significance of Automation in Compliance Processes

Automation plays a critical role in ensuring efficient and effective compliance with GDPR, particularly in the context of web and mobile applications. One of the primary benefits of automation is the ability to streamline and standardize compliance processes, reducing the risk of human error and ensuring consistent application of data protection measures across various platforms and services (Tankard, 2016). By automating tasks such as consent management, data subject request handling, and privacy policy updates, organizations can significantly enhance their ability to meet GDPR requirements in a timely and accurate manner (Voigt & von dem Bussche, 2017).

The use of automation tools also enables organizations to manage large volumes of data more effectively. For instance, machine learning algorithms can be employed to identify and categorize personal data within an organization's systems, facilitating compliance with data inventory and mapping requirements. These tools can also help in detecting potential compliance issues, such as unauthorized data access or processing activities, allowing for prompt corrective actions (Tikkinen-Piri, Rohunen, & Markkula, 2018). This proactive approach not only helps in maintaining compliance but also minimizes the risk of data breaches and associated penalties.

Furthermore, automation supports the implementation of data protection by design and by default principles, which are central to GDPR compliance. Automated systems can be configured to enforce privacy settings and data minimization practices by default, ensuring that only necessary data is collected and processed (Tankard, 2016). Additionally, automated privacy policy captions can provide real-time updates to users about changes in data processing practices, enhancing transparency and user trust. This dynamic approach to privacy management aligns with the GDPR's emphasis on continuous monitoring and improvement of data protection practices (Voigt & von dem Bussche, 2017).

Overall, the significance of automation in GDPR compliance lies in its ability to enhance operational efficiency, ensure consistency, and provide robust mechanisms for monitoring and managing data protection practices. By leveraging automation, organizations can not only achieve compliance more effectively but also foster a culture of privacy and trust among their users.

1.7. Objectives and Scope of the Research

The primary objective of this research is to explore how automation can be leveraged to ensure efficient compliance with GDPR, specifically through the implementation of automated privacy policy captions in web and mobile applications. This research aims to identify the key benefits and challenges associated with automating privacy policy captions and to develop best practices for their effective implementation.

The scope of this research includes an in-depth analysis of current GDPR requirements related to privacy policies and user consent, as well as the technical and operational aspects of integrating automated captions into web and mobile application interfaces. The study will examine various automation tools and technologies, assess their applicability and effectiveness, and provide case studies of successful implementations.

Additionally, this research will explore the impact of automated privacy policy captions on user experience and trust, and how they contribute to overall compliance efforts. By providing a comprehensive understanding of the role of automation in GDPR compliance, this research aims to offer valuable insights and practical guidelines for organizations looking to enhance their data protection practices through automation.

2. GDPR requirements for privacy policies

2.1. Key GDPR Provisions Relevant to Privacy Policies

The General Data Protection Regulation (GDPR) includes several key provisions that directly impact the creation and management of privacy policies for web and mobile applications. One of the fundamental principles is the requirement for transparency, which mandates that organizations must provide clear, concise, and easily accessible information about their data processing activities to users (Voigt & von dem Bussche, 2017). This includes detailing the types of personal data collected, the purposes for which the data is processed, and any third parties with whom the data may be shared.

Another crucial provision is the necessity for explicit and informed consent from users before collecting and processing their personal data. The GDPR requires that consent must be freely given, specific, informed, and unambiguous, meaning that organizations must ensure users are fully aware of what they are consenting to and must be able to withdraw consent at any time (Tankard, 2016). This is particularly important for web and mobile applications that often collect large amounts of personal data, necessitating robust mechanisms for obtaining and managing user consent.

Additionally, the GDPR emphasizes the rights of data subjects, including the right to access, rectify, and erase their personal data, as well as the right to data portability and the right to object to certain types of data processing (Tikkinen-Piri, Rohunen, & Markkula, 2018). Privacy policies must clearly outline these rights and provide straightforward instructions on how users can exercise them. This ensures that users have control over their personal data and can hold organizations accountable for their data handling practices.

These key provisions underscore the importance of designing privacy policies that are not only compliant with GDPR but also user-friendly and transparent. By adhering to these requirements, organizations can build trust with their users and demonstrate their commitment to protecting personal data.

2.2. Mandatory Elements of Privacy Policies

Privacy policies under the GDPR must include several mandatory elements to ensure comprehensive compliance and transparency. Firstly, organizations are required to provide their identity and contact details, including those of any data protection officers (DPOs) appointed to oversee compliance efforts (Voigt & von dem Bussche, 2017). This ensures that users can easily contact the organization regarding any data protection concerns.

Secondly, the purposes for which personal data is collected and processed must be clearly stated. This includes specifying the legal basis for processing, such as consent, performance of a contract, legal obligation, vital interests, public interest, or legitimate interests pursued by the data controller or a third party (Tikkinen-Piri, Rohunen, & Markkula, 2018). By clearly outlining the purposes and legal grounds, organizations can provide users with a better understanding of how their data will be used.

Another critical element is detailing the categories of personal data collected. This involves describing the types of data collected directly from users and any data obtained from third-party sources (Tankard, 2016). Furthermore, organizations must disclose any recipients or categories of recipients to whom the personal data may be disclosed, ensuring transparency regarding data sharing practices.

Additionally, privacy policies must inform users about their rights under the GDPR. This includes the right to access, rectify, and erase personal data, the right to restrict or object to processing, the right to data portability, and the right to lodge a complaint with a supervisory authority (Tikkinen-Piri, Rohunen, & Markkula, 2018). Providing clear instructions on how users can exercise these rights is essential for compliance.

Lastly, organizations must include information on data retention periods or the criteria used to determine these periods. Users should be informed about how long their data will be stored and the reasons for specific retention periods (Voigt & von dem Bussche, 2017). By including these mandatory elements, privacy policies can effectively communicate key information to users and demonstrate a commitment to GDPR compliance.

2.3. User Consent and Data Subject Rights

User consent and data subject rights are central to the GDPR, emphasizing the importance of user autonomy and control over personal data. Under the GDPR, obtaining user consent is a critical requirement. Consent must be freely given, specific, informed, and unambiguous, and it must be given through a clear affirmative action, such as opting-in via a checkbox (Tankard, 2016). This ensures that users are fully aware of what they are consenting to and can make informed decisions about their personal data.

Furthermore, the GDPR grants data subjects several rights to empower them with control over their personal data. One of the fundamental rights is the right to access, which allows individuals to obtain confirmation as to whether their personal data is being processed and, if so, to access the data and receive information about the processing activities (Voigt & von dem Bussche, 2017). This transparency is crucial for building trust between users and organizations.

Another important right is the right to rectification, which enables data subjects to request the correction of inaccurate or incomplete personal data. This ensures that organizations maintain accurate and up-to-date information about their users (Tikkinen-Piri, Rohunen, & Markkula, 2018). Additionally, the right to erasure, also known as the right to be forgotten, allows individuals to request the deletion of their personal data when it is no longer necessary for the purposes for which it was collected or if the data subject withdraws consent (Voigt & von dem Bussche, 2017).

The GDPR also includes the right to restrict processing, which permits individuals to limit the processing of their personal data under certain circumstances, and the right to data portability, which allows data subjects to receive their personal data in a structured, commonly used, and machine-readable format and to transmit it to another data controller (Tikkinen-Piri, Rohunen, & Markkula, 2018). Finally, the right to object enables users to object to the processing of their personal data on grounds relating to their particular situation, including objections to profiling and direct marketing activities (Tankard, 2016).

