www.americanosp.com/index.php/AJBSP ISSN online: 3064-7568

American Journal of Business Science Philosophy

Strengthening Artificial Intelligence Governance through Ethical Handling of Sensitive Data: An Applied Study on Text Classification and Differential Privacy

Ziad Abdullah Alotaibi[1*] and Ziyad Ibraheem AlZaidan[2]

[1]*College of Engineering, Qassim University, Buraydah, Saudi Arabia. Email: contact@ziadabdullah.com*
[2]*Onaizah Colleges, Onaizah, Saudi Arabia. Email: ziyadxp@gmail.com*
**Corresponding author: contact@ziadabdullah.com*

Paper type: Article

Received: 05 June 2025
Revised: 23 July 2025
Accepted: 24 July 2025
Published: 29 July 2025

Citation: Alotaibi, Z. A., & AlZaidan, Z. I. (2025). Strengthening artificial intelligence governance through ethical handling of sensitive data: An applied study on text classification and differential privacy. American Journal of Business Science Philosophy, 2(2), 298–314. https://doi.org/10.70122/ajbsp.v2i2.42

Abstract

This research develops a comprehensive hybrid framework to enhance Artificial Intelligence governance by ethically managing sensitive textual data through advanced classification techniques. Focusing on natural language processing (NLP) applications, the study integrates rule-based systems, logistic regression, and transformer-based models, notably BERT, to address the challenges of identifying and handling sensitive information within complex and ambiguous linguistic contexts. Experimental results demonstrate that the hybrid model attains an overall classification accuracy of 91%, with precision and recall scores of 89% and 94%, respectively, achieving an F1-score of 92%. These metrics reflect the model's robustness in real-world scenarios where explicit textual indicators are often lacking. Individually, the rule-based approach excels in precision (98.6%) for clearly identifiable sensitive content, logistic regression ensures perfect recall (100%), capturing all sensitive instances albeit with increased false positives, and the BERT model achieves perfect precision, effectively minimizing false alarms. The hybrid approach synergizes these strengths, resulting in a balanced and reliable classification system. The study further explores the integration of differential privacy via a differentially private logistic regression model using the diffprivlib library, assessing privacy-utility trade-offs at varying privacy budgets (ε = 3, 5, 6). Results reveal that stronger privacy guarantees (lower ε) reduce classification accuracy (78% at ε=3), while looser privacy constraints (ε=6) approach non-private model performance (97% accuracy). These findings underscore the potential of combining hybrid NLP models with differential privacy to deliver scalable, trustworthy, and privacy-preserving AI systems. The proposed framework holds significant relevance for sensitive domains such as healthcare, public administration, and corporate governance, where balancing data privacy and AI performance is critical. Future research should extend these findings by exploring additional privacy configurations and validating the approach against diverse real-world datasets to optimize the equilibrium between privacy protection and analytical effectiveness.

Keywords: artificial intelligence; governance; sensitive data; ethical system

1. Introduction

In the mid-20th century, Alan Turing, one of the pioneering figures in computer science and artificial intelligence, posed a question that has since shaped the trajectory of technological innovation: "Can machines think?" While initially a philosophical inquiry, this question ignited decades of research aimed at creating machines capable of mimicking human intelligence. Over the years, significant advances in computing power, algorithms, and data availability have transformed this abstract hypothesis into a tangible reality. Today, artificial intelligence is deeply embedded in various aspects of society, with natural language processing (NLP) standing out as one of its most impactful domains. NLP enables machines to understand, interpret, and generate human language, facilitating applications such as virtual assistants, machine translation, sentiment

analysis, and automated content moderation, bringing human-computer interaction closer to natural communication.

The rapid expansion and adoption of AI-driven language models have brought about unprecedented opportunities but also profound challenges, particularly regarding privacy and data protection. These models often rely on large datasets that contain personal and sensitive information, raising concerns about how such data is collected, processed, and safeguarded. Without proper governance, AI systems risk violating privacy rights, perpetuating biases, and making decisions that lack transparency and accountability. Consequently, the question of how to regulate and govern AI technologies, ensuring they operate ethically and responsibly, has become increasingly urgent. AI governance encompasses a set of principles, policies, and technical measures designed to guide the ethical development and deployment of AI systems. Among the many tools available, differential privacy has emerged as a promising technique to protect individual data confidentiality while allowing AI models to learn from aggregated data. Differential privacy introduces controlled noise to datasets or model parameters, preventing the identification of specific individuals even when models are exposed to adversarial attacks. Despite its promise, integrating differential privacy into natural language models presents a critical dilemma: the noise added to protect privacy can reduce the accuracy and reliability of the models, potentially limiting their effectiveness in real-world applications.

This research is situated at the intersection of AI governance, privacy protection, and natural language processing. It aims to rigorously analyze the impact of applying differential privacy on the performance of NLP models, specifically in text classification tasks. By employing a hybrid multi-layer approach that combines linguistic rules, logistic regression based on TF-IDF representations, and transformer-based models like DistilBERT, this study explores how privacy-preserving techniques influence the ability of models to process and classify texts accurately. Beyond measuring performance metrics such as accuracy, precision, recall, and F1-score, the research also investigates the models' adherence to governance principles—examining their capacity to handle sensitive information responsibly and make ethical decisions when faced with uncertainty or ambiguous inputs.

The overarching goal of this study is to strike a balance between two often competing objectives: safeguarding individual privacy and maintaining high standards of AI performance. Achieving this balance is essential for building AI systems that users and regulators can trust, particularly as AI becomes increasingly integrated into sensitive sectors like healthcare, finance, education, and public administration. By providing insights into how differential privacy affects NLP models and proposing practical recommendations for governance frameworks, this research contributes to the ongoing effort to ensure that AI technologies are developed and deployed in ways that respect ethical norms, legal requirements, and societal values.

This study seeks to answer several key questions. First, it investigates what the fundamental principles of AI governance related to privacy protection are, and how these principles can be translated into technical applications such as sensitive data restrictions and differential privacy within natural language processing models. Second, it explores how differential privacy techniques affect the accuracy and efficiency of text classification models in NLP tasks. Third, the research examines the extent to which NLP models can comply with governance controls, particularly in scenarios where the models' confidence in classifying sensitive texts is low. Finally, it identifies the technical and ethical challenges involved in implementing governance and differential privacy controls in AI models and seeks ways to overcome these challenges to achieve an effective balance between performance and compliance. By addressing these questions, the research aims to provide a comprehensive understanding of how governance controls can be effectively integrated into NLP models, promoting AI systems that are both trustworthy and high-performing.

2. Literature Review

2.1. AI Governance and Its Ethical Principles

The discussion explores the theoretical foundations of AI governance, highlighting key concepts and principles that shape the interaction between artificial intelligence technologies and society. It begins by

examining the various types of AI and their impacts, followed by an overview of the evolving regulatory frameworks designed to ensure responsible AI deployment. Central to this exploration are the fundamental ethical principles of fairness, transparency, privacy, and accountability, which guide the development and use of AI systems. The analysis also addresses the practical challenges encountered when implementing governance in natural language processing models, particularly those related to explainability, bias, and regulatory compliance.

2.1.1. Concept of AI Governance and Its Regulatory Development

AI governance refers to the comprehensive framework that regulates the creation, deployment, and use of artificial intelligence systems to ensure they adhere to ethical, legal, and social standards. It encompasses policies, procedures, and controls designed to balance innovation with the protection of individuals, particularly in sensitive sectors such as healthcare, education, and the judiciary, where unregulated AI use could lead to biases, discrimination, or privacy violations.

