future internet

MDPI

Article

UP-SDCG: A Method of Sensitive Data Classification for Collaborative Edge Computing in Financial Cloud Environment

Lijun Zu [1,2,3], Wenyu Qi [4], Hongyi Li [1], Xiaohua Men [3], Zhihui Lu [1,2,*], Jiawei Ye [1] and Liang Zhang [4]

[1] School of Computer Science, Fudan University, Shanghai 200433, China; zulijun@unionpay.com (L.Z.); 22210240089@m.fudan.edu.cn (H.L.); jwye@fudan.edu.cn (J.Y.)
[2] Institute of Financial Technology, Fudan University, Shanghai 200433, China
[3] China UnionPay Co., Ltd., Shanghai 201210, China; menxiaohua@unionpay.com
[4] Huawei Technologies Co., Ltd., Nanjing 210012, China; qiwenyu1@huawei.com (W.Q.); zhangliang1@huawei.com (L.Z.)
* Correspondence: lzh@fudan.edu.cn

Abstract: The digital transformation of banks has led to a paradigm shift, promoting the open sharing of data and services with third-party providers through APIs, SDKs, and other technological means. While data sharing brings personalized, convenient, and enriched services to users, it also introduces security risks, including sensitive data leakage and misuse, highlighting the importance of data classification and grading as the foundational pillar of security. This paper presents a cloud-edge collaborative banking data open application scenario, focusing on the critical need for an accurate and automated sensitive data classification and categorization method. The regulatory outpost module addresses this requirement, aiming to enhance the precision and efficiency of data classification. Firstly, regulatory policies impose strict requirements concerning data protection. Secondly, the sheer volume of business and the complexity of the work situation make it impractical to rely on manual experts, as they incur high labor costs and are unable to guarantee significant accuracy. Therefore, we propose a scheme UP-SDCG for automatically classifying and grading financially sensitive structured data. We developed a financial data hierarchical classification library. Additionally, we employed library augmentation technology and implemented a synonym discrimination model. We conducted an experimental analysis using simulation datasets, where UP-SDCG achieved precision surpassing 95%, outperforming the other three comparison models. Moreover, we performed real-world testing in financial institutions, achieving good detection results in customer data, supervision, and additional in personally sensitive information, aligning with application goals. Our ongoing work will extend the model's capabilities to encompass unstructured data classification and grading, broadening the scope of application.

Keywords: sensitive data; classification and grading; augmentation; synonym mining; financial scenarios

Citation: Zu, L.; Qi, W.; Li, H.; Men, X.; Lu, Z.; Ye, J.; Zhang, L. UP-SDCG: A Method of Sensitive Data Classification for Collaborative Edge Computing in Financial Cloud Environment. *Future Internet* **2024**, *16*, 102. https://doi.org/10.3390/fi16030102

Academic Editor: Paolo Bellavista

Received: 26 February 2024
Revised: 12 March 2024
Accepted: 14 March 2024
Published: 18 March 2024

1. Introduction

With the advent of the big data era, data have been recognized as essential production factors. To promote the data factor market, ensuring data security is a fundamental requirement. In this context, sensitive data pertain to information that, if disclosed or compromised, has the potential to inflict significant harm upon individuals or society. Sensitive data encompass personal privacy information, such as names, phone numbers, bank account numbers, ID numbers, addresses, passwords, email addresses, educational backgrounds, and medical records. Additionally, this includes enterprise data that are not suitable for public disclosure, such as the company's operational details, IP address lists, and network structure.

Effectively addressing the diverse and constantly evolving compliance requirements poses a formidable challenge. As the digital transformation gains momentum, numerous

countries' laws and regulations, coupled with security requirements stipulated by industry organizations (e.g., PCI DSS [1], SOX [2], HIPAA, GDPR [3], CCPA [4], etc.), underscore the importance of identifying and classifying sensitive data as the initial step in data protection. Enterprises are confronted with the task of streamlining their compliance workflows by leveraging simplified technology environments and pre-built templates. This necessitates understanding the precise locations of their data and determining whether additional safeguards are necessary. It also involves identifying both structured and unstructured sensitive data, both locally and in the cloud, that fall under regulatory scrutiny. Subsequently, these data must be categorized and cataloged for ongoing vulnerability monitoring.

In recent years, the financial industry has witnessed a rapid acceleration of the open banking model, where data applications are shared between banks and third-party service providers. More than 30 countries and regions worldwide have already adopted or are in the process of adopting this model [5]. Open banking offers numerous advantages, including enhanced customer experiences, the creation of new revenue streams, and the establishment of sustainable service models in markets with limited access to traditional banking services [5]. However, open banking also presents significant challenges, particularly concerning data security. The shared data encompass user identity information, financial transaction details, property, and other sensitive information. This extensive data sharing deepens the risk of data leakage and misuse [6].

To enhance the security of open banking data, we propose a sensitive data processing technique in a cloud-edge collaborative environment, as depicted in Figure 1. Firstly, financial institutions in the central cloud of a bank need to conduct a comprehensive assessment of their data assets to create a visual map of sensitive data before sharing with external parties. Secondly, the data application side (third-party organizations) deploys a regulatory outpost on the edge to ensure the security and compliance of open banking data. The Regulatory Sentinel is an independent software system designed to monitor every step of data operations performed by the application side, including storage, retrieval, and sharing. It also incorporates sensitive data identification, anonymization, watermarking, and records all user data operations for log auditing, leakage detection, data flow mapping, and situational awareness of data security [7].

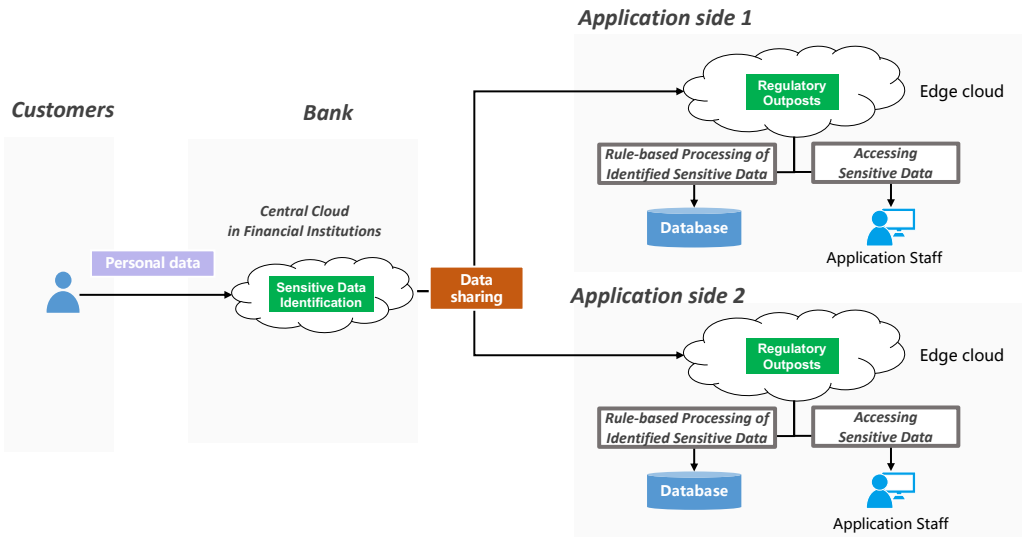

Figure 1. Cloud-Edge collaborative framework for sensitive data processing.

From the description of the regulatory outpost, it is evident that it deeply integrates into the data processing workflow of the application side, leveraging the characteristics of edge-based data processing. To avoid compromising the overall data processing experience

and incurring significant costs for the application side, the deployment of the Regulatory Sentinel should meet the following requirements:

1. Elastic scalability of resources: As data processing by the application side requires computational resources, which fluctuate with varying data volumes, the deployment should allow for elastic scalability of resources to minimize investment costs for the application side.
2. Low bandwidth utilization cost and reduced data processing latency: The data traffic accessed by the application side needs to pass through the regulatory outpost. It is crucial to ensure low bandwidth utilization costs and reduced data processing latency to minimize any impact on the application side's user experience.
3. Ensuring data compliance: In the context of open banking, the application side tends to locally store open banking data, necessitating compliance checks on these data to prevent potential leaks. As shown in Figure 2, a way is given for the application side to perform operations such as data desensitization and watermarking locally to enhance data security, in which data classification and grading is the basis.

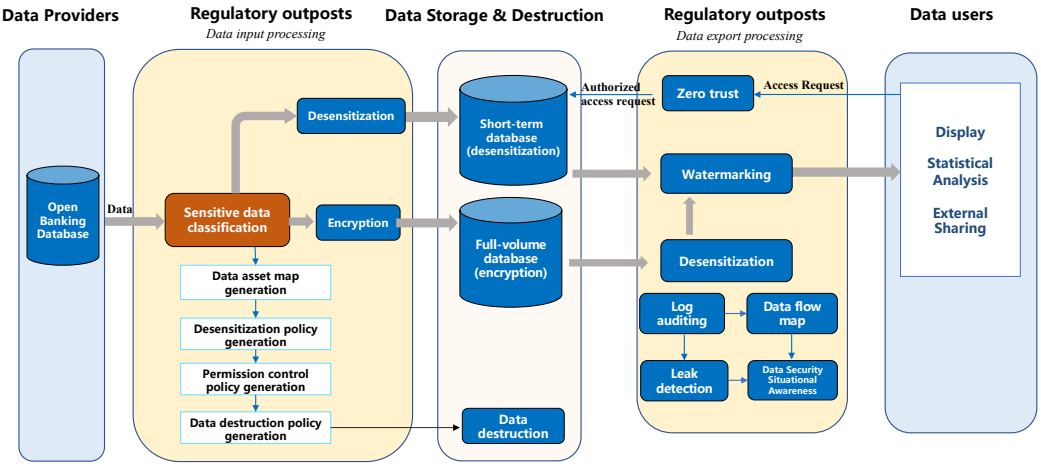

Figure 2. Data processing workflow in regulatory outposts within edge cloud scenarios. We conducted a comprehensive study on various security issues within open banking and proposed a data security framework. In this paper, our primary focus is on the issue of classifying sensitive data. Security measures such as watermarking are addressed in other works [7].