These rights collectively ensure that users have robust control over their personal data, promoting transparency, accountability, and trust in the digital ecosystem.

2.4. Transparency and Accountability in Data Handling

Transparency and accountability are fundamental principles enshrined in the GDPR, designed to ensure that organizations handle personal data responsibly and ethically. Transparency requires organizations to be open about their data processing activities, providing clear and accessible information to individuals about how their data is collected, used, stored, and shared. This is achieved through comprehensive privacy policies and notices that are easy for users to understand (Voigt & von dem Bussche, 2017). The GDPR mandates that these documents must include detailed information on the identity of the data controller, the purposes of processing, the types of data collected, and the rights of data subjects (Tikkinen-Piri, Rohunen, & Markkula, 2018).

Accountability, on the other hand, obligates organizations to take responsibility for their data processing activities and to implement measures that ensure compliance with GDPR requirements. This includes adopting data protection policies, conducting data protection impact assessments (DPIAs), and appointing data protection officers (DPOs) where necessary (Tankard, 2016). Organizations must also maintain records of processing activities, demonstrating their commitment to data protection principles and their ability to comply with GDPR obligations (Voigt & von dem Bussche, 2017).

Additionally, the principle of accountability under the GDPR requires organizations to implement appropriate technical and organizational measures to secure personal data. This includes employing data encryption, pseudonymization, and ensuring data integrity and confidentiality (Tikkinen-Piri, Rohunen, & Markkula, 2018). Regular audits and monitoring of data processing activities are also essential to identify and address potential risks and compliance issues proactively (Tankard, 2016).

Transparency and accountability not only help in complying with legal requirements but also in building trust with users. When organizations are transparent about their data practices and demonstrate accountability, they foster a

sense of security and confidence among users, which is crucial in today's data-driven environment (Voigt & von dem Bussche, 2017).

2.5. Case Studies of Non-Compliance Penalties

The enforcement of the GDPR has led to several high-profile cases of non-compliance, resulting in significant penalties for organizations that failed to meet the regulation's requirements. One notable case involved Google, which was fined €50 million by the French data protection authority, CNIL, for lack of transparency, inadequate information, and lack of valid consent regarding personalized ads (Voigt & von dem Bussche, 2017). The fine was a clear signal of the GDPR's stringent stance on ensuring that users are fully informed and consent is appropriately obtained.

Another significant case was the British Airways data breach, where the UK Information Commissioner's Office (ICO) proposed a fine of £183 million. British Airways was found to have poor security arrangements that led to the exposure of personal data of approximately 500,000 customers (Tikkinen-Piri, Rohunen, & Markkula, 2018). This case highlighted the importance of robust security measures and the severe repercussions of failing to protect user data adequately.

Marriott International also faced a substantial fine of £99 million following a data breach that exposed the personal data of approximately 339 million guests. The breach was a result of Marriott's failure to undertake sufficient due diligence when it acquired Starwood Hotels, whose reservation database had been compromised years before (Tankard, 2016). This case underscored the necessity of thorough data protection assessments during mergers and acquisitions.

These cases demonstrate the broad scope of GDPR enforcement and the severe penalties that can result from non-compliance. They serve as critical reminders to organizations about the importance of transparency, obtaining valid consent, and implementing robust security measures to protect personal data. Furthermore, these penalties emphasize the need for ongoing vigilance and adherence to GDPR requirements to avoid significant financial and reputational damage (Voigt & von dem Bussche, 2017).

2.6. Common Privacy Policies Conflict SCENARIOs

- Smart Home emergence adopting IOT technology using similar developments procedure in mobile developments.
- Mobile Android application developments

2.6.1. Concepts Familiarization

- Privacy policy
- Privacy Notices
- GDPR
- Call Graph and Intermediate Representation
- Abstract syntax Tree AST
- Inter-procedure call flow graph ICFG

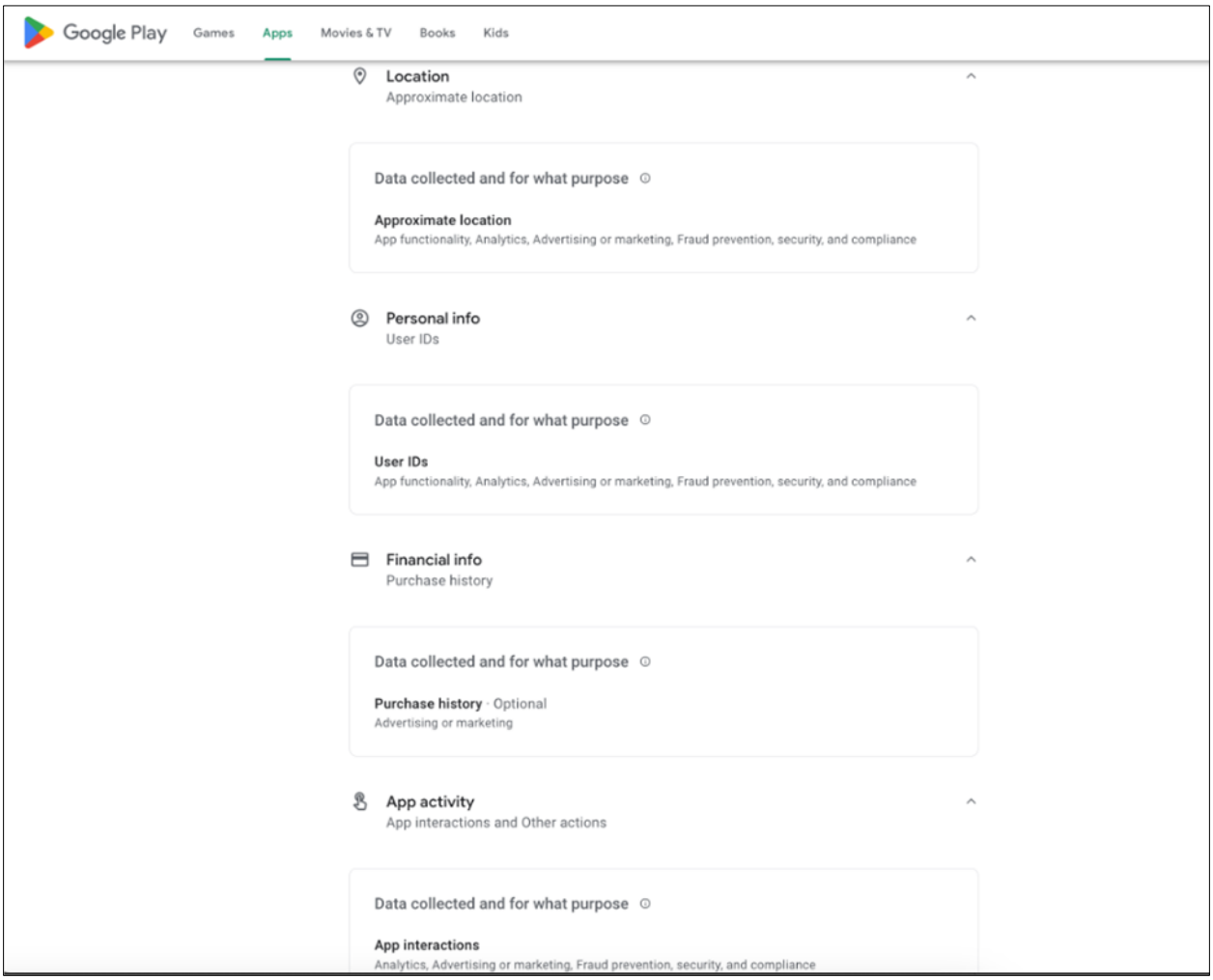
2.7. Privacy Policy

A comprehensive document that outlines how an organization

- Collects
- Uses
- Discloses
- Retain
- Users right exercise
- Manages user data

It provides a detailed overview of the data processing practices and the rights of the users.