At the international level, the Organization for Economic Co-operation and Development (OECD) introduced the first broad AI governance framework in 2019, outlining five key principles: inclusive growth, respect for human rights, transparency, accountability, and technical robustness. In Europe, the EU Artificial Intelligence Act (European Parliament, 2024) represents the first legally binding regulatory framework specifically for AI. It classifies AI applications according to risk levels—ranging from low to high—and mandates stringent controls for high-risk systems, including privacy safeguards, thorough documentation of data origins, and explainability requirements.

In the United States, the White House Office of Science and Technology Policy (OSTP, 2022) issued the "Blueprint for an AI Bill of Rights," emphasizing similar core values such as privacy protection, non-discrimination, and accountability. These regulatory developments mark a clear transition from voluntary ethical guidelines toward enforceable governance mechanisms. This evolving landscape provides a foundation for assessing how AI models, especially natural language models, respect and uphold users' digital rights in practice.

2.1.2. Artificial Intelligence, Its Types, and Their Impact on Privacy

Artificial Intelligence (AI) encompasses a range of technologies that enable machines to replicate human capabilities such as learning, analysis, and decision-making. AI is generally categorized into several types based on its complexity and scope. The most common type today is Narrow AI, which is designed for specific tasks like image recognition, natural language processing, or content recommendation. While Narrow AI offers numerous benefits, it poses privacy risks because it often relies on vast amounts of user data, including sensitive information that may not always be adequately protected.

A more recent advancement is Generative AI, which creates new content—such as text, images, or videos—by learning patterns from training data. Although it drives innovation and creativity, Generative AI raises complex privacy concerns. Since it can potentially reproduce or reveal sensitive personal data unintentionally, it challenges traditional data protection measures and confidentiality safeguards. At a more theoretical level, Artificial General Intelligence (AGI) represents AI systems capable of performing any intellectual task a human can do. Although still in the research phase, AGI carries significant privacy and security implications due to its ability to process and analyze enormous datasets autonomously, which could lead to unprecedented privacy violations.

The concept of Superintelligence refers to an AI surpassing human intelligence across all domains, currently a philosophical and scientific idea. This form raises serious concerns about loss of control over autonomous systems, with potential for extensive privacy breaches and human rights violations. In practical terms, Narrow and Generative AI currently pose the greatest privacy risks due to their widespread adoption in daily technologies like voice assistants and translation tools. These systems collect and process large volumes of personal data, often without robust protection, exposing users to risks such as unauthorized tracking, data

breaches, and unethical use. Generative AI's potential to inadvertently replicate sensitive information further complicates these challenges. Additionally, the increasing complexity of advanced AI models, such as those based on deep learning, makes interpreting decisions and tracking data flow difficult, heightening the risk of privacy violations. Moreover, AI models may perpetuate bias and discrimination when trained on unbalanced or opaque datasets, which not only undermines fairness but also threatens individual privacy and human rights. Consequently, it is vital to implement stringent governance frameworks alongside privacy-preserving techniques like differential privacy, encryption, and federated learning. These measures help strike a balance between leveraging AI's benefits and safeguarding personal data and privacy.

2.1.3. Ethical Principles in Designing Intelligent Models

In the development of intelligent models, adherence to ethical principles is essential to ensure that systems function fairly, transparently, and responsibly. One of the core principles is justice, which requires that AI systems remain unbiased across social, religious, gender, ethnic, or age dimensions. A fair system must not flag queries related to topics like "women" as sensitive simply due to the mention of gender, unless the content itself justifies such classification. This ensures respect for diversity and avoids perpetuating discrimination.

Another foundational principle is transparency, which refers to the need for clarity in how the system processes inputs and generates outputs. Users and developers should be able to understand or trace the logic behind a model's prediction or classification, even if the underlying architecture is complex. Closely related to this is explainability, which emphasizes the importance of enabling both experts and general users to grasp the reasoning behind a model's decision. This is particularly critical in high-stakes domains where misclassifications can have serious consequences.

Lastly, accountability ensures that a responsible party is identified for the outcomes produced by the model. This includes acknowledging and addressing any harm or unintended consequences resulting from the model's deployment. Ethical AI governance frameworks, such as those proposed by the European Commission, stress the importance of integrating these principles to foster trust, fairness, and reliability in AI-driven systems.

2.1.4. Challenges in Applying Governance to Natural Language Models

The practical implementation of governance principles in Natural Language Processing (NLP) models presents several complex challenges, particularly due to the sensitive and personal nature of the textual data these models handle. A primary issue is the lack of explainability, as deep learning models such as GPT and BERT operate through millions of parameters that function as opaque decision layers, making it difficult to trace or justify specific outcomes. This opacity contradicts the transparency requirements outlined in most governance frameworks.

Another major challenge is identifying and mitigating bias in training datasets, especially when such data is harvested automatically from the internet without proper documentation. In these cases, models may inadvertently adopt and reproduce discriminatory associations—for instance, linking particular names with specific nationalities or behaviors—which directly violates ethical standards centered on fairness and justice. The unintentional reinforcement of societal biases highlights the need for rigorous auditing during the model training phase.

Furthermore, legal frameworks such as the European Union AI Act mandate the integration of "control points" within models—mechanisms that can trigger automatic shutdowns if errors in handling sensitive data occur or if trust metrics fall below defined thresholds. However, most current NLP models do not support such features by default, underscoring a gap between regulatory expectations and available technical capabilities. These challenges call for the development of robust testing and experimental environments that can assess both the technical performance and ethical alignment of language models in real-world applications.

2.2. Differential Privacy and Its Applications in Natural Language Processing (NLP) Models

Differential privacy has become a leading innovation in ensuring individual privacy while enabling the effective use of data for model training and analysis. It provides a mathematical guarantee that personal information remains protected, even when data is processed in aggregate. This protection is especially crucial in sensitive domains where data confidentiality is paramount. The discussion begins by outlining the foundational concepts and theoretical principles of differential privacy that enable privacy-preserving data analysis. It then examines how this framework can be integrated into natural language processing (NLP) models, where the risk of exposing personal or sensitive information is particularly high. Emphasis is placed on how differential privacy allows for secure data utilization without compromising individual anonymity. In addition to the conceptual framework, the analysis reviews real-world applications of differential privacy in NLP, including its role in training large language models such as BERT and GPT. Technical challenges are also explored, particularly the trade-off between maintaining strong privacy guarantees and achieving high model accuracy. The overall aim is to clarify how privacy-preserving mechanisms can be effectively implemented within NLP systems while minimizing the impact on their performance.

2.2.1. The Concept of Differential Privacy and Its Importance in Data Protection

Differential privacy is a rigorous mathematical framework designed to safeguard individuals' privacy during the analysis of datasets containing sensitive information. Introduced by Dwork et al. in 2006, the concept aims to allow researchers and developers to extract valuable insights from data without disclosing any identifiable information about specific individuals. This ensures that privacy remains intact, even when large-scale data analysis is performed.

The core principle of differential privacy is that the inclusion or exclusion of any single record in a dataset does not significantly influence the outcome of any analysis or query. As a result, it becomes nearly impossible to detect or reconstruct personal information about any individual. This feature is especially vital in artificial intelligence and natural language processing applications, where models often process text that may contain confidential or personally identifiable information (Abadi et al., 2016).

