Journal of Information Systems Engineering and Management
2026, 11(1s) e-ISSN: 2468-4376 https://www.jisem-journal.com/ **Research Article**

AI/ML-Based Data Sensitivity Classification: A Technical Framework

Avinash Reddy Thimmareddy
Osmania University, India

ARTICLE INFO

Received: 20 Jan 2026

Revised: 23 Jan 2026

ABSTRACT

Increasingly complex information system environments are posing mounting challenges to organizations in determining and defending sensitive data. Conventional rule-based classifications have not been effective in managing various schemas, ambiguous metadata, and dynamic data structures, which typify new enterprise contexts. This architecture introduces an automated system that uses techniques of artificial intelligence and machine learning to categorize data sensitivity on a large scale. The framework uses multiple-dimensional feature extraction using column names, descriptions, table contexts, data types, and semantic embeddings. Transformer-based and classical models, such as BERT, TFIDF, and ensemble classifiers, respectively, convert textual metadata into representations that can be used to make predictions. Multi-class architecture separates Personally Identifiable Information, Protected Health Information, and general sensitivity categories and fits various regulatory requirements in GDPR, HIPAA, and CCPA frameworks. Strict testing based on precision, recall, F1-score, and confusion matrix testing ensures the production-quality performance on uneven datasets that are characteristic of enterprise data catalogs. The framework saves a lot of manual classification effort and also provides high accuracy, which can allow the enforcement of policies automatically, faster compliance initiatives, and more mature data governance. Application in the healthcare, financial services, and technology sectors has shown a significant payoff in the form of lower compliance risks, lower operational overhead, and improved data protection capacity to facilitate digital transformation goals.

Keywords: Data Sensitivity Classification, Machine Learning, Natural Language Processing, Personally Identifiable Information, Protected Health Information

1. Introduction

Companies in the present era are faced with massive amounts of data holding information of various sensitivity levels. Detection and separation of sensitive information, such as Personally Identifiable Information (PII) and Protected Health Information (PHI), has become mission-critical in order to remain compliant, administer data correctly, and ensure security. The data protection expenses, which have been lot, have reached shocking amounts. Fresh industry analyses show how businesses now suffer massive financial hits when sensitive material gets exposed or handled carelessly. Healthcare providers take especially brutal financial beatings thanks to tough regulations surrounding patient records and demanding HIPAA compliance standards.

Current data environments bring staggering complexity—information bounces between cloud setups, local servers, and mixed infrastructures. This situation calls for much smarter classification tools that can roll with different contexts and meanings. This piece lays out a thorough AI/ML framework for automated data sensitivity classification that taps into natural language processing and machine learning methods to smartly group data using several contextual clues. This strategy tackles head-on the central problem of spotting sensitive material at a massive scale while keeping both sharp accuracy and smooth operations across different company settings. It is crucial to understand that this classification framework operates on metadata rather than actual data content. The machine learning models receive inputs such as column names, column descriptions, table names, table descriptions, and data type specifications—not the data values stored within those columns. This metadata-driven approach enables efficient classification at scale without requiring access to potentially sensitive data contents, thereby maintaining privacy during the classification process itself. The models learn to recognize patterns in how sensitive data is labeled, described, and structured within database schemas, rather than analyzing the actual stored values.

2. Feature Engineering for Data Sensitivity Detection

Effective data sensitivity classification begins with extracting meaningful features that capture both semantic and structural characteristics of data elements. Building a robust classification framework requires systematic approaches that consider multiple dimensions of data attributes through comprehensive metadata analysis A thorough data classification setup covers finding, grouping, and tagging data based on how sensitive it is, what business value it holds, and what regulations demand [2]. This structural backbone matters for solid data management and protection game plans. Setups like these usually blend metadata checks, content reviews, and context reviews to figure out the right classification tiers, making certain that touchy material gets proper protection while keeping data reachable for legit business needs.

Checking column names works as a main feature-pulling mechanism. Tags like "ssn", "dob", and "email" give instant hints about possible data touchiness. These naming patterns often mirror company standards, database blueprints, and field-specific lingo that link strongly with particular data groups needing protection. The checking stretches past basic keyword hunting to grab pattern spotting in compound column labels, short forms, and specialized vocabulary that might flag touchy stuff. Parsing column descriptions throws in another key dimension to feature pulling. Descriptive metadata boxes like "Customer Social Security Number" or "Patient Date of Birth" spell out clear context about what data fields are and what purpose they serve. These descriptions, when sitting in data dictionaries or schema paperwork, dish out rich text for natural language processing methods to yank out meaning and purpose.