2.7.1. Example for google play store privacy notices



2.8. Privacy Notices

A fair processing notice or privacy statement, is a concise and user-friendly summary of the key information from the privacy policy.

It is designed to inform users about data practices in a more accessible format.

2.9. GDPR

The General Data Protection Regulation (GDPR) is a comprehensive data protection and privacy regulation enacted by the European Union (EU) that became enforceable on May 25, 2018.

It applies to all member states of the EU and the European Economic Area (EEA) and has global implications for businesses that process the personal data of EU Residents.

2.9.1. Nine key aspects of GDPR for Data Management

- **Territorial Scope Personal Data**: This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not.
- **Data Subject Rights:** The owner's right to access, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object, and the right not to be subject to a decision based solely on automated processing.

- **Lawful Basis For Processing:** The data subject has given consent to the processing of his or her data for one or more specific purposes. Processing is necessary for the performance of a contract to which the data subject is party or to take steps at the request of the data subject before entering into a contract.
- **Consent:** Where the processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to the processing of his or her data. If the data subject's consent is given in the context of a written declaration that also concerns other matters, the request for consent shall be presented in a manner that is distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language.
- **Data Breach Notification:** In the case of a personal data breach, the controller shall without undue delay and, where feasible, not later than 72 hours after having become aware of it, notify the personal data breach supervisory authority competent by Article 55, unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons.
- **Data Protection Officers (Dpos) :** The data protection officer shall be designated based on professional qualities and, in particular, expert knowledge of data protection law and practices and the ability to fulfill the tasks referred.
- **Privacy By Design And By Default:** You should design any system, service, product, and/or business practice to protect personal data automatically. With privacy built into the system, the individual does not have to take any steps to protect their data — their privacy remains intact without them having to do anything.
- **Penalties & Fines:** Among other things, intentional infringement, a failure to take measures to mitigate the damage that occurred or lack of collaboration with authorities can increase the penalties. For especially severe violations, listed in Art. 83(5) GDPR, the fine framework can be up to 20 million euros, or in the case of an undertaking, up to 4 % of their total global turnover of the preceding fiscal year, whichever is higher

3. Methodology

3.1. Research Design

The research design for this study focuses on exploring and validating the use of automation to ensure efficient compliance with GDPR through the generation of privacy policy captions in web and mobile applications. This design encompasses both qualitative and quantitative methodologies to provide a comprehensive understanding of the processes and technologies involved.

- **Qualitative Approach**: The qualitative aspect of the research involves an in-depth examination of current practices and challenges associated with GDPR compliance in the technology industry, specifically concerning privacy policies. This includes a thorough review of literature, case studies, and expert interviews to gather insights into the existing landscape of privacy policy creation and the common pitfalls organizations face. The qualitative data will help identify key areas where automation can significantly enhance compliance efforts and improve user transparency.
- **Quantitative Approach:** The quantitative component focuses on the technical implementation and evaluation of automated privacy policy caption generation. This involves using static code analysis and deep learning models to analyze application code and generate corresponding privacy policy captions. The study will employ a controlled experimental setup where various web and mobile applications are subjected to these automated processes. Metrics such as accuracy, consistency, and user comprehension of the generated captions will be measured and compared against manually created captions to assess the effectiveness of the automation.
- **Justification for Chosen Design:** The mixed-methods approach is justified as it provides a holistic view of the problem by combining the strengths of both qualitative and quantitative research. The qualitative data offers context and depth, highlighting the real-world challenges and opportunities, while the quantitative data provides empirical evidence of the effectiveness of the automated solutions. This comprehensive approach ensures that the research findings are robust, actionable, and grounded in both theory and practice.
- **Addressing Research Objectives**: The chosen design aligns with the research objectives by systematically investigating the role of automation in GDPR compliance. It explores the current challenges in privacy policy creation, evaluates the potential of static code analysis and deep learning models for automated caption generation, and validates the practicality and effectiveness of these automated solutions in real-world scenarios. Through this design, the research aims to provide clear guidelines and best practices for organizations seeking to enhance their GDPR compliance efforts through automation.

3.2. Data Collection Methods

The data collection methods for this research are designed to gather comprehensive and relevant data necessary to evaluate the effectiveness of automating privacy policy captions for GDPR compliance in web and mobile applications. The methods involve a combination of primary and secondary data sources, focusing on the use of static code analysis and deep learning models.

3.2.1. Primary Data Sources

Static Code Analysis

- Description: Static code analysis involves examining the source code of web and mobile applications without executing them. This method is used to extract information about the app's data handling practices, identifying how personal data is collected, processed, and stored.
- Tools and Software:
  - MobSF (Mobile Security Framework):** An automated tool that performs static and dynamic analysis of mobile applications, helping to identify security vulnerabilities and privacy issues.
  - QARK (Quick Android Review Kit):** A tool designed to scan Android applications for potential security vulnerabilities and provide information about the app's data handling practices.
  - AndroGuard: A tool that performs static analysis of Android applications to extract information about the app's code and data flows.
  - SonarQue: A code quality management tool that helps in identifying and fixing vulnerabilities, bugs, and code smells in software applications.
  - Jadx-gui: A tool used to decompile Android DEX files to Java source code, enabling detailed analysis of the application's code structure.

Deep Learning Models

- Description: Deep learning models, particularly neural machine translation (NMT) models, are used to generate privacy policy captions from the analyzed code. These models learn to translate code segments into human-readable privacy policy snippets.
- Tools and Software
  - Neural Machine Translation (NMT) Models: Encoder-decoder architectures used for sequence-to-sequence learning, applied to generate captions from code analysis results.
  - Hugging Face's Gardio: A deep learning framework that can be used to implement and fine-tune NMT models for generating privacy policycaptions.
- Secondary Data Sources
  - Literature Review: A comprehensive review of existing literature on GDPR compliance, privacy policies, and the use of automation and deep learning in software development. This includes academic papers, industry reports, and case studies that provide insights into current practices and challenges.
  - Case Studies: Examination of real-world examples of GDPR non-compliance penalties and successful implementations of automated privacy solutions. These case studies provide context and validation for the research findings.
- Data Collection Process
  - Static Code Analysis
    - Collect source code samples from a diverse set of web and mobile applications.
    - Use the aforementioned tools to perform static analysis, extracting detailed information about data handling practices.
    - Identify and document instances of personal data collection, processing, and storage.
  - Deep Learning Model Training
    - Prepare training datasets consisting of code segments and their corresponding privacy policy captions.
    - Train NMT models using these datasets, fine-tuning them to improve the accuracy and relevance of the generated captions.
    - Apply the trained models to the analyzed code samples to generate privacy policy captions.
  - Evaluation and Validation
    - Compare the generated captions with manually created captions to assess accuracy, consistency, and user comprehension.
    - Conduct user studies to gather feedback on the clarity and usability of the generated captions.
    - Validate the results against GDPR compliance requirements to ensure they meet regulatory standards.

By employing these data collection methods, the research aims to gather robust and comprehensive data necessary to evaluate the feasibility and effectiveness of automating privacy policy captions for GDPR compliance in web and mobile applications.

3.3. Data Analysis Techniques

The data analysis techniques employed in this research aim to evaluate the effectiveness of automating privacy policy captions through static code analysis and deep learning models. The analysis focuses on assessing the accuracy, consistency, and compliance of the generated captions with GDPR requirements.