The significance of differential privacy lies in its ability to enable organizations to leverage the power of big data while maintaining strict privacy standards. By doing so, it fosters user trust, encourages ethical data usage, and ensures compliance with global data protection regulations such as the General Data Protection Regulation (GDPR) in Europe (Voigt & von dem Bussche, 2017).

2.2.2. Applications of Differential Privacy in Natural Language Processing Models

Despite the technical complexities involved, recent research demonstrates that incorporating differential privacy into natural language processing (NLP) models—such as GPT and BERT—is both feasible and effective in reducing the risk of personal data leakage. Kairouz et al. (2019) confirm that deep learning models can be successfully trained under differential privacy frameworks while maintaining acceptable levels of performance. This underscores the growing potential of privacy-preserving AI systems in sensitive data environments.

A prominent application of this approach is "Differentially Private Federated Learning," where models are trained across decentralized data sources without centralizing the data. This method enhances user privacy and minimizes the risk of data breaches by eliminating the need to transfer sensitive information to a single location (Kairouz et al., 2019). Nevertheless, some challenges persist. Research by Li et al. (2021) indicates that raising privacy levels—typically by adding noise to the data—can negatively impact the accuracy of models, especially in language processing tasks that require high precision and context sensitivity.

These findings highlight the need for hybrid strategies, like the one adopted in the current study, which aim to strike a practical balance between classification accuracy and privacy protection. By layering rule-based,

statistical, and deep learning techniques, such methodologies can help address the privacy-performance trade-off and make privacy-aware NLP applications more viable in real-world settings.

2.2.3. Regulatory and Legal Challenges Related to Privacy in Artificial Intelligence

Legal regulations such as the European General Data Protection Regulation (GDPR) emphasize the importance of protecting personal data and reinforcing individuals' rights, with strict penalties imposed for non-compliance. Similarly, the California Consumer Privacy Act (CCPA) enforces greater transparency regarding the collection and use of personal data, requiring organizations to clearly disclose their data practices (Voigt & von dem Bussche, 2017). These frameworks demand that AI developers adopt robust privacy-preserving techniques to ensure legal adherence and uphold user rights.

Among these techniques, differential privacy has gained significant attention as a key requirement for AI systems managing sensitive data. The European AI Act further supports this direction by offering a regulatory framework to assess the privacy-related risks associated with AI applications, particularly those processing personal or sensitive information (European Parliament, 2024). However, despite these legislative advances, a practical gap persists between regulatory expectations and real-world implementations. Many current AI models still lack sufficient mechanisms to guarantee full compliance with privacy and governance standards (Smuha, 2021). This situation highlights the need for further exploration of privacy-centric methodologies, particularly the integration and evaluation of differential privacy in the design, training, and testing phases of language models. Addressing this gap is central to achieving meaningful alignment between ethical governance and technological application—an objective that guides the current study.

3. Practical Framework

This part of the study focuses on the applied dimension by exploring the effectiveness of artificial intelligence models—particularly those based on natural language processing (NLP)—in managing sensitive information within the boundaries of AI governance principles. It seeks to provide a balanced assessment of AI's capabilities from both ethical and technical standpoints, emphasizing how privacy can be safeguarded while leveraging AI in real-world scenarios. The framework also outlines the practical challenges and potential advantages of embedding governance mechanisms into NLP environments. The content is organized into two primary components. The first explores an experiment that implements governance controls by classifying textual data into "sensitive" and "non-sensitive" categories. This is achieved through a multi-model approach, featuring a hybrid system that integrates rule-based logic with automated machine learning processing. The objective is to test the model's adherence to ethical standards and evaluate its performance under these constraints. The second component introduces the concept of differential privacy in the context of training AI models. It investigates how privacy-preserving techniques affect model performance and accuracy, offering a comparative analysis between protected and unprotected model scenarios to highlight the trade-offs between privacy and efficiency.

3.1. Part One: Applying Governance through Data Classification and Training on Sensitive Data

To uphold ethical governance and ensure data privacy, the training process was carefully designed using synthetic data that mimics real-world scenarios without containing any actual sensitive information. This approach allowed realistic model training while preserving user confidentiality. A set of clear and consistent rules was established for identifying sensitive data, supported by a structured evaluation framework that regularly assesses model performance using key metrics such as Accuracy, Recall, Precision, and F1-score.

Confidence thresholds were applied strategically, allowing adjustments to reduce the likelihood of misclassification and minimize associated risks. Furthermore, data balancing was prioritized to maintain an equal representation of sensitive and general data within the training set, thus preventing model bias. The rule-based component was made adaptable, enabling ongoing updates to the classification logic based on performance feedback to improve outcomes. A hybrid evaluation strategy combining rules with statistical and

deep learning models was employed to enhance classification accuracy while maintaining the flexibility needed to address complex or novel inputs effectively.

3.1.1. First: Experimental Conditions

The experiments were carried out on a local computer equipped with moderate hardware specifications to simulate a practical and accessible environment. A lightweight and efficient programming setup was designed to ensure ease of replication by governance bodies or research entities without requiring advanced or high-cost infrastructure. The system used for experimentation was running Windows 11 as the operating system, supported by an Intel Core i5 processor and 8 GB of RAM. The programming environment relied on Python version 3.13.1, which provided a flexible and widely supported platform for developing and testing the hybrid classification model. A variety of libraries and tools were employed to support the development and execution of the hybrid model for sensitive text classification. Pandas was used extensively for data handling tasks such as organizing data into DataFrames, loading datasets, cleaning records, and reviewing classification results. For implementing rule-based classification, the re module (regular expressions) was critical. It enabled the matching of specific patterns within the text to identify whether content was sensitive or general, based on predefined linguistic rules.

Scikit-learn served as the core machine learning library, particularly useful for training the Logistic Regression model. It was also used to convert text into numerical form through the TF-IDF technique, as well as to evaluate key performance metrics such as accuracy and precision. For the deep learning component, the transformers library from Hugging Face was utilized, especially the DistilBERT model, which allowed for high-level semantic analysis and contextual understanding of the input text. Torch (PyTorch) provided the backend framework for executing deep models efficiently, with GPU support for improved performance.

To support model reuse and efficiency, Joblib was used for saving and loading trained models and vectorizers, reducing the need for retraining during multiple runs. Additionally, the csv module allowed storage of evaluation outputs and classifications in structured files, simplifying later review or integration into reports and documentation. The overall objective of this experiment was to create a hybrid classification system governed by ethical AI standards. The model aimed to prevent the output or misuse of sensitive text by categorizing content as either "sensitive" or "general." This was accomplished through a multi-layered architecture integrating linguistic rules, a TF-IDF-based logistic regression layer, and a pre-trained deep learning model. The methodology balanced the need for high classification accuracy with interpretability and operational efficiency, while also ensuring the system remained adaptable and compliant with data privacy standards.

3.1.2. Second: Experiment Implementation Stages

The implementation of the experiment was structured through carefully defined stages, beginning with the preparation of a synthetically generated dataset to simulate realistic yet anonymous scenarios involving sensitive and non-sensitive content. The data was categorized into general, sensitive, and mixed types to ensure a comprehensive testing ground for model robustness. A hierarchical three-layer classification architecture was then developed, integrating rule-based logic, statistical learning via logistic regression, and a deep learning component using the DistilBERT transformer. This multi-tiered approach enabled a progressive evaluation mechanism—each layer filtered inputs based on clarity and confidence, escalating ambiguous cases to more advanced models for accurate and ethically guided classification.