Checking table-level context brings a layered understanding by looking at table names and descriptions like "Patient Medical Records" or "Customer Financial Transactions." This pins down the field and likely sensitivity tier of all data pieces inside. Higher-up context really helps sort out columns with bland or vague names that might mean either touchy or regular information based on what table holds them. Checking data types brings in structural limits that trim down classification options. Tech data types like string, date, numeric, and binary formats drop hints about what kind of stored material sits there. Mixing data type patterns with length limits, format specs, and value range boundaries builds powerful telling-apart features for classification models.

The effectiveness of a data classification framework depends on how well it integrates technical metadata with business context and regulatory requirements [2]. Organizations implementing these frameworks must balance automation with human oversight, ensuring machine learning models receive sufficient training data representing the full spectrum of sensitivity categories relevant to their operational context. The framework must accommodate evolving schema definitions and adapt to changing regulatory landscapes that continuously introduce new categories of protected information. Historical data reveals that most enterprise data remains unclassified, creating significant compliance gaps and security vulnerabilities [3]. This underscores the critical need for automated classification capabilities that can process large-scale data repositories efficiently.

The feature extraction process operates exclusively on metadata elements. The classification models do not access or analyze the actual data values stored in database columns. Instead, the models consume structural and descriptive metadata including column identifiers (e.g., "ssn", "customer_email"), textual descriptions from data dictionaries (e.g., "Customer Social Security Number"), table-level context (e.g., "Patient_Medical_Records"), and technical specifications such as data types (STRING, DATE, INTEGER) and constraints (length, format patterns). This metadatacentric approach allows classification to occur without exposing sensitive data contents, making the framework suitable for privacy-preserving data governance initiatives.

Feature Category	Description	Classification Value
Column Name	Identifiers like ssn, dob, and email indicate sensitivity	Primary indicator for automated detection
Column Description	Metadata text explaining the field purpose and content	Provides explicit contextual information
Table Description	Higher-level context establishing domain and sensitivity scope	Enables hierarchical classification logic
Data Type	Technical type constraints (string, date, numeric, binary)	Narrowing classification possibilities
Semantic Embeddings	NLP-generated dense vector representations	Captures deep semantic relationships

Table 1: Metadata Feature Categories for Data Sensitivity Detection [3, 4]

3. NLP-Based Feature Extraction Techniques

Natural language processing serves as the technological foundation for transforming raw metadata (column names, descriptions, table contexts) into numerical representations that machine learning models can process. Recent advances in transformer-based language models have fundamentally changed how systems interpret textual metadata from database schemas and data catalogs. Pretraining techniques allow models to learn rich contextual representations from large text collections, which are then adapted for specific tasks like data sensitivity classification [5]. These models use bidirectional attention mechanisms that enable each word to consider all other words in the input sequence, creating representations that capture nuanced semantic relationships and contextual dependencies.

Transformer-Based Approaches: BERT and similar transformer architectures generate contextualized embeddings where each word's representation depends on its surrounding context rather than having a fixed meaning [5]. This contextual understanding is essential for data classification tasks because terms like "ID" or "number" have different sensitivity implications depending on surrounding words and their position within schema hierarchies. The models process input text through multiple layers of self-attention and feed-forward networks, progressively building increasingly abstract representations that capture both syntactic structure and semantic meaning. When applied to data classification, these transformer embeddings enable models to recognize that terms like "customer_email" and "user_electronic_address" are semantically similar despite having different surface forms.

Fine-tuning pre-trained language models on domain-specific datasets improves classification performance by adapting general linguistic knowledge to specialized vocabularies and naming conventions common in particular industries. Healthcare-focused models learn to recognize medical terminology, diagnosis codes, and clinical abbreviations that indicate PHI, while financial services models become attuned to account identifiers, transaction descriptors, and regulatory terminology. The computational requirements for transformer-based approaches remain substantial, with inference typically requiring specialized hardware or cloud-based GPU resources to maintain acceptable latency for real-time classification scenarios. Organizations must carefully balance performance benefits against infrastructure costs and deployment complexity when selecting appropriate NLP techniques.

