Cloud Data Leakage Prevention mit Methoden der automatischen Datenklassifizierung

Anna Hamberger
Fakultät für Informatik
Technische Hochschule Rosenheim
Rosenheim, Germany
anna.hamberger@stud.th-rosenheim.de

***Zusammenfassung*—Die Cloud wird aufgrund ihrer Flexibilität, Skalierbarkeit und kostengünstigen Ressourcenverwaltung immer beliebter. Mit dem Einsatz von Cloud-Diensten steigt jedoch auch die Komplexität. Zudem verschiebt sich die Verantwortung zum Nutzer. Diese Faktoren können in Unternehmen zu unbeabsichtigten Datenlecks führen. Der Einsatz von Systemen zur Verhinderung von Datenlecks in der Cloud wird immer wichtiger, um Datenschutzverletzungen in der Cloud zu verhindern. Diese Systeme sind darauf ausgelegt, sensible Daten zu erkennen und zu klassifizieren sowie Verstöße gegen Sicherheitsrichtlinien zu verhindern und zu melden. Ein wesentlicher Aspekt dabei ist die automatische Klassifizierung von sensiblen Daten. Durch die zunehmenden Datenmengen im Kontext von Big Data sind effiziente Technologien und Methoden erforderlich. Der Fortschritt im Bereich künstlicher Intelligenz (KI) bietet hierbei einen vielversprechenden Ansatz. KI-basierte Methoden zur automatischen Datenklassifizierung können lernen, sensible Informationen zu erkennen und zu klassifizieren. Die Verwendung von Methoden wie k-NN, Boosting, Clusteranalyse und Sprachverarbeitungsmodelle haben dabei gute Ergebnisse erzielt. Cloud-Data-Leakage-Prevention-Systeme können auf Basis der klassifizierten Daten verschiedene Sicherheitsmaßnahmen ergreifen und so den Schutz vor unbeabsichtigten Datenlecks verbessern.**

***Index Terms*—machine learning, data classification, data leakage prevention, cloud security**

I. Einführung

Im Jahr 2022 gaben bereits 84% der befragten 552 Unternehmen in Deutschland an, Cloud-Dienste zu verwenden [1]. Cloud Computing hat sich während der digitalen Transformation zu einem wichtigen Bestandteil der Informationsverarbeitung entwickelt. Die Beliebtheit von Cloud-Diensten nimmt stetig zu, da sie die effiziente Speicherung großer Datenmengen, den schnellen Zugang zu Ressourcen und den nahtlosen Datenaustausch ermöglichen. Mit der Verbreitung digitaler Technologien in der Gesellschaft und in Unternehmen werden zunehmend mehr Daten geteilt. Unternehmen nutzen die Vorteile von Cloud-Diensten, um diese umfangreichen Datenmengen zu sammeln und zu verarbeiten. Die Möglichkeit, Daten in Echtzeit zu teilen, verbessert Geschäftsprozesse und erleichtert die unternehmensinterne Zusammenarbeit [2].

Da Informationen einen großen Wert für ein Unternehmen haben, ist ihr Schutz von größter Bedeutung. Beim Sammeln von Daten ist ein Unternehmen zudem verpflichtet, diese vor Diebstahl, Verlust und Missbrauch zu schützen. Zahlreiche Datenschutzgesetze und -vorschriften, wie die EU-Datenschutz-Grundverordnung (DSGVO), wurden eingeführt, um sensible Daten wie personenbezogene Informationen zu sichern. Ziel dieser Vorschriften ist es, klare Richtlinien für das Sammeln von Daten festzulegen und Einzelpersonen eine vergleichsweise hohe Kontrolle über ihre personenbezogenen Daten zu gewährleisten [3]. Unternehmen sind also sowohl intrinsisch als auch extrinsisch motiviert, Daten zu schützen. Intern liegt die Motivation darin, das Vertrauen der Kunden zu erhalten und die betriebliche Kontinuität sicherzustellen, während externe Faktoren wie gesetzliche Vorschriften und der Wettbewerbsdruck zusätzliche Anreize bieten, Sicherheitsstandards zu wahren.

Ein Hauptziel der Informationssicherheit besteht darin, die Offenlegung von Daten gegenüber Unbefugten zu verhindern. Datenlecks können jedoch nicht immer verhindert werden. Diese Bedrohung kann von böswilligen externen Akteuren ausgehen, die versuchen, an sensible Daten zu gelangen. Gleichzeitig können auch interne Mitarbeiter eine Gefahr darstellen, indem sie Informationen absichtlich oder unbeabsichtigt preisgeben [4]. Bereits im Jahr 2018 zeigten Studien, dass 53% der befragten Unternehmen in den letzten 12 Monaten von Insider-Angriffen betroffen waren. Dabei sind interne Bedrohungen oft schwerwiegender als externe, da sie in der Regel schwieriger zu erkennen sind [5]. Die Offenlegung von sensiblen Daten kann erheblichen Schaden verursachen. Unternehmen können ihren Wettbewerbsvorteil verlieren, ihr Image beeinträchtigen, Umsatzeinbußen erleiden oder sogar Geldstrafen und Sanktionen erhalten.