3.3.1. Static Code Analysis

- Objective: To extract detailed information about an application's data handling practices from its source code.
- Process:
  - Code Disassembly: Using tools like MobSF, QARK, AndroGuard, SonarQue, and Jadx-gui to disassemble DEX/ODEX/APK files and obtain an abstract syntax tree (AST) representation of the source code.
  - Call Graph Analysis: Constructing inter-procedure call flow graphs (ICFG) to understand the relationships and data flows between different parts of the application.
  - Data Flow Analysis: Identifying instances of personal data collection, processing, and storage by analyzing the flow of data within the application. This includes tracing data from its source (e.g., user input) to its various sinks (e.g., storage, network transmission).

3.3.2. Deep Learning Model Application

- Objective: To generate privacy policy captions from the analyzed code segments.
- Process:
  - Model Training: Training neural machine translation (NMT) models using a dataset of code segments paired with corresponding privacy policy captions. The models learn to translate technical code into human-readable privacy statements.
  - Caption Generation: Applying the trained NMT models to the code analysis results to generate privacy policy captions. The models produce captions that summarize the data handling practices identified in the code.
  - Post-Processing: Refining the generated captions to ensure they are clear, concise, and compliant with GDPR requirements.

3.3.3. Accuracy and Consistency Evaluation

- Objective: To assess the accuracy and consistency of the generated privacy policy captions.
- Process
  - Manual Comparison: Comparing the generated captions with manually created privacy policy captions to evaluate their accuracy. This involves checking for the presence of key information and ensuring that the captions correctly reflect the data handling practices.
  - Consistency Check: Ensuring that the generated captions are consistent across different parts of the application. This includes verifying that similar data handling practices are described in a similar manner in different captions.

3.3.4. User Comprehension and Usability Testing

- Objective: To evaluate the clarity and usability of the generated privacy policy captions from the perspective of end-users.
- Process
  - User Surveys and Interviews: Conducting surveys and interviews with a sample of users to gather feedback on the generated captions. Users are asked to rate the clarity, comprehensibility, and usefulness of the captions.
  - Usability Testing: Observing users as they interact with the generated captions within the application interface. This helps identify any areas where the captions may be unclear or difficult to understand.

3.3.5. Compliance Validation

- Objective: To ensure that the generated privacy policy captions meet GDPR compliance requirements.
- Process

- Regulatory Checklist: Using a checklist based on GDPR provisions to verify that the captions include all mandatory elements, such as the identity of the data controller, purposes of processing, data subject rights, and data retention periods.
- Expert Review: Having legal and data protection experts review the generated captions to ensure they comply with GDPR requirements and provide accurate and complete information to users.
- By employing these data analysis techniques, the research aims to thoroughly evaluate the feasibility and effectiveness of automating privacy policy captions, ensuring they are accurate, consistent, user-friendly, and compliant with GDPR requirements.

3.4. Implementation Framework

The implementation framework outlines the steps and processes involved in developing and integrating automated privacy policy captions into web and mobile applications to ensure efficient compliance with GDPR. This framework combines static code analysis and deep learning models to generate accurate and user-friendly privacy policy captions.

3.4.1. Initial Setup and Configuration

- Tool Selection: Identify and configure the tools required for static code analysis and deep learning model implementation. This includes MobSF, QARK, AndroGuard, SonarQue, Jadx-gui for static analysis, and neural machine translation (NMT) models such as those available on Hugging Face's Gardio.
- Environment Setup: Set up the development environment with the necessary software, libraries, and dependencies. Ensure that the environment supports the integration of static analysis tools and deep learning models.

3.4.2. Data Collection and Preparation

- Source Code Collection: Collect source code samples from a diverse set of web and mobile applications, focusing on applications that handle personal data.
- Data Annotation: Annotate the source code with information about data handling practices, identifying points where personal data is collected, processed, and stored. This annotated data will be used to train and validate the deep learning models.

3.4.3. Static Code Analysis

- Code Disassembly: Use tools like Jadx-gui to disassemble DEX/ODEX/APK files and obtain an abstract syntax tree (AST) representation of the source code.
- Call Graph and Data Flow Analysis: Construct inter-procedure call flow graphs (ICFG) to map the relationships and data flows between different parts of the application. Identify instances of personal data collection, processing, and storage.
- Information Extraction: Extract relevant information about data handling practices from the analyzed code, including the types of personal data collected, purposes of processing, and data sharing practices.

3.4.4. Deep Learning Model Training

- Dataset Preparation: Prepare a training dataset consisting of code segments paired with their corresponding privacy policy captions. This dataset should include diverse examples to ensure the model can generalize well.
- Model Training: Train the NMT models using the prepared dataset. Fine-tune the models to improve the accuracy and relevance of the generated captions.
- Model Validation: Validate the trained models using a separate validation dataset. Assess the models' performance in terms of accuracy, consistency, and comprehensibility of the generated captions.

3.4.5. Caption Generation

- Integration with Static Analysis Output: Integrate the output of the static code analysis with the trained deep learning models. Use the models to generate privacy policy captions based on the analyzed code segments.
- Post-Processing: Refine the generated captions to ensure clarity, conciseness, and compliance with GDPR requirements. This may involve adjusting the wording, formatting, and structure of the captions.

3.4.6. Implementation into Application Interfaces

- User Interface Integration: Integrate the generated privacy policy captions into the user interface of web and mobile applications. Ensure that the captions are presented at relevant points of user interaction, such as during account creation or when requesting permissions for data access.

- Real-Time Updates: Implement mechanisms for real-time updates to the privacy policy captions. This ensures that any changes in data handling practices are promptly reflected in the captions, maintaining ongoing compliance with GDPR.

3.4.7. Evaluation and Iteration

- Accuracy and Consistency Evaluation: Continuously evaluate the accuracy and consistency of the generated captions by comparing them with manually created captions and assessing user feedback.
- User Feedback and Usability Testing: Conduct user surveys, interviews, and usability testing to gather feedback on the clarity and usability of the generated captions. Use this feedback to iteratively improve the caption generation process.
- Compliance Monitoring: Regularly review the generated captions for GDPR compliance, using a regulatory checklist and expert reviews to ensure ongoing adherence to legal requirements.

3.4.8. Documentation and Best Practices

- Comprehensive Documentation: Document the entire implementation process, including tool configurations, data preparation steps, model training procedures, and integration workflows.
- Best Practices: Develop best practice guidelines for automating privacy policy captions, based on the findings and insights gained from the implementation and evaluation process. Share these guidelines with the broader community to promote the adoption of effective automation techniques for GDPR compliance.
- This implementation framework provides a structured approach to developing and integrating automated privacy policy captions, leveraging static code analysis and deep learning models to enhance GDPR compliance in web and mobile applications.

3.5. Validation and Testing

Validation and testing are critical components of this research, ensuring that the automated privacy policy captions generated through static code analysis and deep learning models are accurate, compliant, and user-friendly. This section outlines the methods and procedures used to validate and test the effectiveness of the automated captions.

3.5.1. Accuracy and Consistency Validation

- Manual Comparison: Compare the generated privacy policy captions with manually created captions. This involves checking for the presence of key information, such as the types of personal data collected, the purposes of data processing, data sharing practices, and user rights.
- Procedure: Select a representative sample of web and mobile applications. For each application, generate privacy policy captions using the automated system and manually create captions by experts. Evaluate the accuracy by comparing these sets of captions.
- Metrics: Measure accuracy in terms of completeness (coverage of all required elements), correctness (factual accuracy), and relevance (appropriateness of the information provided).