Classical NLP Methods: Traditional NLP approaches provide practical alternatives that balance performance with computational efficiency and model interpretability. Automated sensitive data classification systems using privacy-aware machine learning models have demonstrated that conventional feature engineering combined with ensemble learning techniques can achieve competitive accuracy while maintaining transparency in decision-making processes [6]. These approaches typically employ Term Frequency-Inverse Document Frequency (TF-IDF) vectorization to convert textual metadata into numerical feature vectors that capture the relative importance of terms across the collection of column names and descriptions. The resulting sparse vectors serve as input to classification algorithms including logistic regression, support vector machines, and random forest ensembles that learn decision boundaries separating different sensitivity categories.

Random forest classifiers offer particular advantages for data classification tasks by handling nonlinear relationships between features, providing feature importance rankings that aid model interpretation, and delivering consistent performance across diverse dataset characteristics. The ensemble nature of random forests, which aggregates predictions from multiple decision trees trained on different bootstrap samples of the data, provides natural uncertainty quantification and reduces overfitting risks.

Feature engineering for classical approaches extends beyond simple term frequencies to include n-gram features (capturing adjacent word combinations), character-level patterns (indicating common prefixes or suffixes in sensitive column names), and structural features (such as column position within tables or relationships to other columns through foreign key constraints) [6]. These manually engineered features encode domain expertise and data modeling conventions that help classification models identify sensitivity patterns even with limited training data.

Technique Type	Examples	Advantages	Considerations
Transformer Models	BERT, contextualized embeddings	High accuracy, semantic understanding, generalization	Requires specialized hardware, longer inference time
Classical Methods	TF-IDF, Random Forest, Logistic Regression	Computational efficiency, interpretability, and fast deployment	May need more feature engineering
Hybrid Approaches	Combined transformer and classical features	Balanced performance and efficiency	Increased implementation complexity

Table 2: NLP Technique Comparison for Feature Extraction [5, 6]

4. Multi-Class Classification Architecture

The classification setup rolls out a fancy multi-class grouping system that tells apart different sensitivity tiers while fitting the tangled rule landscape running data protection across regions and fields. Protected Health Information classification brings one-of-a-kind headaches thanks to superspecific definitions laid out by healthcare privacy rules and the tech maze of medical data systems. PHI covers individually identifiable health information created, gathered, kept, or sent by covered groups and business partners, including demographic data, medical records, health insurance details, and account information when tied to healthcare services [4]. Building safe AI systems for healthcare workflows needs careful thinking about privacy-keeping tricks and tough classification setups that can reliably flag PHI across different data structures.

Classification models built for PHI spotting must handle the special traits of healthcare data like tangled coding systems, including ICD-10 diagnosis codes, CPT procedure codes, and LOINC laboratory tags that need specialized training material and field knowledge to read right [7]. Electronic health record systems pack organized data in relational databases next to messy clinical notes, imaging metadata, and scanned papers, forcing classification moves that can work across multiple data flavors. The time dimension of healthcare data throws in complexity, since how touchy material is might shift over time, with some data bits getting less touchy after set keeping periods or patient death. Classification systems must also account for roundabout identifiers that might not individually make up PHI but can help re-identification when mixed with other data bits, needing sharp inference skills to judge re-identification dangers.

Personally Identifiable Information classification covers a wider span of data flavors past healthcare, including financial account numbers, government identifiers, biometric data, online identifiers like IP addresses and cookies, and different contact material forms. What PII means and covers swings wildly across rule frameworks, with GDPR using the personal data concept meaning information touching identified or identifiable natural persons, while CCPA uses personal information meaning information that identifies, touches, or could reasonably link with a particular consumer or household [7]. Classification models must therefore support flexible taxonomy layouts that can fit region-specific demands and company policy twists.

The multi-class classification move enables detailed grouping past simple binary touchy or regular tags, backing regulatory compliance workflows that need different handling steps for different data flavors. Financial rules might demand specific keeping periods and access controls for account transaction data, while privacy laws slap on consent demands and right-to-deletion duties for personal identifiers. The classification output packs not just category tags but also confidence numbers that enable threshold-based workflows where high-confidence guesses trigger automated policy application while lower-confidence guesses are sent to human checkers for validation [4]. This mixed move balances automation speed with accuracy demands, knowing that classification slip-ups can carry major compliance and operational hits.