Hence, the automated classification and grading of sensitive data in the financial sector are garnering increasing attention. Firstly, financial institutions should conduct a comprehensive assessment of their own data landscape to achieve a visualized map of sensitive data assets before engaging in data sharing. Secondly, for third-party service providers collaborating with financial institutions, it is imperative that they enhance their data security management capabilities in accordance with government regulatory requirements and contractual agreements with financial institutions, which include encrypting sensitive data during storage or implementing data anonymization techniques, with the prerequisite being the prompt identification and classification of sensitive data transmitted during the collaboration process.

From the aforementioned scenario, the automated classification and categorization of sensitive data in the financial domain is a fundamental capability of the technology platform. Currently, the financial industry employs two primary methods for data classification and grading. One involves manual classification, which spans multiple departments, leading to a lengthy and inefficient process, and it lacks reusability, posing limitations on its scalability and adaptability. Another relies on automated classification and grading based on pattern matching, utilizing internally constructed data dictionaries. However, this approach suffers from low accuracy rates, especially when dealing with incomplete data dictionaries.

Building upon the aforementioned challenges, we present a data classification and grading framework to the financial industry which adheres to the relevant industry standards. Our framework encompasses both structured and unstructured data classification and grading. For structured data, we introduce a novel sensitive data classification and grading algorithm named UP-SDCG, leveraging self-enrichment and broadening techniques. Additionally, we enhance the financial data hierarchical classification by employing an augmentation model to expand keywords and lexicons, which significantly boosts the accuracy and recall of data classification and grading. Furthermore, we incorporate a synonym discrimination model to further expand the keywords and dictionaries in the industry data hierarchical classification library, resulting in improved accuracy and recall of data classification and grading. In our future work, we aim to further develop a scheme for classifying and grading unstructured sensitive data. This scheme will also support the coarse-grained classification of document data containing sensitive information. Additionally, we will propose a fine-grained classification approach to identify the types of sensitive data and their corresponding levels within the document.

Our research makes the following contributions:

- We propose a financial data classification and grading framework and a self-enlarging structured sensitive data classification and grading algorithm named UP-SDCG, with a synonym discrimination model innovatively introduced to further expand keywords and lexicons.
- Testing on real-world financial industry data, UP-SDCG outperforms existing public cloud algorithms in terms of accuracy and recall for sensitive data classification and grading.
- We further propose unstructured sensitive data classification and grading design scheme and scenario analysis.

2. Related Work

When it comes to data classification and grading, distinct approaches are employed to classify and grade various data structures. Data can be categorized into two main types based on their structure: structured and unstructured. Structured data are typically stored within databases, encompassing data types and field designations. This structured nature aids in effective data classification and grading, demanding meticulous categorization. Moreover, this process necessitates classifying and grading outcomes for individual columns. Conversely, unstructured data commonly appear in formats such as logs and documents, encompassing contextual semantics. Leveraging Natural Language Processing (NLP) methods facilitate semantic analysis, unveiling concealed sensitive information within the document. Within this context, the classification granularity can vary between broad and detailed. Broad classification involves furnishing classification results for the entire document, while fine-grained classification mandates identifying specific sensitive data types contained within the document alongside their corresponding levels.

2.1. Structured Sensitive Data Classification

Guan X. et al. [8] conducted a comprehensive investigation into the classification approach for structured sensitive data in the realm of electric power. They introduced a hierarchical identification technique, which initially identifies attributes within the database as sensitive data and subsequently categorizes them based on the specific characteristics of the sensitive information. Furthermore, different levels of sensitivity are assigned to these attributes in accordance with the varying permissions of the involved users. On the other hand, Rajkamal M. et al. [9] focused on safeguarding data stored in the cloud by extracting sensitive data components and post-encryption using Attribute-Based Encryption (ABE), and proposed a classification technique based on fuzzy rule analysis to effectively categorize attributes within structured data. In the healthcare domain, Ray S. et al. [10] combined domain experts and expert systems to assign sensitivity scores to attributes, enabling the

identification of sensitive data through techniques such as random sampling and multiple scanning, eliminating the need for data cleansing before identification.

However, the standalone accuracy of rule-based sensitive data classification methods presents limitations due to the dynamic nature of industry attributes and linguistic context, leading to instances where the algorithm may miss identifying the sensitive data. Mouza C. et al. [11] devised a strategy that involves semantically designating which concepts constitute sensitive information, thereby ascertaining sensitive content within structured data. Subsequently, the attributes in the database that semantically correspond to these concepts are retrieved. While this algorithm exhibits robust performance on smaller datasets, its scalability to larger datasets is a challenge. In response to the limitations of individual sensitive attribute detection, Tong Y. [12] introduced correlation rules to identify interconnected sensitive attributes. Similarly, Xiao Y. [13] proposed determining the correlation among sensitive attributes in structured data through a multidimensional bucket grouping technique, which enables the establishment of sensitive categories and levels based on attribute correlations.

Chong P. [14] employed machine learning techniques, including Bert models and regular expressions, for real-time active identification, classification, and validation of sensitive data. Similarly, Silva P. [15] harnessed NLP tools (NLTK, Stanford, and CoreNLP) to identify and validate personally identifiable information within datasets. Recent advancements have embraced deep learning-based NER models, showcasing their potential in automatic feature discovery for enhanced classification or detection [16]. Furthermore, Park J. et al [17] introduced NER techniques for structured data, constructing the Text Generation Module (TG Module) and Named Entity Recognition Module (NER Module) to generate sentences and recognize entities, respectively. While the application of AI models has indeed enhanced the accuracy of sensitive data recognition to a certain extent, the models often lack domain-specific knowledge at their inception. For instance, they may overlook the recognition of synonyms, leading to suboptimal performance in real-world engineering applications.

2.2. Unstructured Sensitive Data Classification

Jiang H. et al. [18] explored the use of text categorization methods, employing TF-IDF for feature extraction and initially evaluated Bayesian, KNN, and SVM classifiers for the classification of medically sensitive data. Adam Považane [19] employed document classification based on data confidentiality, comparing the performance of commonly used text classification algorithms across resume, legal document, and court report datasets. Notably, both Huimin Jiang's and Považane's studies limited test data classification to binary sensitive and non-sensitive categories. In contrast, Yang R. et al. [20] presented a sophisticated label distribution learning classification approach that aimed to categorize power data into six main categories and twenty-three subcategories but lacked specific experimental outcomes. Additionally, Gambarelli G. et al. [21] focused on personal information in their study of sensitive data. Their model consisted of three stages: SPeDaC1 for sentence classification as sensitive or non-sensitive, SPeDaC2 for multi-class sentence categorization, and SPeDaC3 for detailed labeling with 61 distinct personal data categories. It is important to highlight that empirical findings indicated reduced effectiveness in the model's fine-grained classification performance.

Dias M. et al. [22] endeavored to extract and categorize unstructured Portuguese text containing sensitive data. They constructed a named entity recognition module to identify sensitive information, such as personal names, locations, emails, and credit card numbers, within the text. In contrast, García-Pablos A. [23] introduced a deep learning model, BERT, to identify and categorize sensitive data in Spanish clinical text, aiming to recognize various types of sensitive information, including dates, hospital names, ages, times, doctors, genders, kinships, locations, patients, and occupations. However, there exists potential for further refinement and enhancement of the observed experimental outcomes.

2.3. Data Classification and Grading Framework

In the realm of data classification and grading, scholars typically commence their endeavors by establishing policies to ensure data compliance, which serves as a foundational step in constructing programs and frameworks for data classification and grading. For instance, Aldeco-Perez et al. introduced a compliance analysis framework based on data source and data usage, aligned with the UK Data Protection Act [24,25]. Their approach focuses on averting the misuse of personal sensitive data and evaluating the propriety of its utilization. However, they overlooked the potential risk stemming from the exposure of personal sensitive data under unforeseen circumstances. Subsequently, Yang M presented the Gen-DT scheme [26], which leverages legal statutes to establish an external knowledge base. They employed a generalization-enhanced decision tree algorithm to categorize data into regulatory and non-regulatory types. Nonetheless, this scheme solely dichotomizes data without specifying the sensitive classification and corresponding levels within regulatory data, which poses an inconvenience for implementing distinct protective measures for varying levels of sensitive data. Addressing the challenge of information extraction from regulations, Elluri L constructed a knowledge graph by automatically extracting information from GDPR and PCI DSS [27,28]. Building upon this, Yang M. [29] introduced the GENONTO framework, which autonomously extracts data classification and grading information from enacted regulations to construct a knowledge base. These frameworks expedite the extraction of classification and grading data from regulatory guidelines, facilitating their application in our module following calibration. Expanding beyond compliance considerations, academics have introduced additional metrics to optimize classification and grading outcomes. For instance, Wang J. et al. [30] introduced data value evaluation indicators to enhance data grading results within the context of classification criteria. This optimization was assessed within the new energy automobile industry.

3. Methodology

3.1. Data Classification Framework

Our proposed framework for sensitive data classification and grading consists of four key modules, as illustrated in Figure 3. These modules are the preprocessing module, the classification and grading module, the result presentation module, and the comprehensive analysis module.

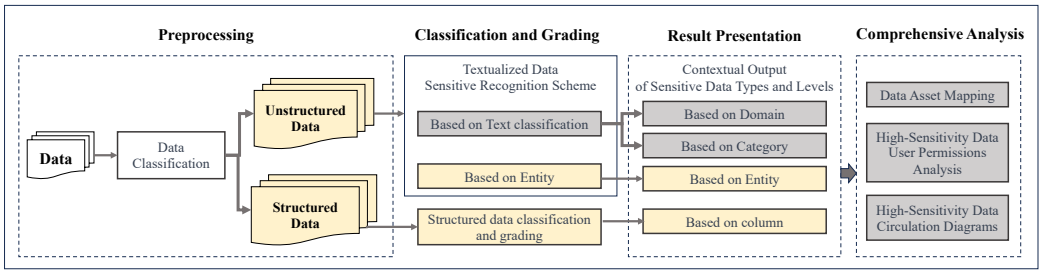

Figure 3. Data classification framework. The highlighted module in the figure is the main focus of this work.