3.5.2. User Comprehension and Usability Testing

- User Surveys and Interviews: Conduct surveys and interviews with a diverse group of users to gather feedback on the clarity, comprehensibility, and usability of the generated captions.
- Procedure: Present users with both automated and manually created captions. Use structured questionnaires and interview guides to elicit feedback on their understanding, ease of use, and trust in the information provided.
- Metrics: Measure user comprehension through questions about the key points covered in the captions. Assess usability based on user ratings of clarity, length, and readability.
- Usability Testing: Observe users interacting with the generated privacy policy captions within the application interface. Identify any areas where the captions may be unclear, difficult to understand, or inconveniently placed.
- Procedure: Conduct usability testing sessions where users perform typical tasks that involve interacting with privacy policy captions, such as signing up for an account or granting permissions. Record and analyze their interactions and feedback.
- Metrics: Measure usability through task completion rates, time taken to understand the captions, and user satisfaction scores.

3.5.3. Compliance Validation

- Regulatory Checklist: Use a checklist based on GDPR provisions to verify that the generated captions include all mandatory elements. Ensure that captions address the identity of the data controller, purposes of processing, types of data collected, data retention periods, and user rights.
- Procedure: Develop a comprehensive checklist covering all GDPR requirements related to privacy policies. Apply this checklist to the generated captions to identify any omissions or non-compliance issues.
- Metrics: Measure compliance by counting the number of checklist items successfully addressed in each caption.
- Expert Review: Have legal and data protection experts review the generated captions to ensure they comply with GDPR requirements and provide accurate and complete information to users.
- Procedure: Select a panel of experts with experience in GDPR compliance and data protection. Provide them with the generated captions and the corresponding application contexts. Gather their feedback and suggestions for improvement.
- Metrics: Measure the level of compliance based on expert ratings and qualitative feedback on the accuracy, completeness, and legal sufficiency of the captions.

3.5.4. Real-Time Monitoring and Continuous Improvement

- Real-Time Updates: Implement mechanisms for real-time updates to the privacy policy captions to reflect any changes in data handling practices. This ensures ongoing compliance with GDPR requirements.
- Procedure: Develop a monitoring system that tracks changes in application code and automatically updates the corresponding privacy policy captions. Regularly review and validate these updates.
- Metrics: Measure the timeliness and accuracy of updates in response to changes in data handling practices.
- Continuous Improvement: Use the findings from validation and testing to iteratively improve the automated caption generation process. Address any identified issues and enhance the accuracy, clarity, and compliance of the captions.
- Procedure: Establish a feedback loop where validation and testing results inform model retraining, tool refinement, and process adjustments. Conduct periodic reviews to assess progress and implement improvements.
- Metrics: Measure improvement by tracking key performance indicators over time, such as accuracy rates, user satisfaction scores, and compliance levels.

By employing these validation and testing methods, the research ensures that the automated privacy policy captions are not only accurate and compliant but also user-friendly and effective in conveying essential information to users. This rigorous approach helps build trust and confidence in the automated solutions, promoting wider adoption and enhancing GDPR compliance in web and mobile applications.

3.6. Ethical Considerations

Ethical considerations are paramount in research involving personal data, particularly in the context of GDPR compliance and the automation of privacy policy captions. This section outlines the ethical principles and practices adhered to throughout the research to ensure the protection of personal data, respect for user rights, and compliance with legal and ethical standards.

3.6.1. Data Privacy and Security

- Confidentiality: Ensure that all personal data collected during the research is treated with the highest level of confidentiality. Access to data is restricted to authorized personnel only, and data is anonymized where possible to protect the identities of individuals.
- Procedure: Implement robust data security measures, including encryption, access controls, and anonymization techniques, to safeguard personal data throughout the research process.
- Data Minimization: Collect only the data that is necessary for the purposes of the research. Avoid collecting excessive or irrelevant data that does not contribute to the research objectives.
- Procedure: Review data collection methods and datasets to ensure that only essential data is included. Regularly audit data collection practices to maintain compliance with the principle of data minimization.

3.6.2. Informed Consent

- User Consent: Obtain informed consent from all participants involved in the research, particularly those whose data is being used for testing and validation of the automated privacy policy captions.

- Procedure: Provide clear and comprehensive information to participants about the nature of the research, the data being collected, how it will be used, and their rights. Obtain explicit consent before collecting any personal data.
- Withdrawal of Consent: Allow participants to withdraw their consent at any time without any negative consequences. Ensure that their data is promptly removed from the study if they choose to withdraw.
- Procedure: Implement a straightforward process for participants to withdraw their consent and ensure that their data is deleted or anonymized immediately upon request.

3.6.3. Transparency and Accountability

- Transparency: Maintain transparency about the research objectives, methods, and findings. Share relevant information with stakeholders and participants in a clear and accessible manner.
- Procedure: Regularly update stakeholders and participants on the progress of the research, providing them with summaries of findings and any changes to the research plan.
- Accountability: Ensure accountability in data handling practices by establishing clear roles and responsibilities for data protection. Appoint a data protection officer (DPO) or equivalent to oversee compliance with ethical standards and legal requirements.
- Procedure: Define and document the responsibilities of all team members involved in data handling. Conduct regular training sessions on data protection and ethical practices.

3.6.4. Compliance with Legal and Ethical Standards

- GDPR Compliance: Adhere to all GDPR requirements, including data subject rights, lawful basis for processing, and data breach notifications. Ensure that the research methodology aligns with GDPR principles of data protection by design and by default.
- Procedure: Conduct regular compliance audits to verify adherence to GDPR requirements. Implement technical and organizational measures to integrate data protection into the research process from the outset.
- Ethical Review: Subject the research proposal to an ethical review by an independent ethics committee or institutional review board (IRB). Obtain ethical approval before commencing the research.
- Procedure: Prepare and submit a detailed research proposal outlining the ethical considerations and data protection measures. Address any feedback or concerns raised by the ethics committee.

3.6.5. Addressing Potential Risks

- Risk Assessment: Identify and assess potential risks to data privacy and security throughout the research. Develop mitigation strategies to address these risks proactively.
- Procedure: Conduct a thorough risk assessment at the beginning of the research and periodically review and update it. Implement risk mitigation measures, such as enhanced security protocols and contingency plans.
- Impact on Participants: Consider the potential impact of the research on participants, particularly in terms of their privacy and data protection rights. Take steps to minimize any negative impact.
- Procedure: Engage with participants to understand their concerns and preferences regarding data use. Ensure that the research is designed to minimize any adverse effects on their privacy and rights.

By adhering to these ethical considerations, the research ensures the protection of personal data, respects user rights, and maintains compliance with legal and ethical standards. This ethical framework not only safeguards the integrity of the research but also builds trust and confidence among participants and stakeholders, promoting responsible and ethical data handling practices.

3.7. Limitations of the Study

While this research aims to provide a comprehensive evaluation of automating privacy policy captions for GDPR compliance in web and mobile applications, several limitations must be acknowledged. Understanding these limitations helps contextualize the findings and highlights areas for future research and improvement.

3.7.1. Data and Tool Limitations

- Diverse Application Contexts: The study focuses on a selected set of web and mobile applications. While these applications are chosen to represent a variety of data handling practices, they may not capture the full diversity of applications in the market.
- Impact: Results may not be fully generalizable to all types of web and mobile applications, especially those with unique or highly specialized data processing practices.