3.1.2.1. Data Preparation

The dataset used in the experiment was synthetically generated to simulate both general and sensitive questions in a hypothetical manner, deliberately distant from real-world data. This was done in a Python environment by assigning general identifiers combined with random text segments, ensuring a diverse set of samples without any risk of matching actual sensitive information. The dataset was divided into three main categories: general information (such as hobbies, favorite colors, and sports), sensitive information (including

personal identifiers like names, IDs, and emails), and mixed questions containing information indirectly, designed to test the model's flexibility. The distribution comprised 200 sensitive items, 200 general items, and 100 mixed items.

3.1.2.2. Experiment Architecture and Adopted Model Design

A hierarchical, three-layer hybrid architecture was adopted for classifying sensitive texts, combining expert-driven, statistical, and deep learning models based on decisiveness and confidence levels (Figure 1). The first layer is a rule-based system built on manually crafted rules using regular expressions (Regex). This layer serves as the initial filter to classify texts that directly and clearly match defined patterns. If a text is ambiguous or does not meet any rule, it is passed on to the next layer for further evaluation. The second layer employs statistical classification using a binary logistic regression model trained on TF-IDF representations of the texts with n-grams (1,2). This model is trained on balanced data to minimize bias and is only used when the rule-based layer cannot make a definitive classification. The model's confidence score determines whether the input proceeds to the deep learning layer. The final layer uses a deep learning approach with a DistilBERT transformer model. This pre-trained model is fine-tuned on the experiment's dataset and activated only when the statistical model fails to reach a confident decision. DistilBERT's ability to interpret complex linguistic contexts allows the model to capture subtle nuances, ensuring accurate classification in challenging cases.

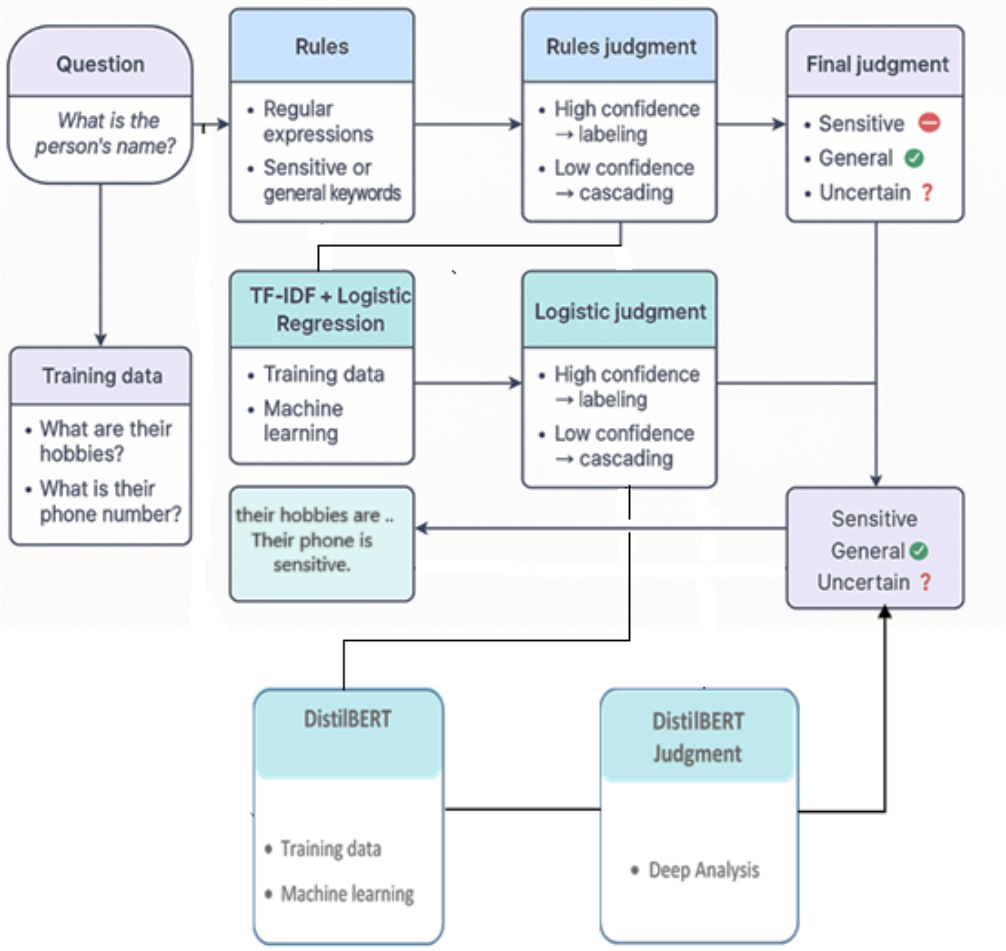

Figure 1. Three-layer hybrid architecture model.

4. Methodology

This study adopts a multifaceted methodology combining analytical, descriptive, and experimental approaches to thoroughly investigate the integration of AI governance and differential privacy within natural language processing (NLP) models. The analytical component focuses on understanding how governance principles—particularly those related to privacy protection—are operationalized in AI systems. This involves

a detailed examination of textual data classification processes, utilizing a hybrid multi-layer model that integrates linguistic rules, a logistic regression classifier based on Term Frequency-Inverse Document Frequency (TF-IDF) features, and a transformer-based model (DistilBERT). Through this analytical lens, the study explores the mechanisms by which AI models manage sensitive data while adhering to governance requirements.

The descriptive methodology complements the analytical approach by systematically documenting the architecture and functioning of the employed models and techniques. This includes elaborating on the role of linguistic rules in pre-processing and categorizing text, explaining how logistic regression is applied for binary classification tasks, and describing the fine-tuning and inference stages of the DistilBERT transformer model. Furthermore, the descriptive aspect extends to the practical experiment designed to generate synthetic datasets representing sensitive, non-sensitive, and ambiguous text categories. These synthetic datasets serve as controlled environments for testing model behavior and compliance with governance standards.

The experimental approach forms the core of the research by empirically assessing the impact of differential privacy mechanisms on model performance and compliance. Differential privacy is implemented by injecting calibrated noise during the training process to mask individual data contributions, thereby safeguarding user privacy. Multiple classification experiments are conducted, comparing the performance of models trained with and without differential privacy constraints. Quantitative performance indicators such as accuracy, precision, recall, and F1-score are calculated to measure the trade-offs between privacy protection and classification efficacy. Additionally, qualitative evaluations are undertaken to interpret the ethical and regulatory implications of observed model behaviors.

By integrating these methodologies, the study not only provides a comprehensive theoretical understanding of AI governance and privacy but also offers practical insights through empirical validation. This combined approach allows for a robust examination of the research hypotheses, emphasizing both the technical feasibility and ethical necessity of applying differential privacy in NLP applications. It ultimately aims to contribute to the development of AI systems that are both effective in their tasks and responsible in protecting individual privacy in compliance with emerging regulatory frameworks.