The preprocessing module is responsible for classifying data into two categories: structured data and unstructured data. Depending on the data type, the classification and grading module applies different processes. For structured data, we designed a specialized model called the structured sensitive data classification and grading model (UP-SDCG). On the other hand, for unstructured data, we devised both coarse-grained and fine-grained data classification and grading schemes. The coarse-grained approach is based on text classification, while the fine-grained approach relies on entity recognition. The result presentation module displays unstructured data with details such as the Category, Type and Entity Type of the text, presented with granularity ranging from coarse to fine. For

structured data, the module inputs the classification and grading results for each column. Finally, in the comprehensive analysis module, we utilized the obtained classification and grading results to build data asset maps and other capabilities that provide users with a deeper understanding of the data landscape.

In the following sections, we present our proposed model, UP-SDCG, which focuses on the classification and grading of structured sensitive data.

3.2. Structured Data Classification Framework

Figure 4 illustrates the components of the structured data classification and grading framework, comprising three main modules: the hierarchical classification library building module, the keyword augmentation module, and the data classification and grading module. The hierarchical classification library building module is responsible for building industry-specific data classification and grading libraries, which are designed based on industry compliance standards. The keyword augmentation module leverages NLP technology to expand the keywords present in the industry data hierarchical classification library. Additionally, it trains the synonym discrimination model to enhance the library's capabilities. Lastly, the data classification and grading module utilizes the keywords, rules, dictionaries, and synonymous discriminative models from the hierarchical classification library. These components collectively enable the module to classify and grade structured data. The resulting output includes sensitive data types and their respective grades organized by columns.

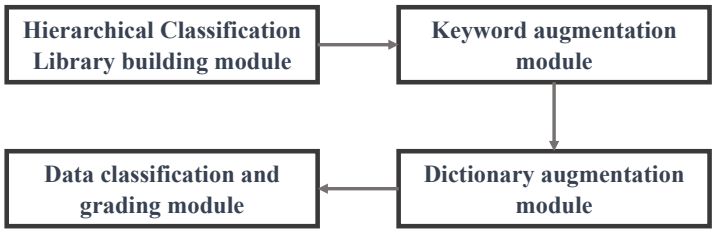

Figure 4. Structured data classification framework.

3.2.1. Library Building Module

To construct UP-SDCG's data classification hierarchy for the financial industry, we followed the guidelines outlined in the Financial Data Security Classification Guidelines (JR/T 0197-2022) [31]. This library encompasses the standard data commonly found in financial institutions, which can be categorized into four Level 1 subcategories, thirteen Level 2 subcategories, seventy-one Level 3 subcategories, and two hundred and seventy-nine Level 4 subcategories.

To extract entity names from the content, including names, genders, nationalities, and so on, we utilized pattern-matching technology. Subsequently, the sensitivity level is based on the identified entity. These entities are categorized into three groups based on expert knowledge and experience: strong rule entities, weak rule entities, and irregular entities.

- Strong Rule Entities: These entities are characterized by explicit and well-defined rules, resulting in minimal recognition errors, including Chinese ID numbers and Chinese cell phone numbers.
- Weak Rule Entities: These entities, including passwords and balances, exhibit some identifiable patterns, but regular expressions alone cannot guarantee complete matching.
- Irregular Entities: Unlike strong and weak rule entities, irregular entities lack discernible patterns or rules, making their identification particularly challenging.

We employed distinct recognition methods tailored to various entity types, as illustrated in Table 1.

Table 1. Entity recognition methods for structured data.

Entity Type	Structured Recognition Method
Strong Rule Entities	Regular Expression
Weak Rule Entities	Keyword + Regular Expression
Irregular Entities	Keyword + Dictionary

Based on the identification concepts outlined above, we developed the financial data hierarchical classification library for UP-SDCG. The structure of this library is presented in Table 2.

Table 2. Illustrative examples of UP-SDCG financial data hierarchical classification library.

Entity Name	Sensitivity Level	Entity Type	Keywords	Features
Name	3	Irregular Entity	Name	Name
Gender	3	Weak Rule Entities	Gender	Gender (Broad)
Gender	3	Strong Rule Entities	-	Gender (Narrow)
Nationality	3	Irregular Entity	Nationality	Country Name
ID Effective Date	3	Weak Rule Entities	ID Effective Date	Date
Enrollment Date	2	Weak Rule Entities	Enrollment Date	Date
Personal Income	3	Weak Rule Entities	Personal Income	Amount
Deposit	2	Weak Rule Entities	Deposit	Amount

We constructed the feature library by extracting content characteristics of entities, including information such as birthdays, the effective date of documents, the expiration date of documents, the date of enrollment, and other entities represented in date format. Similarly, personal income, deposit, credit card cash withdrawal amount, product amount, and other entities are represented in amount format. Additionally, we categorized presentation forms such as name, gender, country, date, and amount to form the comprehensive "Feature Library". This Feature Library comprises three distinct modules: Feature Name, Regular Expression, and Dictionary. The Feature Name within the Feature Library is associated with the features found in the Financial Data Hierarchical Classification Library. For further clarity, please refer to the structure of the Feature Library presented in Table 3.

Table 3. Illustrative examples of UP-SDCG Features Library.

Feature Name	Regular Expression	Dictionary
Name	-	Chinese Name
Gender (Broad)	$Male \| Female \| 0 \| 1 \| 2$	-
Gender (Narrow)	$Male \| Female$	-
Country Name	-	Country Name
Date	$\backslash d\{4\}year(1[0-2] \| [1-9] \| 0[1-9])day$	
	$\backslash d\{4\}(1[0-2] \| [1-9] \| 0[1-9])day$	-
Amount	$\hat{}(-\|\backslash+)?([1-9]\backslash d\{0,9\} \| 0)(\backslash.\backslash d\{1,10\})?$	-

3.2.2. Keyword Augmentation Module

In the design of the identification scheme, we observed that entity identification heavily relies on keywords. However, the lack of uniformity in data dictionaries across different departments and enterprises, as well as the reliance on manual labeling, presented challenges in achieving comprehensive keyword coverage. For instance, when referring to the income situation of an entity, keywords such as "monthly salary", "salary," "wage," "income," "treatment," and "remuneration" may be involved. Therefore, we proposed a keyword augmentation framework with a synonym discrimination model.

- Keyword Augmentation Framework

Keyword augmentation relies on the fundamental concept of synonym mining, which can be broadly categorized into two main types: knowledge-based augmentation and pattern-based augmentation.

As illustrated in Figure 5, knowledge-based augmentation primarily relies on four types of knowledge bases:

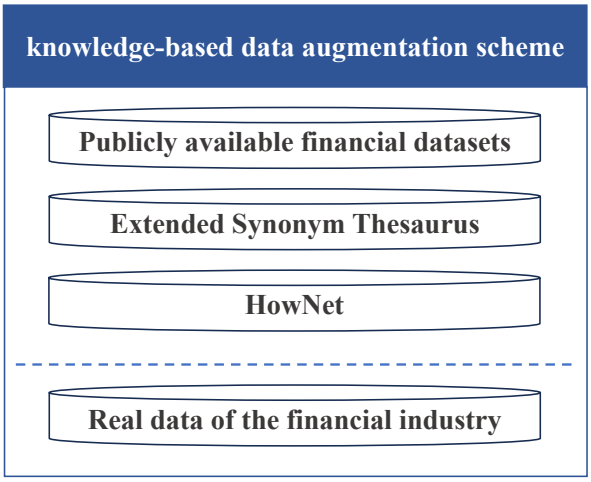

Figure 5. Knowledge-based keyword augmentation.

(1) Publicly available financial datasets: We used publicly accessible financial structured data to accumulate keywords. For example, we extracted statistics provided by the China Banking and Insurance Regulatory Commission;
(2) Extended Synonym Thesaurus: Considering the uniqueness of Chinese synonyms, we employed the Synonym Thesaurus [32] compiled by Mei J. et al. [33] at the Harbin Institute of Technology Information Retrieval Laboratory as the foundation to construct the Extended Synonym Thesaurus. This extended version encompasses nearly 70,000 entries organized in a hierarchical tree-like structure, utilizing a five-level encoding pattern to classify the entries into five tiers: major category, intermediate category, minor category, word group, and atomic word group. Each atomic word group includes one or more synonymous words;
(3) HowNet [34]: KnowNet utilizes tree-like sense-principal graphs and net-like sense-principal graphs to describe lexical properties;
(4) Real data of the financial industry: We incorporated real information from the financial industry, specifically the banking industry interface.

The central concept behind pattern-based keyword augmentation lies in bootstrapping. Bootstrapping is a statistical estimation method that involves inferring the distributional properties of the aggregate by resampling the observed information. The idea of implementing relationship extraction based on semi-supervised learning bootstrapping methods was proposed in Snowball [35]. We introduced this approach to the domain of keyword synonym mining, which comprises the following four sub-steps, as shown in Figure 6.

(1) Preparing the seed word set: This step involves collecting a set of high-quality alias word pairs for the current keyword;
(2) Mining the occurrence patterns: We analyzed the occurrence patterns of both the native names and aliases in a corpus constructed from Wikipedia and the Baidu Encyclopedia. These patterns encompass instances like "X, also known as Y." Furthermore, we utilized the seed word set to facilitate the identification of these patterns;
(3) Generating pattern sets: Based on the identified occurrence patterns, we generated sets of patterns that can be used for further analysis;

(4) Mining synonym pairs: Using the pattern sets, we extracted pairs of synonyms from the corpus. This step expands the range of synonymous terms associated with the designated keyword and facilitates a more comprehensive understanding of its semantic variations.