Organized prediction outputs shaped as JSON objects, smooth integration with downstream data management platforms, access control systems, and data loss prevention tools that eat classification metadata to enforce policies. Throwing in confidence numbers enables fancy decision logic where companies can dial acceptance thresholds based on risk comfort, compliance demands, and operational limits. High-confidence guesses might automatically fire off data masking or encryption, medium-confidence guesses could flag data for periodic checking, and low-confidence guesses might need immediate human classification before data becomes reachable for analytical uses. This stepped response ability lets companies hit comprehensive coverage while managing the operational weight of human review flows.

Category	Scope	Regulatory Framework	Classification Challenges
Personally Identifiable Information (PII)	Names, emails, SSN, phone numbers, addresses	GDPR, CCPA	Geographic variations, indirect identifiers
Protected Health Information (PHI)	Medical records, diagnosis codes, treatment data	HIPAA	Complex coding systems, temporal sensitivity
Sensitive/Non-Sensitive	Binary classification for general protection	Cross-jurisdictional regulations	Context-dependent sensitivity levels

Table 3: Multi-Class Sensitivity Categories [7, 8]

5. Model Evaluation and Performance Metrics

Tough checking of classification models makes certain hitting production-quality bars for accuracy, reliability, and fairness across different data flavors and company contexts. Picking and reading suitable metrics needs grasping the particular traits of data classification jobs, especially the class tilt typically sitting in company data catalogs, where touchy columns make up a minority of total database fields. Classification model checking stretches past simple accuracy measures to cover multiple metrics that grab different sides of model performance, touching operational rollout and business value creation [9].

Precision metrics figure out what chunk of columns tagged as touchy truly hold touchy material, directly hitting operational smoothness by settling how many false alarms create needless data blocks, access denials, and compliance weight. Sharp precision cuts down the load on data users who must navigate access controls and approval workflows, keeping company nimbleness while protecting genuinely sensitive material. The business hit of false alarms packs in delayed analytics projects, slowed data-driven decision speed, and user annoyance that may spark workarounds dodging intended security controls. Companies must set precision targets that mirror operational comfort for false alarms, typically needing sharper precision in spots where data access blocks seriously hamper business flows [9].

Recall metrics measure what chunk of actual touchy columns get correctly flagged by the classification model, directly settling how well data protection programs and compliance stances work. Membership inference attacks and other privacy breaches show that missing even small batches of touchy data bits can build substantial security and compliance dangers [8]. Bad actors exploit unprotected material to trash individual privacy or grab unauthorized system access. Each miss stands for a potential compliance violation, with regulatory enforcement moves increasingly hitting companies that fail to spot and properly protect personal data. The financial and reputation hits of unspotted touchy data stretch past immediate breach costs to pack in regulatory fines, lawsuit expenses, and long-term brand damage, hitting customer trust and business ties [1].

F1-score serves up a harmonic mean of precision and recall, dishing out a single metric that grabs the balance between these competing goals, especially valuable when stacking different models or tuning classification thresholds. The metric really matters for lopsided datasets where accuracy alone can fool, since a simple sorter that tags all columns as non-touchy might hit high accuracy while totally bombing at spotting any touchy data [9]. Deep analysis of handling lopsided datasets shows that solid classification needs careful attention to class spread, sampling game plans, and checking metrics that properly weight minority class performance [10]. Tricks like synthetic minority oversampling, classweighted loss functions, and threshold tweaking let models hit a better balance between precision and recall on lopsided data.

Confusion matrix analysis serves up detailed insights into classification patterns by tallying true positives, true negatives, false positives, and false negatives across all sensitivity buckets in multi-class situations. This detailed breakdown shows systematic error patterns like frequent confusion between linked buckets like financial PII and general business identifiers, or wrong classification of edge cases like partial dates or business email addresses using off-brand domains [10]. Looking at confusion patterns guides targeted improvements through focused training data gathering, feature building boosts, and model architecture tweaks, tackling particular weaknesses. The loop process of analyzing confusion matrices, spotting error patterns, rolling out targeted improvements, and re-checking performance drives ongoing model boost and adjustment to shifting data landscapes.

Companies rolling out classification models must set up comprehensive checking frameworks that judge performance across multiple dimensions beyond total accuracy metrics. Field-specific checking needs testing on representative samples from each data source, business unit, and regulatory context touching company operations. Time validation makes certain models keep performance as data schemas shift, new systems get added, and naming habits change over time. Fairness checking looks at whether classification accuracy swings systematically across different data fields or sensitivity buckets, potentially flagging training data gaps or model biases needing fixes. Mixing number metrics with quality error analysis lays the groundwork for trustworthy classification systems that reliably protect touchy material while backing legit data uses.