4.1. Testing Methodology

The testing methodology for the hybrid model involved using a manually selected dataset that encompassed a broad range of cases, including those that clearly fit predefined rules as well as those that did not. This comprehensive selection ensured that the model's ability to handle both straightforward and ambiguous texts was thoroughly evaluated. During testing, the processing sequence for each text was carefully tracked through the model's three stages: starting with the rule-based layer, followed by logistic regression, and finally the BERT model if needed. This tracking allowed for a detailed understanding of how each input was classified and at which stage the decision was made. Performance metrics were assessed not only for the overall hybrid model but also individually for each stage in the processing sequence. This granular evaluation provided insights into the contribution and effectiveness of each classification layer, highlighting areas of strength and opportunities for further refinement within the hybrid approach.

4.2. Criteria for Selecting Questions and Data

The selection of questions and data for the study was guided by specific criteria aimed at ensuring the robustness and effectiveness of the model (Table 1). A primary consideration was achieving balance, which involved maintaining an equal number of samples across the main categories of sensitive and general data. This balance is crucial to prevent bias during training and to allow the model to learn and differentiate effectively between the two classes. Another important criterion was linguistic coverage. The dataset was designed to include a wide variety of vocabulary and phrases, covering diverse expressions and terminologies. This diversity helps the model develop comprehensive coverage, improving its ability to handle different textual styles and nuances encountered in real-world scenarios. To test the model's flexibility, cases that presented ambiguity were deliberately included. These are examples that are difficult to classify using

straightforward rule-based methods, challenging the model to make accurate decisions in less clear-cut situations. This approach ensures the model is not overly reliant on rigid rules and can adapt to complex inputs. Finally, contextual sensitivity was considered by incorporating questions whose classification as sensitive or general could change depending on the surrounding context. This aspect evaluates the model's accuracy in recognizing when sensitivity is conditional, reflecting real-life complexities in data classification and enhancing the model's practical relevance.

Table 1. Criteria for selecting questions and data.

Criterion	Details
Balance	Equal number of samples across the main categories (sensitive, general).
Linguistic Coverage	Inclusion of diverse vocabulary and phrases to ensure comprehensive model coverage.
Ambiguity	Inclusion of cases difficult to classify by rules to measure model flexibility.
Contextual Sensitivity	Inclusion of questions whose sensitivity changes depending on context to evaluate accuracy.

5. Model Governance and Calibration

The model was developed with a strong commitment to ethical AI governance principles, ensuring responsible and transparent decision-making throughout its operation. One key aspect of this approach is that no final classification is made when confidence levels are low unless the input has been evaluated by the deep transformer model. This safeguard helps prevent premature or uncertain decisions that could compromise accuracy or privacy. Priority is also given to interpretable and explainable decisions, with the model first relying on rule-based methods, followed by logistic regression. These layers provide clear, understandable reasoning behind classifications, which is essential for accountability and trust, especially in sensitive domains. The use of deep learning models, such as the transformer-based BERT, is reserved only for cases where simpler methods fail, minimizing computational costs and reducing potential privacy risks associated with complex models. Additionally, the model's architecture is designed for flexibility and adaptability. Each classification layer can be easily replaced or updated independently, which enhances the system's correctability and allows for ongoing improvements. This modularity supports continuous refinement and alignment with evolving ethical standards, technological advances, and domain-specific requirements.

5.1. Model Deployment

The deployment of the hybrid model follows a layered approach to classification. Initially, the input text is processed through the rule-based layer, which attempts to classify the text using predefined linguistic patterns. If the rule layer does not produce a result, the input is passed on to the logistic regression model for further analysis. Should the logistic model also fail to classify the text, the input is finally processed by the BERT model, which handles the most complex and ambiguous cases. Model sensitivity is regulated by a confidence threshold—commonly set at 0.6—to balance between classification confidence and coverage. Additionally, the rule sets are designed to be flexible and customizable, allowing modifications tailored to specific domains such as healthcare, education, or other specialized fields.

Table 2 outlines the key performance metrics used to evaluate the hybrid model. Overall model accuracy measures the proportion of correct classifications out of the total samples processed. Layer accuracy assesses the performance of each classification layer independently, providing insight into their individual effectiveness. The BERT layer access rate reflects the percentage of cases resolved only at the final layer, highlighting the complexity of those inputs. The "uncertain" rate indicates the proportion of texts for which no definitive classification was made, serving as a measure of model hesitation. Finally, the average confidence per layer quantifies the certainty of the model's decisions across different stages.

For training and evaluation, the dataset was carefully curated to include a variety of content types, combining texts and questions related to both personal data—such as names, emails, and phone numbers—and more general information like hobbies, skills, and favorite colors. Generative codes were employed to simulate realistic data distributions that mirror what might be encountered in diverse real-world settings, ensuring a balanced representation of sensitive and general data. In total, 500 records were generated, providing sufficient coverage for the models to learn effectively while maintaining computational efficiency. Prior to

training and testing, each record was labeled as "sensitive" or "general" according to established privacy and sensitivity criteria to guarantee accurate classification outcomes.

Table 2. Performance metrics for the hybrid model.

Metric	Description
Overall Model Accuracy	Number of correct classifications ÷ total samples
Layer Accuracy	Performance evaluation of each layer separately
BERT Layer Access Rate	Percentage of texts resolved only at the last layer
"Uncertain" Rate	Percentage of texts with no final classification
Average Confidence per Layer	Measures hesitation of models during classification

6. Analysis of Hybrid Model Results

The performance analysis of the hybrid model in the context of implementing governance standards and information classification reveals robust capabilities, particularly in handling complex and ambiguous text cases (Table 3 and Figure 2). With an overall accuracy of 91%—excluding uncertain cases—the model demonstrates a strong ability to classify sensitive and non-sensitive information even when inputs do not conform strictly to predefined rules. This level of accuracy is especially noteworthy in real-world scenarios where textual ambiguity and rule inconsistencies are common challenges.

Table 3. Results of hybrid model.

Metric	Value
Accuracy	0.91
Precision	0.89
Recall	0.933852
F1-Score	0.914286

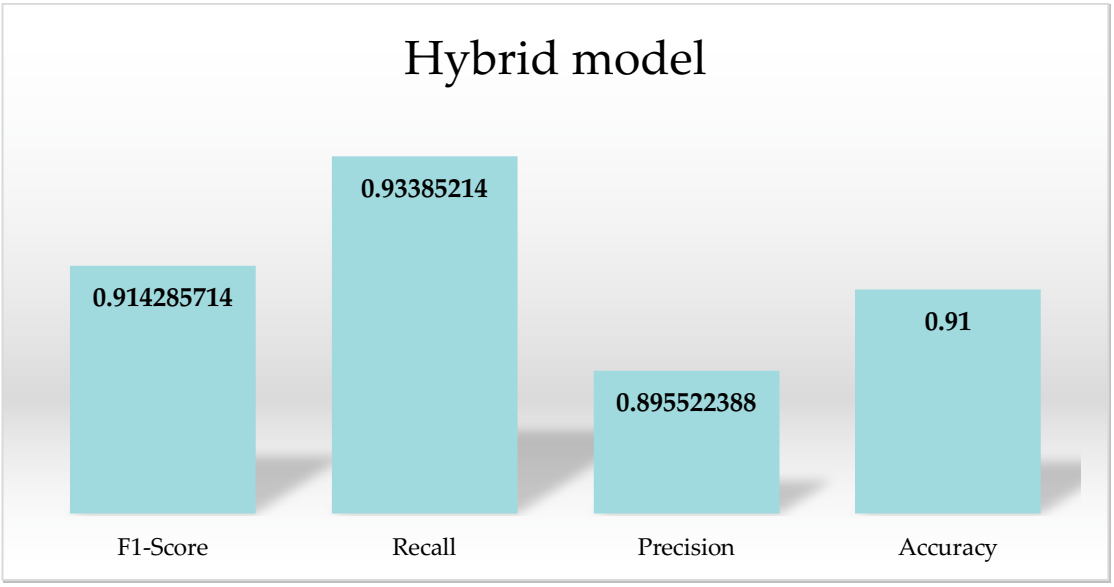

Figure 2. Results of hybrid model.