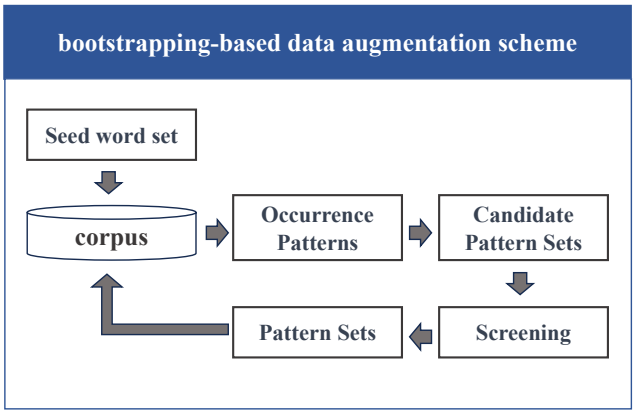

Figure 6. Pattern-based keyword augmentation.

- Synonym Discrimination Model

We developed a synonymy discrimination model classifier to determine whether a word can be added to a certain keyword collection. The construction process is as follows:

(1) We extracted keyword sets $\{S_1, S_2, \ldots, S_n\}$ from the existing UP-SDCG Financial Data Hierarchical Classification Library. Each keyword set S_i consists of several words with similar meanings;
(2) For each keyword set S_i, we employed knowledge-based and pattern-based keyword expansion techniques to extract the top five similarity words $\{t_{i1}, t_{i2}, \ldots, t_{i5}\}$. These similarity words are used to construct the keyword candidate set;
(3) We labeled the candidate words in the keyword candidate set. Words belonging to this keyword set were labeled as 1, while those not belonging to it were labeled as 0. Candidate keywords labeled as 1 were then expanded into the keyword set, resulting in the expanded keyword set. To train the classifier, we generated a collection of keyword training set-instance pairs from the pattern-based augmented keyword set. For each keyword set S_i, we randomly retained an instance $t_{pos} \in S_i$ and constructed a positive set of instance samples (S_i, t_{pos}) where the label y_{pos} was 1. For each positive sample (S_i, t_{pos}), we generated a negative sample (S_i, t_{neg}) by randomly selecting a negative instance t_{neg} where the label y_{neg} was 0. Following the research [36] experimental analysis, for each positive instance sample, we constructed five negative instance samples as shown in Table 4.

Table 4. Training data and labeling.

Candidate Keywords	Keywords Set	Label
t_{11}	S_1	0
t_{12}	S_1	1
...	...	...
t_{i1}	S_i	1
t_{i2}	S_i	1
...	...	...
t_{n4}	S_n	0
t_{n5}	S_n	0

(4) Next, we constructed the keyword set-candidate word classifier. We set the keyword set as S_i, the candidate word as t_{ij}, and the corresponding label as y_{ij}. We followed work [36] for similar candidate word discrimination through scores. First, we used $q(*)$ to quantify the degree of set similarity:

$$q(S_i) = g(\sum_{i=1}^{m} f_1(x_i)) \tag{1}$$

where $S_i = t_1, t_2, \ldots, t_m$ was put into the embedding layer to obtain the embedding vector $f_1(S_i) = x_1, x_2, \ldots, x_m$ and after that the original score representation was obtained and $g(*)$ represents the post-transformer, we then used a fully connected neural network with three hidden layers to transform the obtained vectors into scores. Then, we computed the difference between the set S_i and the set $S_i \cup \{t_{ij}\}$, and transformed it into a probability to determine the similarity between t_{ij} and keyword in S_i:

$$P(t_{ij} \in S_i) = \phi(q(S_i \cup t_{ij}) - q(S_i)) \tag{2}$$

where $\phi(*)$ is the sigmoid function. The model was optimized by minimizing the loss function:

$$loss(t_{ij}) = \begin{cases} -\log(max(1 - P(t_{ij} \in S_i), \alpha)) & \text{if } y_{ij} = 0 \\ -\log(max(P(t_{ij} \in S_i), \alpha)) & \text{if } y_{ij} = 1 \end{cases} \tag{3}$$

$$loss = \sum_{i=1}^{n} \sum_{j=1}^{5} loss(t_{ij}) \tag{4}$$

where t_{ij} belongs to S_i while y_{ij} equals to 1, and y_{ij} is 0 when t_{ij} does not belong to S_i. We set the parameter α to 10^{-5} to prevent the loss function from yielding infinite values.

By employing the aforementioned method, we optimized using an Adam optimizer with an initial learning rate of 0.001 and set the dropout to 0.5.

3.2.3. Dictionary Augmentation Module

In the Financial Data Hierarchical Classification Library of UP-SDCG, the detection of certain entities using regular expressions presents challenges. To address this issue, we need to construct specific dictionaries for entities such as names, app names, car brands, and others. Unlike keywords, the words within these dictionaries do not have exact semantic matches but tend to appear within similar contextual structures. Leveraging this characteristic, we propose a word2vec-based augmentation scheme, illustrated in Figure 7, to enhance the detection capabilities.

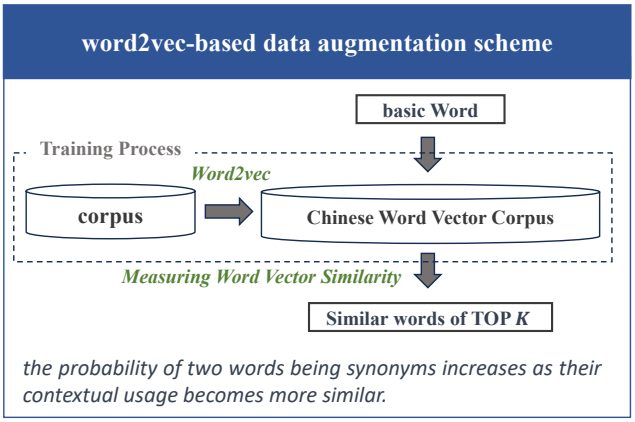

Figure 7. Word2vec-based dictionary augmentation scheme.

Word2vec [37] is a word embedding technique introduced by Google, which aims to represent abstract words as vectors in the real number domain. The method comprises two architectures: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts the current word based on its context, whereas Skip-gram predicts the context given the current word. To enhance training efficiency, Word2vec introduces two training algorithms: Hierarchical Softmax and Negative Sampling. Word2vec's ability to capture synonymy between words proves advantageous in dictionary construction. In our study, we utilized word2vec to train a Chinese word vector library specifically tailored to the financial domain. By combining publicly available Chinese word vectors from the industry, we constructed a dictionary using word vector similarity. The process is as follows:

(1) Utilize open-source pre-trained word vectors, such as Tencent AI Lab Embedding Corpus for Chinese Words and Phrases [38], Stanford GloVe Embeddings [39], fastText word vectors [40];
(2) Load the Embedding model with the selected pre-trained word vectors and fine-tune it using the financial corpus, which includes financial reports, financial news messages, etc.;
(3) Subsequently, extract similar words from the fine-tuned word vectors using cosine distance to calculate the distance between words and construct the dictionary.

3.2.4. Data Classification and Grading Module

In this module, we present the fundamental principle of quantifying information quantity. Specifically, in structured data, when column A and column B have an equal number of rows and pertain to the same type of sensitive data, the difference $\in$ in the information they provide falls within a certain range [41]. This can be formulated as follows:

$$|H(A) - H(B)| \leqslant \in \tag{5}$$

Here, we introduce $H(*)$ as a measure function to quantify the amount of information provided by each column of data.

$$H(X) = -\sum_{x \in A} P(x) \log P(x) \tag{6}$$

In Equation (6), $0 \leq P(x) \leq 1$, $\sum_{x \in A} P(x) = 1$ and $P(x)$ represent the probability of occurrence for each discrete piece of information.

To apply the basic principle of information quantity quantization, we computed the average amount of information provided by each subclass $\{S_1, S_2, \ldots, S_n\}$ in the data dictionary. This was achieved by defining the equation as follows:

$$S_i = k_{ij}{}_{i=1}^{m_1} \tag{7}$$

where k_{ij} represents a single keyword and m_i (where $m_i \geq 1$) denotes the number of keywords in subclass S_i.

To proceed, we identified columns in the dataset with k_{ij} as the field name and extracted 100 rows from each column to form $c_{ijk}{}_{k=1}^{r_i}$. Subsequently, we calculated the number of discrete information elements q and information entropy H in each class.

$$q_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \left(\frac{1}{r_i} \sum_{k=1}^{r_i} |c_{ijk}| \right) \tag{8}$$

$$H_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \left(\frac{1}{r_i} \sum_{k=1}^{r_i} H(c_{ijk}) \right) \tag{9}$$

The process of automatically classifying and grading structured data based on synonym discrimination and information quantification, as illustrated in Figure 8, involves the following steps:

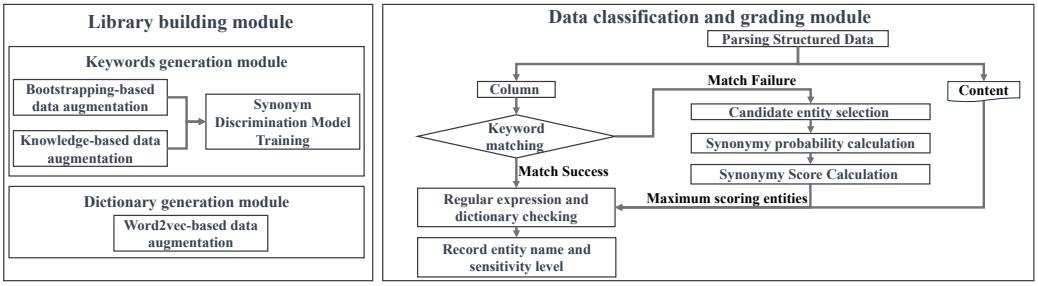

Figure 8. Structured data classification and grading process.