Metric	Purpose	Impact on Operations	Optimization Strategy
Precision	Measures the accuracy of positive predictions	Reduces false positives and unnecessary restrictions	Threshold tuning, feature refinement
Recall	Captures the ability to identify all sensitive data	Minimizes compliance risks from missed data	Class weighting, synthetic oversampling
F1-Score	Balances precision and recall	Provides a single performance	Threshold calibration for
		indicator	imbalanced data

Confusion Matrix	Details classification patterns across categories	Reveals systematic errors for targeted improvement	Error pattern analysis, focused training

Table 4: Model Evaluation Metrics [9, 10]

Conclusion

The machine learning and artificial intelligence data sensitivity classification framework provides a disruptive solution to important flaws in enterprise data governance. Organizations are able to automatically detect and classify sensitive information on a scale that would have been difficult to do manually, by combining the dimensions of multiple feature extraction with advanced natural language processing capabilities. Multi-class architecture is flexible in its ability to accommodate the different definitions of regulations found in healthcare, financial, and privacy regulations, and provides extensibility to organization-specific needs. Transformer-based models are state-of-the-art accurate due to their deep semantic comprehension, whereas classical methods give computationally efficient alternatives that can be deployed in constrained resources. Thorough assessment systems involving precision, recall, F1-score, and confusion matrix analysis are used to ensure that classification systems can be used in production to high standards before actual use. Companies that have adopted this framework realize great gains, such as manual classification work, faster compliance readiness, automatic enforcement of policies, and reduced regulatory exposure due to unidentified sensitive data. The structured prediction results can be easily incorporated with existing information governance frameworks, access control systems, and data loss prevention systems, and developed into end-to-end automated processes. With the escalation of regulatory complexity and the exponential growth of data volumes, intelligent classification systems are emerging as the basic infrastructure that can support sustainable data governance at the enterprise scale. The capabilities will be improved further in the future due to the evolution towards federated learning, explainable artificial intelligence, active learning integration, and real-time streaming classification. The intersection of machine learning advancement and data governance requirements defines a novel model in which organizations will typically identify, categorize, and safeguard sensitive data and derive the utmost analytical advantage without violating privacy, security, and regulation needs amidst more intricate data environments.

References

[1] Abi Tyas Tunggal, "What is the Cost of a Data Breach in 2023?" UpGuard, 2025. [Online]. Available: https://www.upguard.com/blog/cost-of-data-breach

[2] Satori Cyber, "Data Classification Framework: What, Why and How." [Online]. Available: https://satoricyber.com/data-classification/data-classification-framework-what-why-and-how/

[3] Alon Halevy et al., "The Unreasonable Effectiveness of Data," IEEE Computer Society, 2009. [Online]. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf [4] Alation, "PHI-Safe AI: Designing Privacy-First Healthcare Workflows", 2025. [Online]. Available: https://www.alation.com/blog/phi-safe-ai-privacy-healthcare-workflows/

[5] Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[6] Sandra Onami, "Automated Sensitive Data Classification Using Privacy-Aware Machine Learning Models," ResearchGate Publication, 2025. [Online]. Available: https://www.researchgate.net/publication/397942157_Automated_Sensitive_Data_Classification_Using_Privacy-Aware_Machine_Learning_Models

[7] Kai Zhang and Xiaoqian Jiang, "Sensitive Data Detection with High-Throughput Machine Learning Models in Electrical Health Records," PMC Article PMC10785837, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10785837/

[8] Reza Shokri et al., "Membership Inference Attacks against Machine Learning Models," arXiv preprint arXiv:1610.05820, 2017. [Online]. Available: https://arxiv.org/abs/1610.05820

[9] Zeljko Vujovic, "Classification Model Evaluation Metrics," International Journal of Advanced Computer Science and Applications, 2021. [Online]. Available: https://www.researchgate.net/publication/352902406_Classification_Model_Evaluation_Metrics

[10] Babajide Adeoti, "A Comprehensive Analysis of Handling Imbalanced Datasets," International Journal of Advanced Trends in Computer Science and Engineering, 2025. [Online]. Available: https://www.researchgate.net/publication/392631260_A_Comprehensive_Analysis_of_Handling_Imbalanced_Dataset