- Future Research: Expanding the study to include a broader range of applications from different industries and regions would help improve generalizability.
- Tool Limitations: The static analysis tools and deep learning models used in this research have their own limitations in terms of accuracy and capabilities.
- Impact: The tools may not capture all nuances of data handling practices, leading to potential inaccuracies in the generated captions.
- Future Research: Continual improvement and validation of these tools, as well as exploring additional tools and methodologies, could enhance the accuracy and reliability of the results.

3.7.2. Model Training and Dataset

- Training Data Quality: The quality of the deep learning models' output heavily depends on the quality and diversity of the training data.
- Impact: Inadequate or biased training data can result in less accurate or less relevant privacy policy captions.
- Future Research: Building a more comprehensive and diverse training dataset, including more varied examples of data handling practices, can help improve model performance.
- Model Limitations: The neural machine translation (NMT) models used in this study are based on existing datasets and may not fully adapt to new or unseen data processing scenarios.
- Impact: The models might struggle with generating accurate captions for novel or complex data handling practices not well-represented in the training data.
- Future Research: Developing more advanced models, such as those incorporating reinforcement learning or transfer learning, could address these limitations.

3.7.3. Validation and User Testing:

- Sample Size: The sample size for user testing and validation is limited due to resource constraints.
- Impact: A smaller sample size may not provide a comprehensive understanding of user comprehension and usability across diverse user demographics.
- Future Research: Conducting larger-scale user studies with more diverse participant groups would offer more robust insights into user comprehension and usability.
- User Bias: Participants in user studies may have varying levels of familiarity with privacy policies and data protection, potentially influencing their feedback.
- Impact: Results may be biased towards users with more knowledge or interest in data privacy, skewing usability and comprehension findings.
- Future Research: Ensuring a balanced mix of participants with varying levels of familiarity with privacy policies can help mitigate this bias.

3.7.4. Real-Time Adaptability

- Dynamic Environments: Web and mobile applications frequently undergo updates and changes, which may not be immediately reflected in the automated privacy policy captions.
- Impact: Captions generated may become outdated if not updated promptly to reflect changes in data handling practices.
- Future Research: Implementing real-time monitoring and adaptive systems that can quickly update captions in response to application changes would enhance the relevance and accuracy of the captions.

3.7.5. Ethical and Legal Considerations

- Evolving Regulations: GDPR and other data protection regulations continue to evolve, and interpretations of these regulations may vary.
- Impact: The study's findings and methodologies may need continual updates to remain compliant with the latest regulatory requirements.
- Future Research: Regularly reviewing and updating the research framework to align with evolving data protection laws and best practices is essential.

3.7.6. Resource Constraints

- Time and Budget: The research is conducted within certain time and budget constraints, which may limit the depth and breadth of the analysis.
- Impact: Constraints may affect the comprehensiveness of tool evaluations, model training, and user testing.
- Future Research: Securing additional funding and resources could allow for a more extensive and detailed study.

By acknowledging these limitations, this research provides a transparent account of its scope and potential constraints. Addressing these limitations in future studies will be crucial for advancing the field and enhancing the reliability and applicability of automated privacy policy captions for GDPR compliance.

4. Result and discussion

4.1. Software implementation that deals with PII

Call Graph: This represents calling relationships between subroutines in a computer program

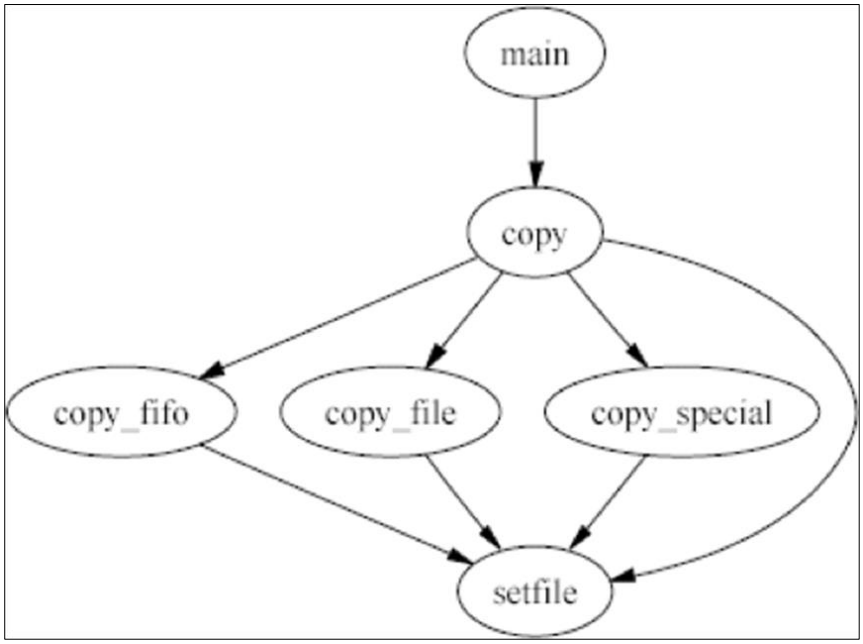


Figure 1 Function Call Hierarchy for the Copy Module

Figure 1 represents the function call hierarchy for the copy module. The main function calls the copy function, which then calls three specialized functions: `copy_fifo`, `copy_file`, and `copy_special`. All three of these specialized functions, in turn, call the `setfile` function. This hierarchical structure shows how different types of file copy operations are managed within the module, with `setfile` being a common function used by the specialized copying functions.

- An **abstract syntax tree** (**AST**) is a data structure used in computer science to represent the structure of a program or code snippet. It is a tree representation of the abstract syntactic structure of text (often source code) written in a formal language. Each node of the tree denotes a construct occurring in the text. It is sometimes called just a **syntax tree**.

Example of code and AST using Preview extension from Vscode

```
import ast
import inspect
def example_function(a, b):
 result = a + b
  return result
# Get the source code of the function
source_code = inspect.getsource(example_function)
# Parse the source code into an abstract syntax tree (AST)
parsed_ast = ast.parse(source_code)
# Display the AST
print(ast.dump(parsed_ast))
```

Figure 2 Abstract Syntax Tree Representation of Python Import Statements

Figure 2 shows a JSON representation of an Abstract Syntax Tree (AST) for a Python script. The AST describes two import statements within a module. The first import statement imports the `ast` module, and the second imports the `inspect` module. Each import node includes metadata such as the type of AST node (`Import`), the column and line offsets (`col_offset`, `end_col_offset`, `lineno`, `end_lineno`), and details of the imported names, which are represented as `alias` nodes. This structure helps in understanding the syntactic elements of the Python script.

4.2. Framework for Automating Privacy Policy Captions

Multiple authors research have approached theses using two main strategy

- Deep learning
- Static code analysis <>

4.3. Static Code Analysis

This is used to analyze computer programs performed without executing them, in contrast with dynamic program analysis, which is performed on programs during their execution.

What they do is disassemble DEX/ODEX/APK files, analyze the code, and extract information about the app.

4.4. Examples of Apps that can perform static analysis out there

- MobSF (Mobile Security Framework),
- QARK (Quick Android Review Kit),
- AndroGuard.
- SonarQue
- Jadx-gui (https://github.com/skylot/jadx)

Andriod API Documentation: Describe the nature of the sensitive information accessed which can be analyzed for in the static analysis.

Class code — Caption from the static analysis tells what the code does for the collection of PII d'ata collected from Andriod Developer documentation. This was used as a basis for search in the APK files static analysis

List of a few Android APIs and classes related to handling sensitive (PII) from the Android documentation Link

- Uses Permission
- LocationManager
- BiometricPrompt
- TelephonyManager
- Content-providers

4.5. Deep learning

Different Authors have used the neural machine translation (NMT) model, NMT is an encoder-decoder Architecture, in some cases known as the Sequence to Sequence (Seq2Seq) model :

4.5.1. Image Captioning

It receives the image as the input and outputs a sequence of words. This also works with videos.