(1) Parsing: Structure the data into two parts: field names and data content;
(2) Field name identification: Utilize the keywords from the financial data hierarchical classification library to match the field names. If a corresponding field name is found, proceed to step 6; otherwise, move to step 3;
(3) Candidate entity selection: Randomly select 100 rows of data (denoted as d_i for $i = 1, 2, \ldots, 100$) and identify the data types, such as numeric value, English character, Chinese character, mixed character and date. Consider entities with the same data type from the Financial Data Hierarchical Classification Library as candidate entities;
(4) Synonym probability calculation: Apply the synonym discriminant model to determine if the field name is synonymous with the keyword set of the candidate entity. Input the candidate entity's keyword set S_i and the field name s into the synonym discriminant model, resulting in the probability $P(s \in S_i)$ that the field name belongs to the keyword set. Iteratively traverse all candidate entities to obtain:

$$\{P(s \in S_1), P(s \in S_2), \ldots, P(s \in S_i)\} \tag{10}$$

(5) Synonym score calculation: Calculate the number of discrete information q and information entropy $H(d)$ of the data:

$$Score = \theta_1 P(s \in S_i) - \theta_2 |q - q_i| - \theta_3 |H(d) - H_i| \tag{11}$$

$$\theta_1 + \theta_2 + \theta_3 = 1 \tag{12}$$

where θ_i represents the weight share of each index. The entity belonging to the keyword set with the highest score becomes the classification result, and step (6) is executed;
(6) Calibration: Perform stratified sampling of the corresponding content of the field name. Apply sensitive rules belonging to the keywords that match successfully in the feature library under its regular items and dictionaries for secondary detection of the sampling results. Recognition is considered successful if the matching rate exceeds the set threshold; otherwise, it is considered a recognition failure;
(7) Output: Output the corresponding entity name and sensitivity level from the Financial Data Hierarchical Classification Library if the recognition is successful. If the recognition fails, output NULL.

4. Experiments

The experiment was divided into two parts. Firstly, we organized a batch of simulation test data tailored to the data characteristics of financial institutions. Using this dataset, we compared the performance of our proposed algorithm with that of existing public algorithms. Secondly, we conducted a verification of the practical effectiveness of our algorithm using real business system data from financial institutions.

We present an automatic classification and grading program written in C++ (with 5000 lines of codes) and tested using financial data. The experiments were conducted on a Windows host with an Intel Core i7-8700 CPU @ 3.20 GHz 3.19 GHz processor. The synonym determination module in this paper was trained using Python 3.

4.1. Evaluation Metrics

This experiment evaluates the data classification and grading model using Precision, Recall, and F1-score. Precision, also known as recall, represents the probability of correct classification results among all the samples classified by the model, and it is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

where TP is the number of samples with correct classification and grading and FP is the number of samples with errors in classification or grading.

Recall represents the probability of correctly classifying and grading the samples that are actually required to be classified and graded from the original sample. It is expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

where FN represents the number of samples that are not classified and graded.

The F1-score considers both Precision and Recall, facilitating their simultaneous maximization and balance. The F1-score is mathematically represented as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{15}$$

4.2. Comparative Analysis

4.2.1. Datasets

The simulation dataset comprises three primary categories: personnel, projects, and contracts. Table 5 provides an overview of the experimental data, presenting relevant details for each category.

Table 5. Overview of simulation data.

Dataset	Row Number	Column Number	Sensitive Columns	Non-Sensitive Columns	Sensitive Type
Personnel Information	22,618	111	19	92	Name, gender, phone number, email address, company name
Project Information	23,208	301	31	270	Information about departments and personnel involved in the project
Contract Information	6351	37	15	22	Contract payment information

- Personnel information: The personnel dataset consists of 111 variables (columns) and 22,618 data points, encompassing details such as the employee's name, gender, work number, cell phone number, email, department, and position.
- Project information: The project dataset contains basic information about the bank's projects, comprising 301 variables and 23,208 data points. This dataset includes information pertaining to project personnel, departments, project budgets, and other relevant factors. It is noteworthy that the dataset contains a substantial amount of missing values.
- Contract information: The contract dataset has 37 variables and 6351 data points that relate to basic contract information as well as supplier information.

In the simulation dataset, the sensitive information in each column was identified through expert auditing, and the distribution of sensitive information in this dataset is illustrated in Figure 9. Among the various data columns, the contract data contained the highest proportion of sensitive information, accounting for over 40% of the dataset. On the other hand, the project data exhibited a relatively smaller percentage of sensitive data, mainly due to a significant number of missing values present in the data.

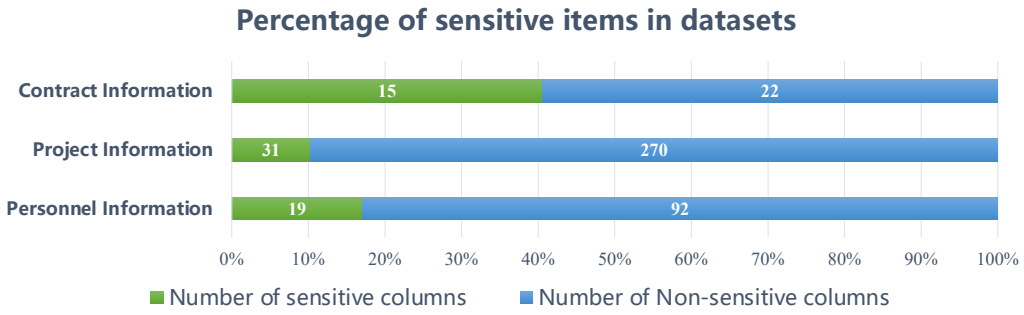

Figure 9. Statistical analysis of sensitive data distribution in simulation dataset.

4.2.2. Experimental Results

Using the test dataset, we conducted a comprehensive comparison of UP-SDCG with the existing sensitive data recognition models commonly employed in the industry. In the subsequent sections, we elaborate on the specifics of the comparison model and present the results of our experiments.

- DSC sensitive data identification model(Alibaba): Including 210 detection rules, the financial classification template in the DSC sensitive data identification model is constructed with reference to the industry standard Financial Data Security Data Security Classification Guide;
- DSGC Sensitive Data Identification Model (Tencent): Use the built-in general classification and grading standard template for identification, which contains 41 detection rules;
- GoDLP (Bytedance): ByteDance's open source tool for sensitive data identification in 2022, which can support structured data and unstructured data identification, with 36 built-in detection rules.

UP-SDCG exhibited exceptional accuracy, surpassing all other three comparison models with a remarkable accuracy rate of over 95% on all three datasets, as depicted in Figure 10. Notably, the DSC achieved high accuracy in recognizing personnel information, boasting a perfect 100% accuracy rate for both personnel information and project information. However, its performance in detecting contract dates was subpar, attributed to the complexity of contract data that often contain various types of date information, such as contract start and end dates. DSC's limitations lie in its inability to correctly classify the granularity of the date categories, resulting in insufficient delineation ability.

In contrast, our model demonstrated fine-grained category classification through the utilization of keyword augmentation techniques, leading to a significant improvement in recognition accuracy. By effectively recognizing and classifying sensitive data, including numerical information like employees' work numbers and identity IDs, our model outperformed DSGC, which has a high misclassification rate for such data. Furthermore, while GoDLP achieved a higher accuracy rate by adhering to stricter rules, it recognized fewer sensitive data instances.

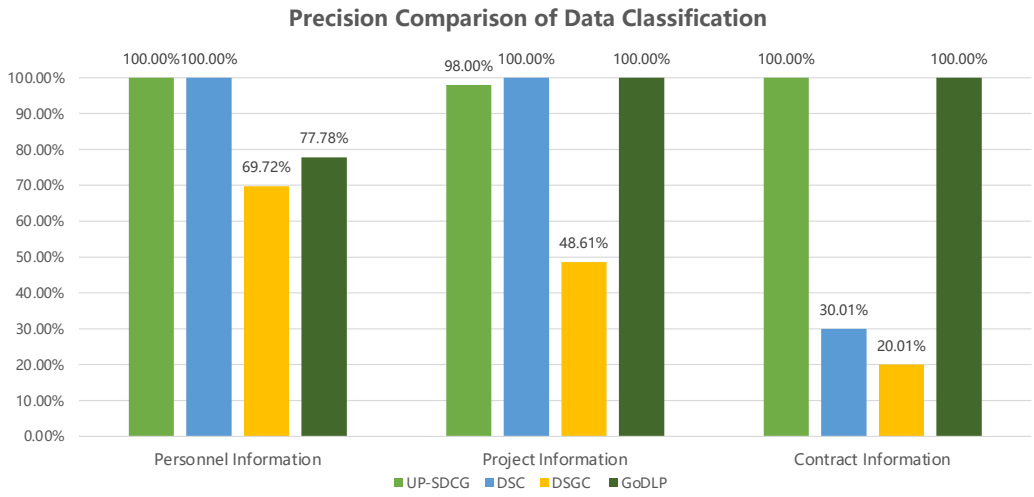

Figure 10. Comparison of precision in data grading and classification models.

UP-SDCG demonstrated a remarkable recall rate of over 94% across all test sets, resulting in fewer omissions, as depicted in Figure 11. By formulating more than 1100 detection rules based on industry standards, UP-SDCG covered a broader range of sensitive data compared to other three comparison models. As a result, its recall rate exhibited significant improvement. A comparison with models DSGC and GoDLP, which utilize generic sensitive data recognition templates revealed the limitations of current generic data classification and grading models in the financial domain. This highlights the crucial role played by domain-specific detection rules in achieving accurate recognition within the financial context.

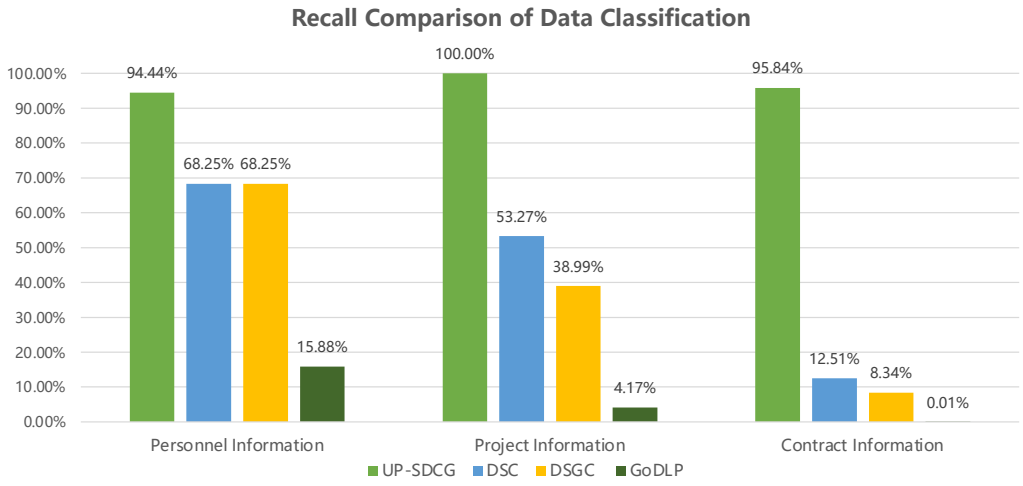

Figure 11. Comparison of recall in data grading and classification models.