The overall accuracy of the hybrid model reached 91% after excluding uncertain cases, reflecting its strong capability to correctly classify both sensitive and non-sensitive texts, even in the presence of ambiguity and complex linguistic structures. This level of accuracy is considered high, especially given that the dataset includes instances that do not align perfectly with predefined classification rules, closely mimicking real-world conditions where not all inputs are straightforward or easily interpreted. The precision of the model was recorded at 89%, indicating that most of the instances identified as sensitive were indeed correctly labeled. This demonstrates the model's effective ability to minimize false positives, which is particularly important in applications that demand a balance between protecting user privacy and maintaining operational flexibility. Moreover, this level of precision can further improve with larger and more diverse datasets.

The model achieved a recall of 94%, signifying that it was successful in identifying the vast majority of truly sensitive cases. This high recall rate reduces the likelihood of overlooking sensitive information—an essential feature in domains like healthcare, finance, and governance, where such errors could lead to serious privacy breaches. Complementing this, the F1 score reached 92%, highlighting the model's ability to maintain a strong balance between sensitivity and specificity by minimizing both false positives and false negatives. A detailed analysis of the confusion matrix reveals that 210 non-sensitive cases and 245 sensitive cases were correctly classified. However, 33 non-sensitive cases were mistakenly marked as sensitive, while 12 sensitive cases were missed. These results show a well-managed distribution of errors, suggesting consistent and controlled model behavior.

In terms of classification methods, approximately 87% of decisions were made using rule-based techniques, showcasing the strength of linguistic pattern recognition. Logistic regression accounted for 11% of the decisions, proving useful in borderline or less clear cases. The BERT model, though used in only 2% of classifications due to confidence thresholds, played a vital role in resolving the most ambiguous inputs, where neither rules nor statistical models were sufficient. Importantly, there were no uncertain cases in this experiment, indicating that the confidence threshold settings were appropriately tuned. This outcome supports the reliability of the model's predictions, offering both consistency and transparency in decision-making processes. In conclusion, an accuracy rate of 91% under conditions that include inconsistent or ambiguous data demonstrates the model's robustness. These findings validate the hybrid approach's practical value and confirm its suitability for real-world applications involving the classification of sensitive information under stringent privacy governance requirements.

7. Analysis of Individual Model

The measurement outputs for the experiment were recorded for each model mentioned in Table 4.

Table 4. Results of individual model.

Metric	Bert	Logistic	Rule-Based
Accuracy	0.886	0.922	0.965
Precision	1.000	0.869	0.986
Recall	0.779	1.000	0.946
F1-Score	0.875	0.930	0.965

7.1. Rule-Based Model (Rule)

The Rule-Based model demonstrated a high accuracy of 96.5%, showcasing the strength of using well-defined pattern rules for classifying texts with clear indicators of sensitivity or generality (Figure 3). This high accuracy highlights the model's reliability in environments where textual cues are explicit and consistent. The model's precision was recorded at 98.7%, indicating that nearly all cases identified as sensitive were indeed sensitive. This strong precision reflects the model's effectiveness in avoiding false positives, making it highly dependable for correctly flagging sensitive information. With a recall rate of 94.6%, the model successfully retrieved the majority of sensitive examples. However, a small percentage of actual sensitive cases went undetected, suggesting slight limitations in handling ambiguous or less explicitly marked data. The positive error rate of 96.5% demonstrates an excellent balance between precision and recall, confirming that the model performs reliably across both metrics. According to the confusion matrix, only 3 general cases were mistakenly classified as sensitive, while 12 sensitive cases were incorrectly labeled as general. These minimal errors affirm the rule-based model's overall accuracy and its practicality for applications with well-defined textual patterns.

7.2. Logistic Regression (Logistic)

The Logistic Regression model achieved an accuracy of 92.2%, which is considered strong but slightly below that of the rule-based model (Figure 4). This minor drop in accuracy can be attributed to the model's reliance on a mathematical representation of textual features during training, which may allow for subtle mismatches due to partial textual differences. In terms of precision, the model recorded a score of 87%, indicating that some non-sensitive cases were incorrectly classified as sensitive. This increase in false positives suggests that

the model prioritizes capturing all possible sensitive cases, even at the risk of over-identifying. A key strength of the Logistic Regression model lies in its recall, which stands at a perfect 100%. This means the model successfully detected every actual sensitive case in the dataset, making it highly reliable for applications where missing sensitive information is unacceptable. The positive error rate of 93% further supports the model's overall robustness, although its tendency toward inclusiveness comes at the expense of precision. According to the confusion matrix, 38 general (non-sensitive) cases were wrongly flagged as sensitive, reinforcing the model's bias toward avoiding false negatives, even if it results in some over classification.

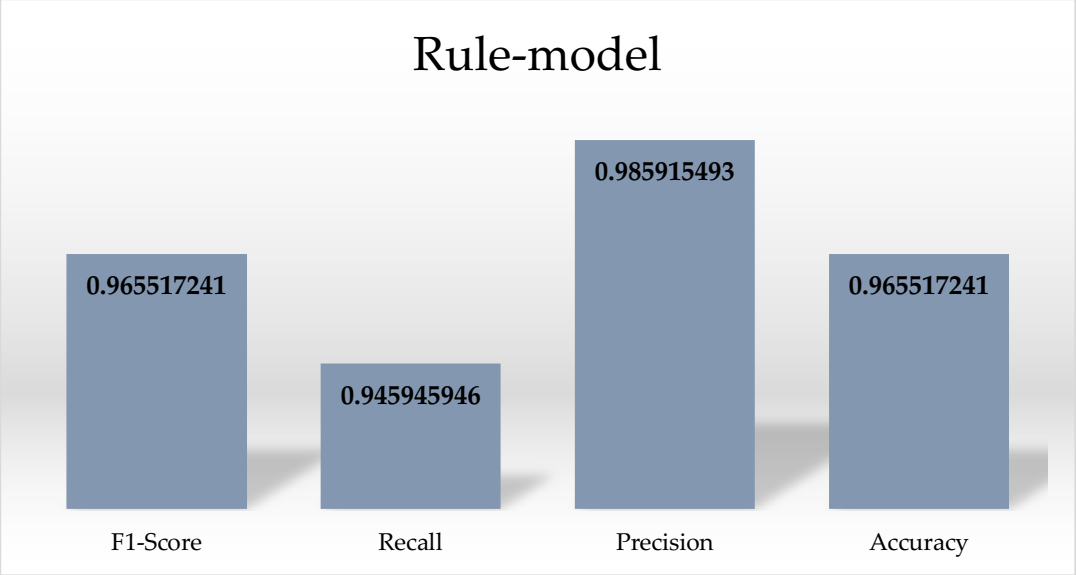

Figure 3. Performance of the Rule-Based model.