4.5.2. Sentiment Analysis

These models understand the meaning and emotions of the input sentence and output a sentiment score. It is usually rated between -1 (negative) and 1 (positive) where 0 is neutral. client's emotions analyzer.

4.5.3. Translation

This model reads an input sentence, understands the message and the concepts, then translates it into a second language.

Another deep learning tool that can be used is

Workflow with Gardio on Hugging Face: *YouTubeLink*

Application Architect

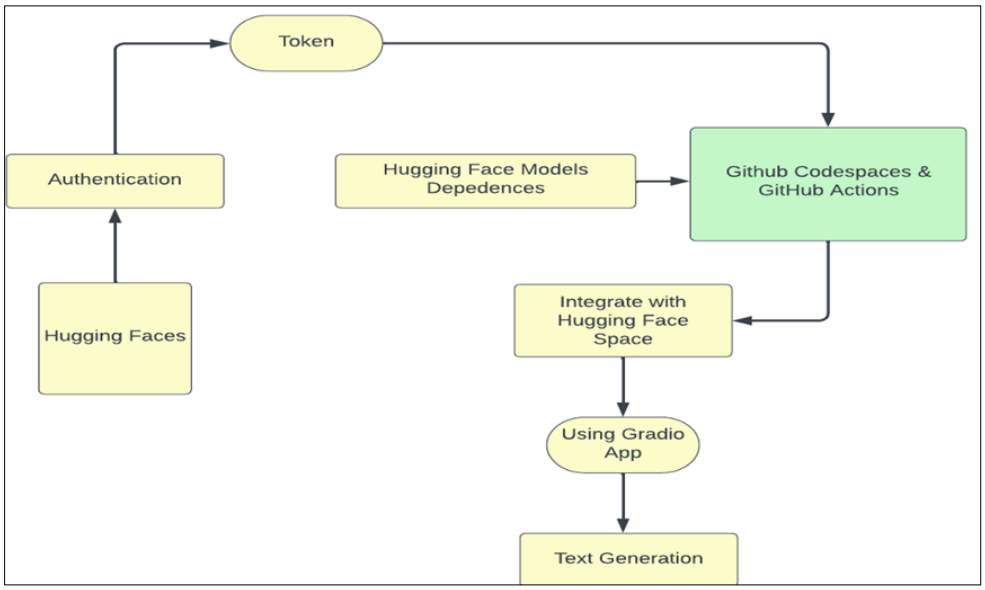


Figure 3 Workflow for Text Generation Using Hugging Face, GitHub Codespaces, and Gradio

Figure 3 illustrates the workflow for generating text using Hugging Face, GitHub Codespaces, and Gradio. It begins with obtaining a token from Hugging Face, which is used for authentication. Once authenticated, the Hugging Face models and dependencies are set up. These are integrated into GitHub Codespaces and GitHub Actions to streamline the development process. The models and dependencies are then integrated with Hugging Face Space, which is utilized by the Gradio app for text generation. The final step in the process is the generation of text using the Gradio application.

4.6. Application Architecture of static analysis and deep learning to generate a privacy policy captions

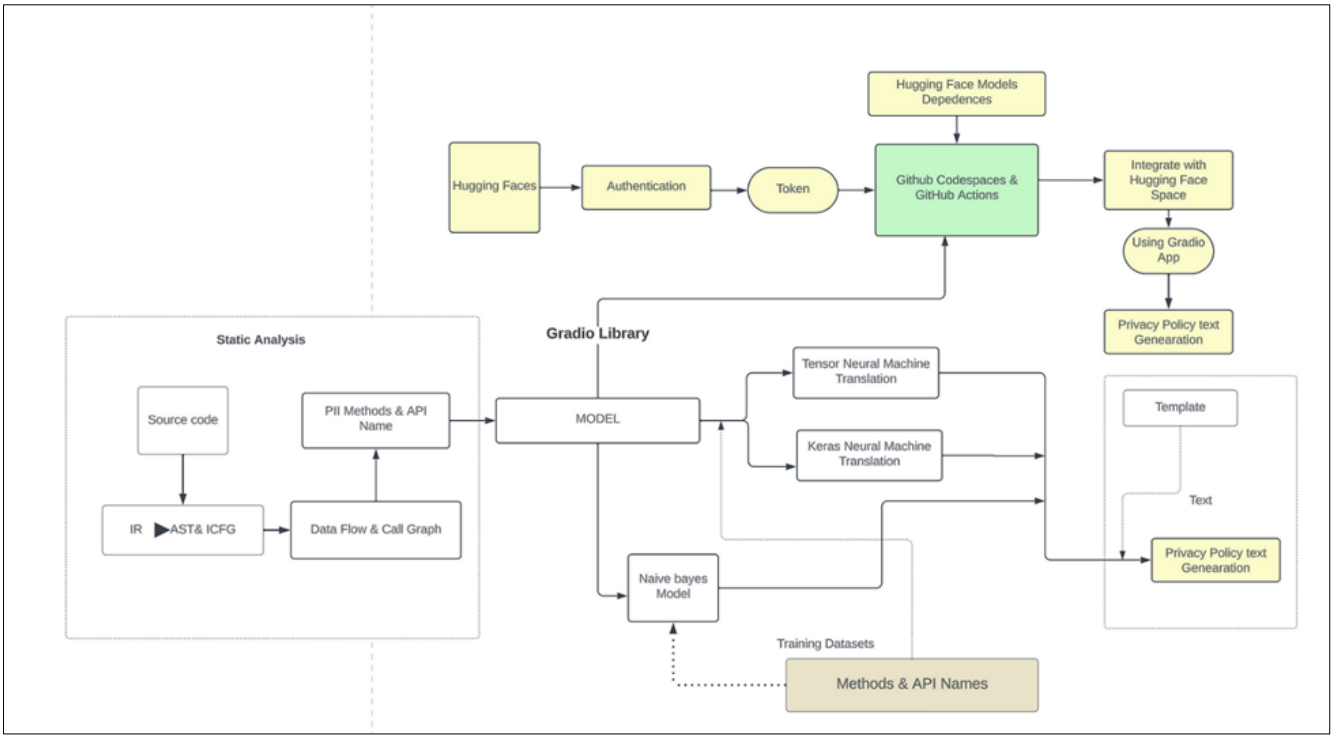


Figure 4 Integrated Workflow for Privacy Policy Text Generation Using Static Analysis, Machine Learning Models, and Gradio Library

Figure 4 illustrates a comprehensive workflow for generating privacy policy text by integrating static analysis, machine learning models, and the Gradio library. Initially, the source code undergoes intermediate representation (IR) through

Abstract Syntax Tree (AST) and Interprocedural Control Flow Graph (ICFG) analysis, which helps in identifying Personally Identifiable Information (PII) methods and API names via data flow and call graphs. These identified elements are then fed into machine learning models, including Tensor Neural Machine Translation, Keras Neural Machine Translation, and a Naive Bayes Model, all housed within the Gradio library and trained with relevant datasets. Authentication and model dependencies are managed through Hugging Face, where a token is generated, and dependencies are handled via GitHub Codespaces and GitHub Actions. The trained models, integrated with Hugging Face Space, are utilized by the Gradio app to generate privacy policy text. This streamlined workflow ensures efficient and accurate automation of privacy policy text generation by leveraging advanced static analysis and machine learning techniques.