By considering both false alarms and leakage cases, we demonstrate the superior performance of UP-SDCG over existing industry models, as illustrated in Figure 12. Specifically, when compared to DSC, which also leverages financial hierarchical classification template recognition and detection based on industry standards, our model achieves a lower leakage rate due to its comprehensive detection rules. Additionally, we incorporated keyword augmentation and expansion techniques, enabling fine-grained and accurate hierarchical classification, thus effectively mitigating false alarm situations. Furthermore, a comparison with generalized sensitive data hierarchical classification models, such as

DSGC and GoDLP, further underscores the advantages of our financial data recognition hierarchical classification rule base construction.

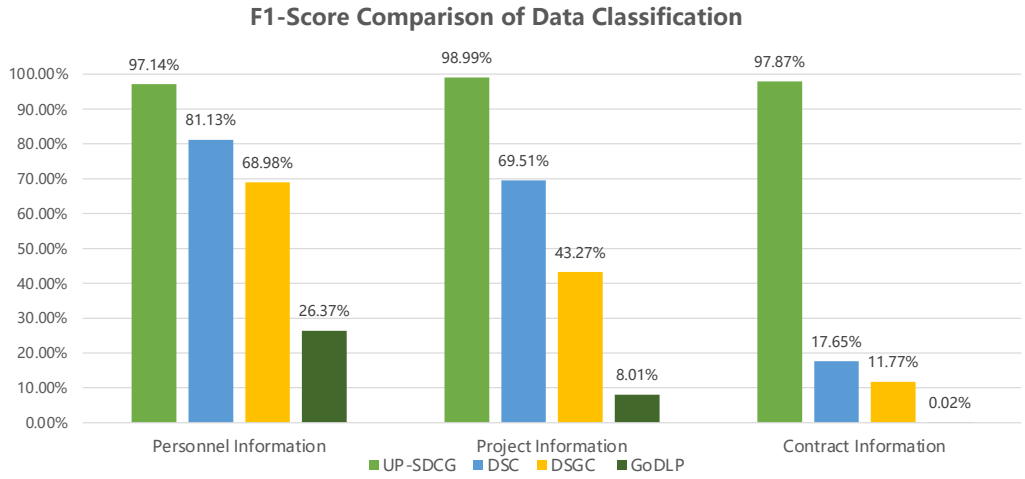

Figure 12. Comparison of F1-score in data grading and classification models.

The comprehensive results of this experiment are presented in Table 6, revealing significant advantages of our method over existing industry models across the three types of test data. Our approach excels in terms of Precision, Recall, and F1-score, which are the three key evaluation metrics used for performance assessment.

Table 6. Performance metrics of different models on various datasets.

Dataset	UP-SDCG			DSC			DSGC			GoDLP		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Personnel Information	100.00%	94.44%	97.14%	100.00%	68.25%	81.13%	69.72%	68.25%	68.98%	77.78%	15.88%	26.37%
Project Information	98.00%	100.00%	98.99%	100.00%	53.27%	69.51%	48.61%	38.99%	43.27%	100.00%	4.17%	8.01%
Contract Information	100.00%	95.84%	97.87%	30.01%	12.51%	17.65%	20.01%	8.34%	11.77%	100.00%	0.01%	0.02%

4.3. Practical Validation

4.3.1. Dataset

The dataset utilized in this study comprises real business data from financial institutions. The experiments were conducted within a secure inner loop environment. The dataset encompasses four major categories, namely customer information, service data, operation management, and financial supervision. A comprehensive overview of the experimental data is presented in Table 7.

Table 7. Overview of the business dataset.

Dataset	Row Number	Sensitive Columns	Non-Sensitive Columns	Sensitive Type
Customer Information	73	12	61	Personal information, such as name, certificate number, income, address, phone number,account password, etc.
Service Data	23,208	103	78	Loans, insurance, bonds, cross-border business, etc.
Operation Management	13	0	13	Personal and financial information
Financial Supervision	3	0	3	Risky assets and capital adequacy

- Customer information: The customer dataset comprises 73 variables, encompassing a wide range of data including customer names, ID numbers, income details, addresses, phone numbers, and account passwords.

- Service data: The service dataset, encompassing information on loans, insurance, bonds, and cross-border transactions, comprises a total of 103 variables.
- Operation management: The Operation management dataset has 13 variables related to personal information as well as company financial information.
- Financial supervision: The financial supervision dataset has three variables related to information on regulatory indicators.

4.3.2. Experimental Results

Our model demonstrated exceptional performance, achieving over 90% precision, recall, and F1-score across all four types of test data, as presented in Table 8. The false alarm and omission cases primarily arose from the following two factors:

- Ambiguous content representation: The data content pertaining to real-world business scenarios lacked clarity, which leads to certain omissions and false alarms in our analysis;
- Data quality challenges: In real business scenarios, we encounter issues such as typos, missing characters, and other irregularities, which contributed to certain omissions in our data processing.

Table 8. Experimental effects of classification and grading.

Dataset	Precison	Recall	F1-Score
Customer Information	100.00%	98.36%	99.17%
Service data	96.05%	93.59%	94.81%
Operation management	91.67%	92.31%	91.99%
Financial Supervision	100.00%	100.00%	100.00%

UP-SDCG demonstrated greater effectiveness in detecting customer information and service data, suggesting that it currently outperforms other models for identifying personal sensitive information. However, when it comes to financial data related to business operations and management, its performance slightly lagged due to the inherent uncertainty in the data structure. Nevertheless, UP-SDCG still achieved precision, recall, and F1-scores surpassing 90%.

4.4. Performance Analysis

We sought to understand the factors influencing the time consumption of UP-SDCG. To investigate this, we analyzed six key variables, specifically row count, column count, sensitive column count, non-sensitive column count, sensitive data percentage, and the computational time taken by the test dataset. These variables are detailed in Table 9.

Table 9. Time consumption examples of the UP-SDCG Model on partial datasets.

Dataset	Rows	Columns	Sensitive Columns	Non-Sensitive Columns	Sensitive Ratio	Time (s)
Personnel Information1	9	50	6	44	0.120	6.351
Personnel Information2	11,280	25	6	19	0.240	5.597
Personnel Information3	11,329	36	8	28	0.222	4.825
Contract Information1	4372	9	3	6	0.333	2.025
Contract Information2	1979	28	12	16	0.429	4.269
Project Information1	20,206	151	10	141	0.066	49.493
Project Information2	3002	150	28	122	0.187	29.860
Bank Data1.csv	48	132	111	21	0.841	6.675
...	...	...	...	...	...	...

Initially, we utilized the Pearson correlation coefficient to quantify the linear relationship between variables. The Pearson correlation coefficient is computed using the following formula:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{16}$$

The resulting correlation values for pairwise variables are computed and presented in Table 10.

Table 10. Pairwise variable correlations.

	Rows	Columns	Sensitive Columns	Non-Sensitive Columns	Sensitive Ratio	Time (s)
Rows	1.000	−0.305	−0.193	−0.235	0.576	−0.008
Columns	−0.305	1.000	0.537	0.834	−0.268	0.779
Sensitive Columns	−0.194	0.537	1.000	−0.017	0.429	−0.032
Non-sensitive Columns	−0.235	0.834	−0.017	1.000	−0.598	0.944
Sensitive Ratio	0.576	−0.268	0.429	−0.598	1.000	−0.421
Time(s)	−0.008	0.779	−0.032	0.944	−0.421	1.000

The strength of the correlation between variables can be determined by the magnitude of the correlation coefficient ρ. When $|\rho| > 0.8$, it signifies a strong correlation, while $0.5 \leq |\rho| < 0.8$ indicates a moderate correlation. For $|\rho|$ values falling within $0.3 \leq |\rho| < 0.5$, the correlation is considered weak, and if $|\rho| < 0.3$, the variables are essentially uncorrelated. Analyzing Table 10, we observe that model execution time exhibits a strong correlation with the number of non-sensitive columns, a moderate correlation with the number of data columns, a weak correlation with the percentage of data sensitivity, and a negligible correlation with the number of data rows and sensitive columns. Although both the count of non-sensitive columns and the number of data columns influence the model's runtime, their correlation coefficient stands at 0.83449, signifying a strong linear correlation. In this context, either one of these factors could be selected for analysis. However, it is important to note that the Pearson correlation coefficient solely addresses linear correlations between variables. Our comprehensive analysis is extended further in Figure 13.

Upon analyzing Figure 13, it became evident that a curvilinear relationship exists between the model's elapsed time and the data sensitivity ratio, represented by the equation $y = \frac{1}{x}$. Consequently, we performed the reciprocal of the sensitivity ratio to derive the column $\frac{1}{sensitivity\_ratio}$. Subsequently, we recalculated Pearson's correlation coefficient with the elapsed time, yielding the results presented in the updated Table 11:

Table 11. Model execution time and correlation coefficients with various variables.

	Rows	Columns	Sensitive Columns	Non-Sensitive Columns	Sensitive Ratio
Time (s)	−0.00756	0.77861	−0.03227	0.94370	−0.80537

$$Time = a + b \cdot insensitivity\_column + \frac{c}{sensitivity\_ratio} \tag{17}$$

Through fitting the execution time, we obtained the fitted equation:

$$Time = -0.578811 + 0.263136 \cdot insensitivity\_column + 0.419442/sensitivity\_ratio \tag{18}$$

The *R-squared* value of this fitted equation is 0.895, and the p-value is 3.02×10^{-48}, indicating a favorable fit.