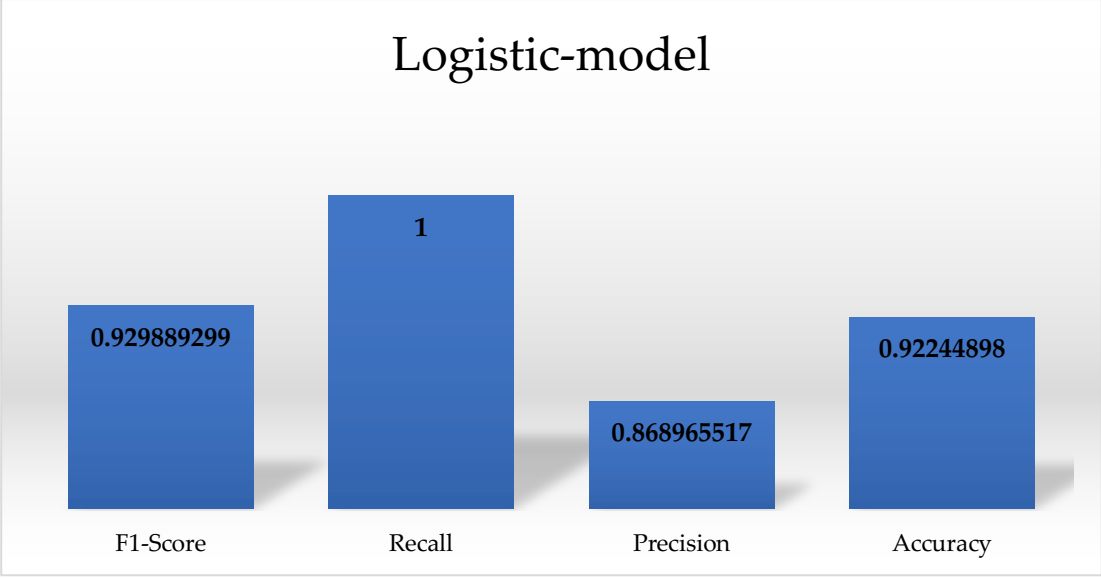

Figure 4. Performance of the logistic regression model.

7.3. BERT Model

The BERT model demonstrated an overall accuracy of 88.6%, which is the lowest among the three models evaluated in this study (Figure 5). This suggests that, despite its powerful language understanding capabilities, BERT faced challenges in consistently achieving precise classifications across all cases. However, the model achieved a perfect precision score of 100%, indicating that every case it identified as sensitive was indeed correct, with zero false positives. This reflects a high level of confidence in the classifications it did make regarding sensitive content. On the other hand, the recall rate stood at 78%, revealing a limitation in the model's ability to detect all sensitive cases. Specifically, 22% of the sensitive items were not identified, which could be problematic in applications where missing sensitive information has serious implications. The

positive error balancing rate of 88% remains solid, though it is impacted by the model's reduced recall. The confusion matrix further illustrates this issue, showing that 57 sensitive cases were missed. This highlights a trade-off in BERT's performance, where high precision comes at the expense of recall, suggesting the need for enhancement in sensitivity detection. Figure 6 presents the comparison among the models.

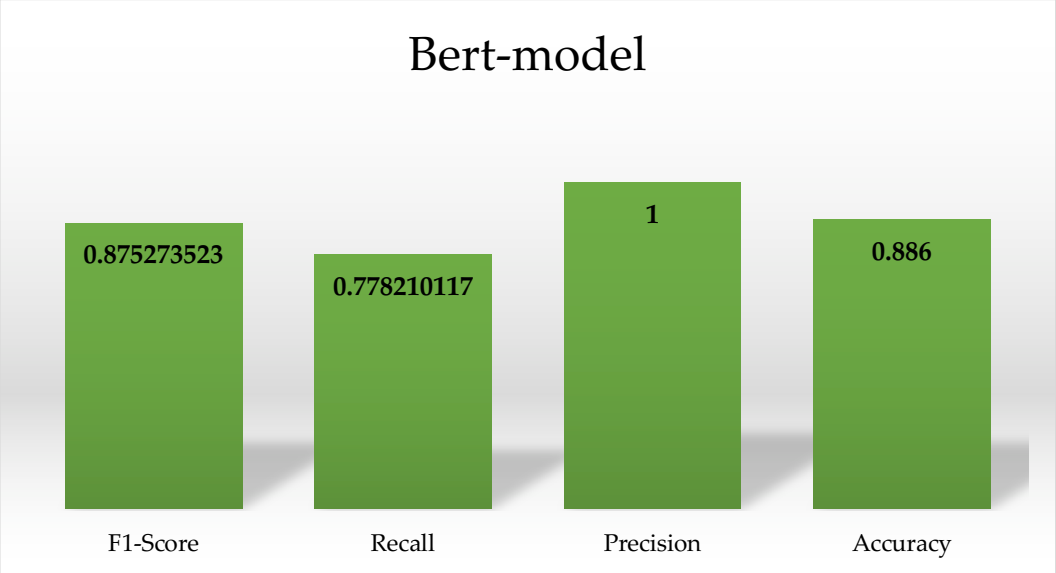

Figure 5. Performance of BERT model.

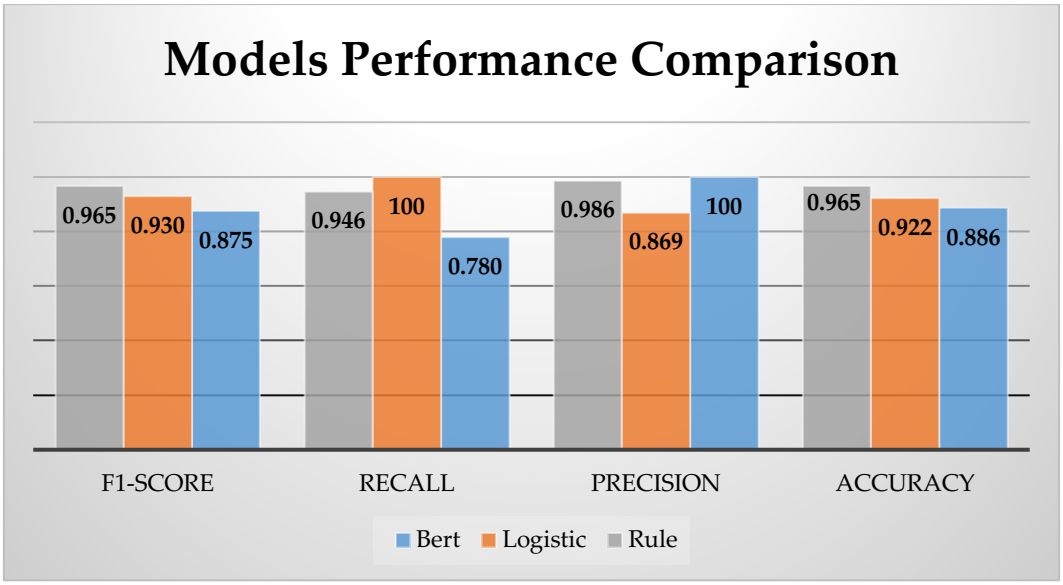

Figure 6. Comparison results of models composing the hybrid model.