4.7. Output of the data from Static Analysis passed to ML model

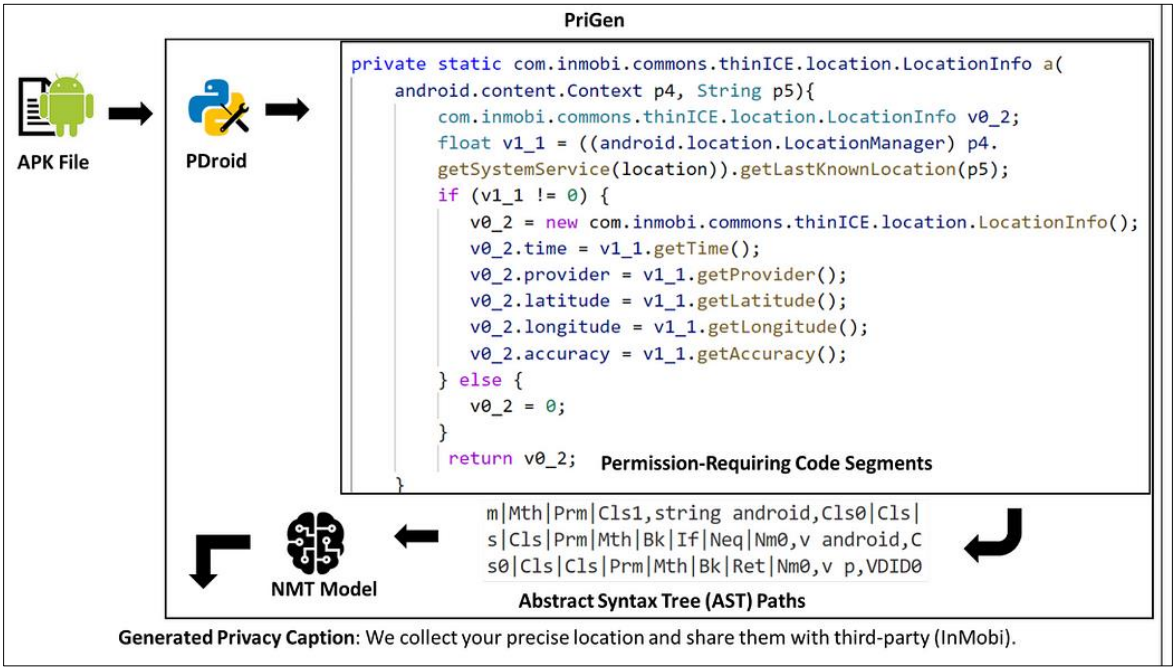


Figure 5 PriGen: Automated Translation of Android Code to Privacy Captions

Figure 5 illustrates the workflow of PriGen, a tool designed to automate the translation of Android application code into privacy captions. The process starts with an APK file, which is analyzed using PDroid to extract permission-requiring code segments. These code segments are converted into Abstract Syntax Tree (AST) paths, which capture the structural and contextual information of the code. The AST paths are then fed into a Neural Machine Translation (NMT) model. This model processes the AST paths and generates a privacy caption. In the example provided, the generated privacy caption explains that the application collects precise location data and shares it with a third party (InMobi). This automated process aids in understanding and documenting the privacy implications of mobile applications.

4.8. Recommendations

Based on the findings of this research, several recommendations can be made to enhance the effectiveness of automating privacy policy captions for GDPR compliance in web and mobile applications.

4.8.1. Improve Training Data Quality and Diversity:

- Recommendation: Develop a more comprehensive and diverse dataset for training deep learning models. This should include varied examples of data handling practices across different types of applications and industries.
- Implementation: Collaborate with data protection authorities and industry partners to gather a wide range of annotated data samples. Regularly update the dataset to reflect new privacy practices and regulatory changes.

4.8.2. Enhance Tool Capabilities:

- Recommendation: Invest in the continuous improvement of static analysis tools and deep learning models to better capture the nuances of data handling practices.

- Implementation: Allocate resources for research and development to refine existing tools and explore new methodologies. Incorporate feedback from developers and users to address tool limitations.

4.8.3. Expand User Testing and Validation:

- Recommendation: Conduct larger-scale user studies with diverse participant groups to gather more robust insights into user comprehension and usability of automated privacy policy captions.
- Implementation: Partner with academic institutions, industry organizations, and user advocacy groups to recruit a diverse set of participants. Use a combination of surveys, interviews, and usability testing sessions to collect comprehensive feedback.

4.8.4. Implement Real-Time Monitoring and Updates:

- Recommendation: Develop real-time monitoring systems that can a Automatically update privacy policy captions in response to changes in application data handling practices.
- Implementation: Integrate monitoring tools that track changes in application code and data flows. Ensure that the system can generate and deploy updated captions promptly to maintain compliance.

4.8.5. Regularly Review and Update Compliance Strategies:

- Recommendation: Continuously review and update compliance strategies to align with evolving GDPR requirements and best practices.
- Implementation: Establish a dedicated team or task force to monitor regulatory developments and update compliance frameworks accordingly. Conduct regular audits and compliance checks to ensure ongoing adherence to GDPR.

4.8.6. Address Ethical and Legal Considerations:

- Recommendation: Prioritize ethical considerations in the development and deployment of automated privacy policy captions. Ensure that all data processing activities comply with legal standards and respect user rights.
- Implementation: Conduct thorough ethical reviews of all research and development activities. Implement robust data protection measures and obtain informed consent from users before collecting or processing their data.

5. Conclusion

This research demonstrates the potential of automating privacy policy captions to enhance GDPR compliance in web and mobile applications. By leveraging static code analysis and deep learning models, organizations can generate accurate, consistent, and user-friendly privacy policy captions that meet regulatory requirements and improve user transparency. The mixed-methods approach employed in this study, combining qualitative insights and quantitative evaluations, provides a comprehensive understanding of the challenges and opportunities associated with automating privacy policy captions. The findings highlight the importance of accurate data handling practices, user comprehension, and continuous compliance monitoring. However, the research also identifies several limitations, including the diversity of application contexts, tool capabilities, and the need for more extensive user testing. Addressing these limitations through the recommended actions will further enhance the effectiveness and reliability of automated privacy policy captions. Overall, this research contributes valuable insights and practical guidelines for organizations seeking to improve their GDPR compliance efforts through automation. By implementing the recommendations, organizations can build trust with users, ensure regulatory compliance, and stay ahead in the evolving landscape of data protection. The adoption of automated solutions for privacy policy generation represents a significant step forward in achieving efficient and effective data protection practices in the digital age.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Albrecht, J. P. (2016). How the GDPR Will Change the World. European Data Protection Law Review, 2(3), 287-289.

[2] Jain, V., Gupta, S. D., Ghanavati, S., & Peddinti, S. T. (n.d.). PriGen: Towards Automated Translation of Android Applications' Code to Privacy Captions.

[3] PriGen: Towards Automated Translation of Android Applications' Code to Privacy Captions by Vijayanta Jain, Sanonda Datta Gupta,Sepideh Ghanavati & Sai Teja Peddinti

[4] Tankard, C. (2016). What the GDPR means for businesses. Network Security, 2016(6), 5-8.

[5] Tikkinen-Piri, C., Rohunen, A., & Markkula, J. (2018). EU General Data Protection Regulation: Changes and implications for personal data collecting companies. Computer Law & Security Review, 34(1), 134-153.

[6] Voigt, P., & von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer.

[7] White & Case. (2018). The Impact of the GDPR One Year On. Retrieved from https://www.whitecase.com.