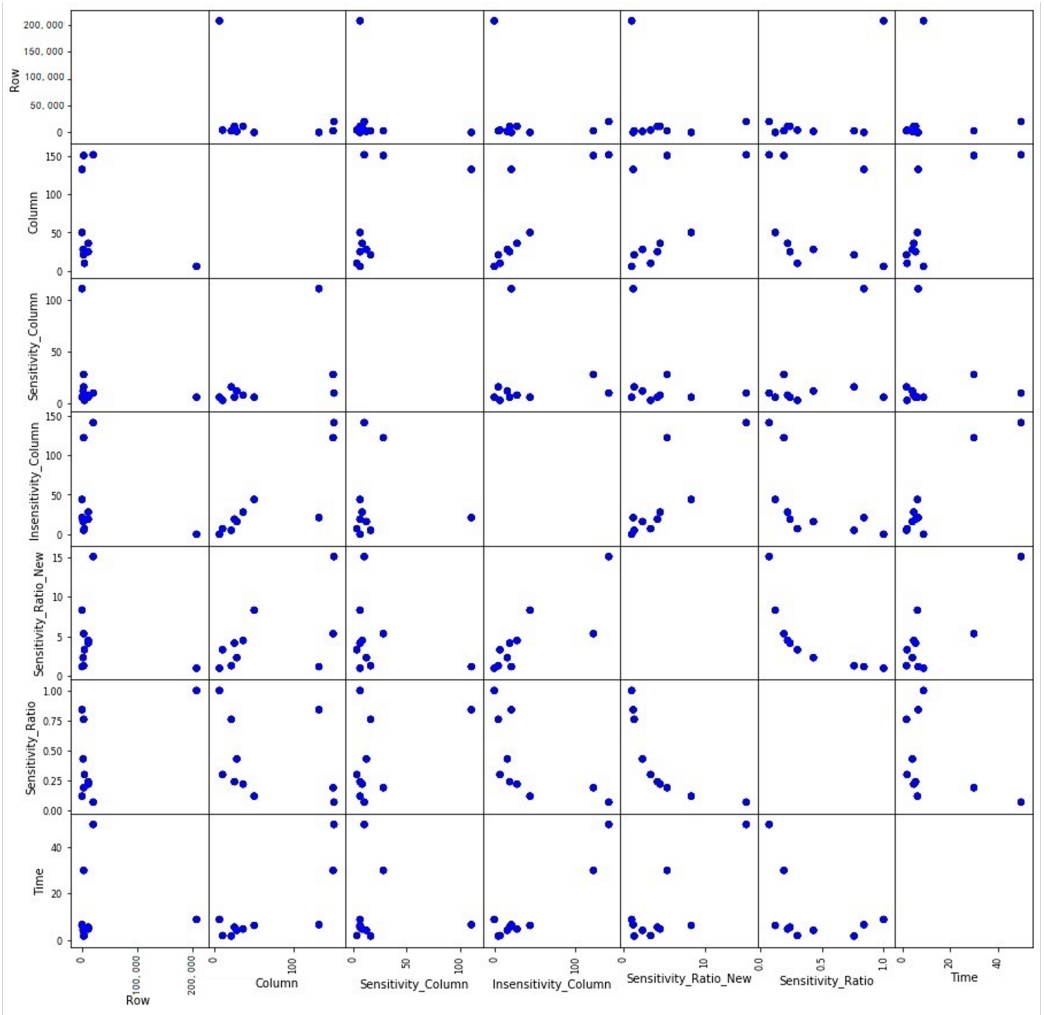

Figure 13. Scatter plots of pairwise variables.

5. Conclusions

Data security is an important basic condition for financial institutions to build a data-based ecology, and how to accurately identify massive data for classification and grading control has become a key issue. Under the overall framework of financial sensitive data classification and grading research work, we propose a self-enlarging and broadening financial sensitive data classification and grading method (UP-SDCG), which combines the traditional recognition technology with NLP technology, effectively solves the problem of low accuracy rate of the traditional recognition technology. The experimental results show that it has a significant advantage of effect compared with other publicly available platform algorithms, and also has been validated in real financial institutions. The results have also been verified in real financial institution business scenarios. Compared to existing classification and grading frameworks, our approach offers a finer granularity, enabling more precise implementation of protective measures tailored to various data types and levels, which significantly mitigates the risk of high-sensitivity data leakage. Our subsequent work will focus on the research of Unstructured Financial Data Classification and Graded Recognition (UP-UDCG), mainly realizing the two major functions of data classification and grading based on text classification and data classification and grading based on entities, and essential research methodology can be referred to in Appendix A. By deploying the sensitive data classification algorithm at the regulatory outposts, we aspire for our work to contribute to enhancing data security in open banking.

Author Contributions: Conceptualization, L.Z. (Lijun Zu); Methodology, L.Z. (Lijun Zu), W.Q. and H.L.; Software, W.Q.; Validation, H.L. and X.M.; Resources, L.Z. (Lijun Zu) and X.M.; Data curation, W.Q.; Writing—original draft, L.Z. (Lijun Zu), W.Q. and H.L.; Writing—review & editing, Z.L.; Supervision, Z.L., J.Y. and L.Z. (Liang Zhang). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Key Research and Development Program of China (Grant 2021YFC3300600).

Data Availability Statement: The data can be shared up on request. The data are not publicly available due to the sensitivity and confidentiality of financial industry data.

Conflicts of Interest: Author Xiaohua Men was employed by Unionpay Co., Ltd. Her position is engineer. Author Liang Zhang employed by Huawei Technologies Co., Ltd. His position is senior engineer. Author Wenyu Qi was employed by Huawei Technologies Co., Ltd. Her position is engineer. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Unstructured Data Classification and Grading Framework

The framework for unstructured data classification and grading comprises two modules: text-based data classification and grading, and entity-based data classification and grading. The text-based module determines the domain and type of the text, while the entity-based module identifies specific entities embedded within the text. As depicted in Figure A1, the framework provides information on the domain of the text, the involved text type (such as diplomas, CVs, insurance policies, etc.), and the sensitive entities present in the text (e.g., ID card numbers, email addresses, cell phone numbers, etc.).

FileName	MD5	Domain	Type	Sensitivity level	Level 1 Sensitive Entity Category	Level 2 Sensitive Entity Category	Level 3 Sensitive Entity Category	Level 4 Sensitive Entity Category
Diploma.pdf	4C1B1AF08F0A06EB	Academic	diploma	4	...	...	Email, cell phone, name	...
Transcript.pdf	9E3CBA4BD8F987CD	Academic		3	...	Province, city	...	...
AttendanceRecord.doc	84ABD7194DD3D07D	Academic		2	...	...	Email, cell phone, name	...
1213.pdf	ED33D2B8CC1BD034	Medical	policy	4	...	...	Email, cell phone, name	...

Entity Text classification

Figure A1. Example of classification and grading of unstructured data.

Appendix A.1. Data Classification Based on Text Classification

Various classification algorithms can be chosen depending on the specific context or scene. We provide an overview of common text classification algorithms, along with their respective applicable scenarios, advantages, and disadvantages, as shown in Table A1.

Table A1. Applications and Pros/Cons of Common Text Classification Algorithms.

Text Classification Algorithm	Suitable Scenarios	Advantages	Disadvantages
FastText	Large sample sizes, multiple categories, tasks with limited semantic understanding	Fast, low computational requirements	Limited semantic understanding
CNN	Tasks requiring some semantic understanding	Captures more, broader, and finer text features	Long training time
Self-Attention	Tasks requiring some semantic understanding	Captures more, broader, and finer text features, long-term dependencies within the text	Long training time
Traditional Machine Learning	Short texts (e.g., messages, microblogs, comments) with less than 150 words	Fast training	Unable to handle long texts
BERT	Limited labeled data scenarios	High accuracy	Long training and prediction time

Appendix A.2. Data Classification Based on Entity

Entity-based unstructured data classification and grading builds upon the principles of structured data classification and grading, employing keywords and patterns to detect sensitive entities. While structured data utilizes a "keyword+dictionary" approach for identifying irregular entities, this method is not suitable for unstructured data. Therefore, we adopt the Named Entity Recognition (NER) model to handle irregular entities. The specific identification process is illustrated in Figure A2.

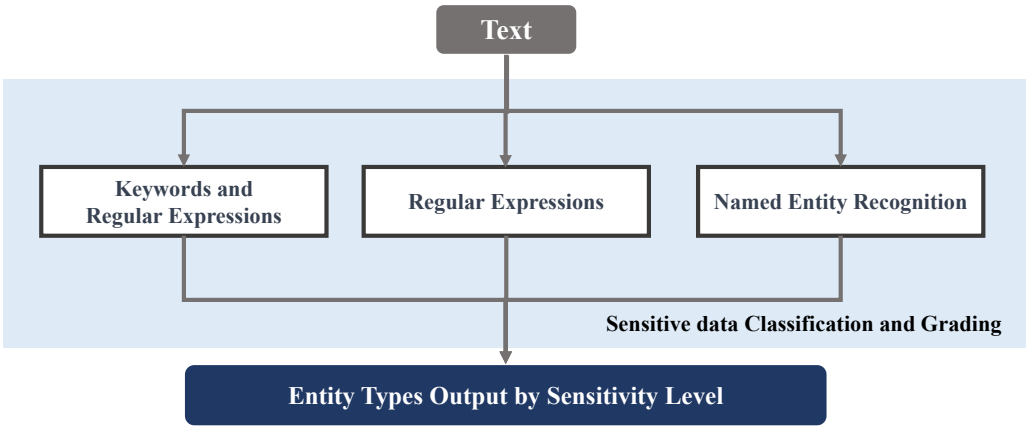

Figure A2. Unstructured Data Classification and Grading Process.

Likewise, the selection of NER models can be tailored to specific scenarios. We provide an overview of various NER models along with their respective applicable scenarios, highlighting their individual strengths and limitations, as shown in Table A2.