8. Part Two: Implementing Governance through Differential Privacy

8.1. Experiment Design

A synthetic textual dataset was utilized in this study, comprising narrative sentences and questions categorized into two distinct classes (Figure 7). Sensitive data, such as names, email addresses, and ID numbers, were labeled with a value of (1), while non-sensitive information, including hobbies, favorite colors, and preferred sports teams, was assigned a value of (0). This binary classification facilitated clear distinctions during model training and evaluation. To improve the diversity of the dataset and simulate more natural language variations, questions were incorporated alongside narrative texts at a proportion of 20%. This addition aimed to mimic real-world communication patterns, where queries often accompany statements, enhancing the model's exposure to varied sentence structures. For effective and unbiased model evaluation,

the dataset was divided into separate training and testing sets. This split ensured that the model was assessed on unseen data, providing a more accurate measure of its generalization capabilities and overall performance.

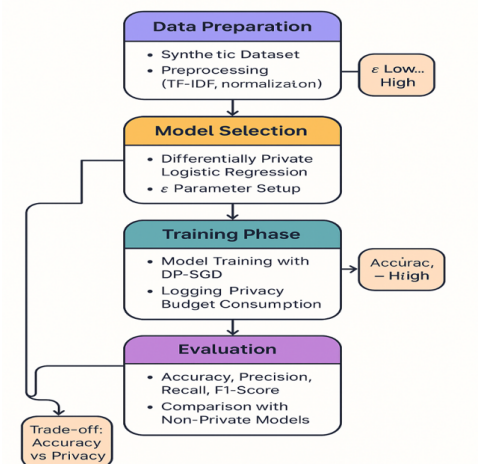

Figure 7. Governance experiment design using differential privacy data preparation.

8.2. Model Selection

The Differentially Private Logistic Regression model was implemented using the diffprivlib library, which offers built-in mechanisms to ensure the protection of sensitive data during model training. This library is specifically designed to comply with differential privacy standards, making it suitable for applications where data confidentiality is a key concern. To examine the effect of privacy levels on model performance, the privacy parameter ϵ (epsilon) was tuned using different values—specifically, 3, 5, and 6. These values were selected to evaluate how varying degrees of privacy influence the accuracy and reliability of the model. By testing across this range, the study was able to observe the trade-offs between strong privacy guarantees and practical model performance.

8.3. Testing Procedure

The model was trained on the preprocessed dataset by transforming textual inputs into numerical representations using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. This method enabled the model to quantify the importance of words within documents, allowing effective feature extraction for classification tasks. During training, various values of ϵ (epsilon) were applied to evaluate their influence on the balance between privacy and accuracy. Each setting offered insights into how different levels of differential privacy impact the model's ability to learn and generalize from the data. Privacy-related warnings encountered during training were addressed either by adjusting the necessary parameters or by accepting a certain level of information leakage, which is known to occur at specific epsilon values. This practical approach ensured that training could proceed without compromising the study's objective of exploring the privacy-accuracy trade-off.

8.4. Results Measurement and Comparison

Increasing the value of ϵ (epsilon) in Differentially Private Logistic Regression directly enhances model performance but comes at the cost of reduced privacy (Table 5). A higher epsilon value allows the model to access more informative data patterns, leading to improved accuracy and predictive capabilities. At lower epsilon values, such as 3, the model offers stronger privacy guarantees, which is critical for sensitive applications. However, this heightened privacy comes with a trade-off in performance, as the model may struggle to generalize effectively, resulting in reduced accuracy. On the other hand, setting epsilon to 6 yields the highest performance among the tested configurations, making the results closely comparable to those of non-private models. Nevertheless, this setting provides weaker privacy protection, which may not be suitable for applications requiring stringent confidentiality.

Table 5. Measurement of results and comparison.

ϵ\epsilonϵ	Accuracy	F1 Score	Notes
3	78%	0.78	High privacy, good performance with accuracy decline
5	82%	0.81	Good balance between privacy and accuracy
6	97%	0.97	Strong performance similar to non-private models, weaker privacy

9. Conclusions

The research highlights the strengths and limitations of various classification models used in identifying sensitive textual content. The rule-based model performs effectively when clear textual indicators are present but struggles with ambiguity. Logistic regression ensures full recall of sensitive cases, though it sometimes incorrectly classifies general content as sensitive. The BERT model, known for its precision, accurately detects sensitive cases but has a lower recall rate, causing it to miss some instances. A hybrid model emerges as a promising approach, offering a balanced performance and a higher degree of confidence, with room for further experimentation and refinement. When incorporating Differentially Private Logistic Regression, setting the privacy parameter (ϵ) is crucial to balance privacy and model accuracy. For highly sensitive applications such as medical or governmental data, lower ϵ values (between 3 and 5) are advisable, even at the cost of some performance degradation. Conversely, in contexts where accuracy holds more significance and moderate privacy is acceptable, higher ϵ values (around 6) are more suitable, delivering results comparable to traditional models. Future research using real-world datasets and broader experimentation is encouraged to establish optimal privacy settings tailored to specific application domains. Moreover, the choice of data preparation and training methods should prioritize maintaining equilibrium between dataset size and model complexity to enhance the effectiveness of differential privacy mechanisms.

Author Contributions:

Conceptualization: Ziad Abdullah Alotaibi, Ziyad Ibraheem AlZaidan.
Data curation: Ziad Abdullah Alotaibi, Ziyad Ibraheem AlZaidan.
Formal analysis: Ziad Abdullah Alotaibi.
Funding acquisition: Ziad Abdullah Alotaibi, Ziyad Ibraheem AlZaidan.
Investigation: Ziyad Ibraheem AlZaidan.
Methodology: Ziad Abdullah Alotaibi.
Project administration: Ziyad Ibraheem AlZaidan.
Resources: Ziad Abdullah Alotaibi, Ziyad Ibraheem AlZaidan.
Software: Ziyad Ibraheem AlZaidan.
Visualization: Ziad Abdullah Alotaibi.
Writing – original draft: Ziad Abdullah Alotaibi, Ziyad Ibraheem AlZaidan.
Writing – review & editing: Ziyad Ibraheem AlZaidan.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available upon request from the authors.

Conflicts of Interest: The author(s) declares no conflicts of interest.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318. https://doi.org/10.1145/2976749.2978318

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732. https://doi.org/10.2139/ssrn.2477899

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2020). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213. https://arxiv.org/abs/2004.07213

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2021). The secret sharer: Evaluating and testing unintended memorization in neural networks. USENIX Security Symposium, 267-284. Retrieved from: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference (TCC), 265–284. https://doi.org/10.1007/11681878_14

European Commission. (2019). Ethics guidelines for trustworthy AI. Retrieved from: https://ec.europa.eu/digital-strategy/en/news/ethics-guidelines-trustworthy-ai

European Parliament. (2024). Artificial Intelligence Act: Regulation laying down harmonised rules on artificial intelligence. EUR-Lex. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hu, L., Habernal, I., Shen, L., & Wang, D. (2023). Differentially Private Natural Language Models: Recent Advances and Future Directions. arXiv. https://arxiv.org/abs/2301.09112

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2019). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), 1–210. https://doi.org/10.1561/2200000073

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2021). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.3010695

Office of Science and Technology Policy (OSTP). (2022). Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. The White House. Retrieved from: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/

Organisation for Economic Co-operation and Development (OECD). (2019). Recommendation of the Council on Artificial Intelligence. OECD Legal Instruments. Retrieved from: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog. Retrieved from: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Smuha, N. A. (2021). From a ‘race to AI’ to a ‘race to AI regulation’: Regulatory competition for artificial intelligence. Law, Innovation and Technology, 13(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide (1st ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. Global Catastrophic Risks, 308-345. Retrieved from: https://intelligence.org/files/AIPosNegFactor.pdf

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.