Table A2. Applicability and Pros/Cons of Common NER Models.

Model	Application Scenarios	Advantages	Disadvantages
BiLSTM + CRF	Large sample data, multiple label categories	Simple model structure, fast training speed	Moderate entity extraction performance
StructBert	Insufficient annotated data	Good entity extraction	Lower overall performance
StructBert + CRF	Small data scenarios	Good entity extraction performance	Lower overall performance

Appendix B. Unstructured Data Classification and Grading Framework

In the module for constructing hierarchical classification libraries, industry experts have the capability to create data hierarchical classification libraries that align with compliance standards and requirements. Table A3 presents the data security compliance standards applicable to China's core industries. This allows for a systematic and structured approach to organizing and managing data in accordance with industry-specific regulations.

Table A3. Data Security Compliance Standards in Key Chinese Industries.

Industry	Compliance Standard	Regulatory Authority
General	"Guidelines for Cybersecurity Standard Practice—Network Data Classification and Grading"	National Information Security Standardization Technical Committee
Industrial	"Guidelines for Industrial Data Classification and Grading"	Ministry of Industry and Information Technology (MIIT) of China

Table A3. *Cont.*

Industry	Compliance Standard	Regulatory Authority
Financial	"Financial Data Security—Data Classification and Grading Guidelines"	People's Bank of China (PBOC)
Financial	"Technical Specifications for Personal Financial Information Protection"	People's Bank of China (PBOC)
Financial	"Guidelines for Securities and Futures Industry Data Classification and Grading"	China Securities Regulatory Commission (CSRC)
Telecommunication	"Method for Data Classification and Grading of Basic Telecommunication Enterprises"	Ministry of Industry and Information Technology (MIIT) of China
Telecommunication	"Guidelines for Identifying Important Data in Basic Telecommunication Enterprises"	Ministry of Industry and Information Technology (MIIT) of China
Medical	"Information Security Technology—Healthcare Data Security Guidelines"	China National Information Security Standardization Technical Committee
Automotive	"Regulations on Automotive Data Security Management"	Ministry of Industry and Information Technology (MIIT) of China

References

1. Seaman, J. *PCI DSS: An Integrated Data Security Standard Guide*; Apress: Berkeley, CA, USA, 2020.
2. George, G. The Public Company Accounting Reform and Investor Protection Act of 2002: Any implications for Australia? *Aust. J. Corp. Law* **2002**, *14*, 286–295.
3. General Data Protection Regulation. General Data Protection Regulation (GDPR)—Final Text Neatly Arranged. Available online: https://gdpr.verasafe.com/ (accessed on 13 March 2024).
4. Pardau, S.L. The California consumer privacy act: Towards a European-style privacy regime in the United States. *J. Technol. Law Policy* **2018**, *23*, 68.
5. Brodsky, L.; Oakes, L. *Data Sharing and open Banking*; McKinsey & Company: Chicago, IL, USA, 2017; p. 1105.
6. Yuan, J. *Practice and Thoughts on Information Security Protection of Open Banks under the New Financial Situation*; Financial Electronification: 2021. Available online: https://www.secrss.com/articles/35541 (accessed on 13 March 2024).
7. Zu, L.; Li, H.; Zhang, L.; Lu, Z.; Ye, J.; Zhao, X.; Hu, S. E-SAWM: A Semantic Analysis-Based ODF Watermarking Algorithm for Edge Cloud Scenarios. *Future Internet* **2023**, *15*, 283. [CrossRef]
8. Guan, X.; Zhou, C.; Cao, W. Research on Classification Method of Sensitive Structural Data of Electric Power. In Proceedings of the 2022 IEEE 12th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 15–17 July 2022; pp. 268–271.
9. Rajkamal, M.; Sumathi, M.; Vijayaraj, N.; Prabu, S.; Uganya, G. Sensitive data identification and protection in a structured and unstructured data in cloud based storage. *Ann. Rom. Soc. Cell Biol.* **2021**, *25*, 1157–1166.
10. Ray, S.; Mishra, K.N.; Dutta, S. Sensitive data identification and security assurance in cloud and IoT based networks. *Int. J. Comput. Netw. Inf. Secur. IJCNIS* **2022**, *14*, 11–27. [CrossRef]
11. Mouza, C.; Métais, E.; Lammari, N.; Akoka, J.; Aubonnet, T.; Comyn-Wattiau, I.; Fadili, H.; Cherfi, S.S.S. Towards an automatic detection of sensitive information in a database. In Proceedings of the 2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications, Menuires, France, 11–16 April 2010; pp. 247–252.
12. Yi, T.; Shi, M. Privacy protection method for multiple sensitive attributes based on strong rule. *Math. Probl. Eng.* **2015**, *2015*, 464731. [CrossRef]
13. Xiao, Y.; Li, H. Privacy preserving data publishing for multiple sensitive attributes based on security level. *Information* **2020**, *11*, 166. [CrossRef]
14. Chong, P. Deep Learning Based Sensitive Data Detection. In Proceedings of the 2022 19th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 16–18 December 2022; pp. 1–6.
15. Silva, P.; Gonçalves, C.; Godinho, C.; Antunes, N.; Curado, M. Using nlp and machine learning to detect data privacy violations. In Proceedings of the IEEE INFOCOM 2020—IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 6–9 July 2020; pp. 972–977.
16. Ma, J.; Zhang, J.; Xiao, L.; Chen, K.; Wu, J. Classification of power quality disturbances via deep learning. *IETE Tech. Rev.* **2017**, *34*, 408–415. [CrossRef]
17. Park, J.s.; Kim, G.w.; Lee, D.h. Sensitive data identification in structured data through genner model based on text generation and ner. In Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things, Sanya, China, 24–26 April 2020; pp. 36–40.
18. Jiang, H.; Chen, C.; Wu, S.; Guo, Y. Classification of Medical Sensitive Data based on Text Classification. In Proceedings of the 2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Yilan, Taiwan, 20–22 May 2019; pp. 1–2.
19. Považanec, A. Comparison of Machine Learning Methods for Sensitive Data Identification. Undergraduate Thesis, Masaryk University, Brno, Czech Republic, 2020.

20. Yang, R.; Gao, X.; Gao, P. Research on intelligent recognition and tracking technology of sensitive data for electric power big data. In Proceedings of the 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Beihai, China, 16–17 January 2021; pp. 229–234.
21. Gambarelli, G.; Gangemi, A.; Tripodi, R. Is Your Model Sensitive? SPeDaC: A New Benchmark for Detecting and Classifying Sensitive Personal Data. *arXiv* **2022**, arXiv:2208.06216.
22. Dias, M.; Boné, J.; Ferreira, J.C.; Ribeiro, R.; Maia, R. Named entity recognition for sensitive data discovery in Portuguese. *Appl. Sci.* **2020**, *10*, 2303.
23. García-Pablos, A.; Perez, N.; Cuadros, M. Sensitive data detection and classification in Spanish clinical text: Experiments with BERT. *arXiv* **2020**, arXiv:2003.03106.
24. Aldeco-Pérez, R.; Moreau, L. A provenance-based compliance framework. In Proceedings of the Future Internet Symposium, Berlin, Germany, 20–22 September 2010; pp. 128–137.
25. Aldeco Perez, R.; Moreau, L. Provenance-based auditing of private data use. In Proceedings of the Visions of Computer Science—BCS International Academic Conference (VOCS), London, UK, 22–24 September 2008.
26. Yang, M.; Tan, L.; Chen, X.; Luo, Y.; Xu, Z.; Lan, X. Laws and regulations tell how to classify your data: A case study on higher education. *Inf. Process. Manag.* **2023**, *60*, 103240. [CrossRef]
27. Elluri, L.; Nagar, A.; Joshi, K.P. An integrated knowledge graph to automate gdpr and pci dss compliance. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Shanghai, China, 15–17 January 2018; pp. 1266–1271.
28. Elluri, L.; Joshi, K.P. A knowledge representation of cloud data controls for EU GDPR compliance. In Proceedings of the 2018 IEEE World Congress on Services (SERVICES), San Francisco, CA, USA, 2–7 July 2018; pp. 45–46.
29. Yang, M.; Chen, X.; Tan, L.; Lan, X.; Luo, Y. Listen carefully to experts when you classify data: A generic data classification ontology encoded from regulations. *Inf. Process. Manag.* **2023**, *60*, 103186.
30. Wang, J.; Wang, L.; Gao, S.; Tian, M.; Li, Y.; Xiao, K. Research on Data Classification and Grading Method Based on After sales Energy Replenishment Scenarios. In Proceedings of the 2022 2nd International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR), Xi'an, China, 25–27 November 2022; pp. 11–15.
31. *JR/T 0197-2022*; Financial Data Security—Guidelines for Data Security Classification. Technical Report. People's Bank of China: Beijing, China, 2020.
32. Quan, Y.; Yuquan, S. Research on Semantic Similarity Calculation Based on the Depth of "Synonymous Treebank". *J. Comput. Eng. Appl.* **2020**, *56*, 48–54.
33. Mei, J.; Zhu, Y.; Gao, Y.; Yin, H. *Synonym Word Forest*; Shanghai Dictionary Press: Shanghai, China, 1983.
34. Dong, Z.; Dong, Q. HowNet—A hybrid language and knowledge resource. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, 26–29 October 2003; pp. 820–824.
35. Agichtein, E.; Gravano, L. Snowball: Extracting relations from large plain-text collections. In Proceedings of the Proceedings of the Fifth ACM Conference on Digital Libraries, San Antonio, TX, USA, 2–7 June 2000; pp. 85–94.
36. Shen, J.; Lyu, R.; Ren, X.; Vanni, M.; Sadler, B.; Han, J. Mining entity synonyms with efficient neural set generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 249–256.
37. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
38. Song, Y.; Shi, S.; Li, J.; Zhang, H. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 2 (Short Papers), pp. 175–180.
39. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
40. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
41. He, W. Intelligent Recognition Algorithm and Adaptive Protection Model for Sensitive Data. Master's Thesis, Guizhou University, Guiyang, China, 2020.