Um das Risiko von Datenschutzverletzungen zu minimieren, werden zunehmend Data-Leakage-Prevention (DLP) Lösungen eingesetzt. Gartner prognostiziert, dass bis 2027 etwa 70% der größeren Unternehmen eine DLP-Lösung einsetzen werden, um die Datensicherheit vor Insider-Risiken und externen Angreifern zu schützen [6]. DLP-Systeme überwachen den Zugriff und Austausch vertraulicher Daten, um unbefugte Offenlegung oder missbräuchliche Nutzung zu erkennen.

Unternehmen sammeln oft große Datenmengen, ohne genau zu wissen, welche Informationen erfasst werden. Auch die Suche nach und der Abruf von personenbezogenen Daten ist dabei eine große Herausforderung. Das erschwert den Schutz der Privatsphäre. DLP-Systeme benötigen Informationen darüber, ob bestimmte Daten besonders schützenswert sind oder

nicht. Im Zeitalter von Big Data ist es jedoch kaum noch möglich, die enormen Datenmengen manuell zu analysieren. Der Fortschritt im Bereich künstlicher Intelligenz (KI) bietet hierbei einen vielversprechenden Ansatz. KI-basierte Methoden zur automatischen Datenklassifizierung können in DLP-Systemen eingesetzt werden, um sensible Informationen zu erkennen.

Aufgrund der neuen Möglichkeiten durch den Einsatz von KI im Bereich Datenschutz liegt der Fokus in dieser Arbeit auf der Anwendung von Methoden der automatischen Datenklassifizierung zur Erkennung sensibler Informationen. Die zentrale Fragestellung dieser Arbeit befasst sich mit der effektiven Erkennung sensibler Daten in großen Datenmengen. Hierbei wird zunächst die Bedrohung durch versehentliche Offenlegung von Daten beschrieben und anschließend die Abwehrmaßnahme 'Data Leakage Prevention' vorgestellt. Dabei liegt der Fokus auf der Erkennung von sensiblen Informationen. Es werden verschiedene KI-basierte Methoden und ihre Funktionsweise im Bezug auf Datenklassifizierung vorgestellt. Anschließend wird deren Einsatz in der Cloud Sicherheit diskutiert.

II. Versehentliche Offenlegung von Daten in der Cloud

Die Cloud Security Alliance (CSA) veröffentlicht jährlich einen Bericht über die größten Bedrohungen der Cloud Security, der auf der Befragung von über 700 Experten basiert. Ziel dieses Berichts ist es, auf Bedrohungen, Risiken und Schwachstellen in der Cloud aufmerksam zu machen. Der aktuellste Bericht von 2022 hebt hervor, dass die Verantwortung für die Sicherheit in der Cloud vermehrt vom Cloud Service Provider zum Cloud-Kunden verlagert wird [7]. Diese Verschiebung erhöht das Risiko von Fehlern aufgrund von Unwissenheit. Ein zentrales Sicherheitsproblem ist die unbeabsichtigte Offenlegung von Cloud-Daten ('Accidental Cloud Data Disclosure').

Die versehentliche Offenlegung von Cloud-Daten ist eine Sicherheitsbedrohung, bei der sensible Informationen unbeabsichtigt öffentlich zugänglich gemacht werden. Das kann durch menschliches Versagen, Konfigurationsfehler oder unzureichende Sicherheitsmaßnahmen verursacht werden [7]. Ein Beispiel ist der Fall von Toyota Motor im Jahr 2023, bei dem persönlichen Daten von Kunden über mehrere Jahre offengelegt wurden. Der Grund war eine Fehlkonfiguration, wodurch die Datenbank in der Cloud öffentlich zugänglich war [8]. Ein weiteres Ereignis im August 2023 betraf die nordirische Polizei, als ein interner Mitarbeiter versehentlich persönliche Daten der aktuellen Beamten auf einer Online-Plattform veröffentlichte, indem er die Datei verwechselte [9].

Die beiden genannten Beispiele verdeutlichen die potenziell schwerwiegenden Folgen von Fehlern, die schnell zu erheblichem Schaden führen können. Mit der steigenden Verantwortung für die Sicherheit in der Cloud seitens der Kunden erhöht sich das Risiko menschlichen Versagens. Social Engineering und Phishing-Attacken stellen eine Gefahr dar, da Mitarbeiter unbeabsichtigt sensible Daten wie Zugangsdaten offenlegen können. Wie im Fall der nordirischen Polizei gezeigt, besteht auch das Risiko, dass Mitarbeiter Daten unwissentlich veröffentlichen. Zu einer ungewollten Offenlegung von Daten kann es auch kommen, wenn Geräte wie Laptops oder Smartphones, die nicht ausreichend geschützt sind, verloren gehen. Der einfache Zugang zu Cloud-Ressourcen kann dazu führen, dass neue Ressourcen oder Dienste ohne ausreichende Sicherheitsüberlegungen genutzt werden, wodurch Daten aufgrund von Fehlkonfigurationen offengelegt werden könnten. Nicht nur menschliches Versagen, sondern auch Schwächen im Zielsystem stellen Risiken dar. Schwache Passwörter, mangelnde Authentifizierung bei sicherheitsrelevanten Systemen und andere Konfigurationsfehler können bewirken, dass Daten in der Cloud unbeabsichtigt offengelegt werden. Ebenso stellen ungeschlossene Sicherheitslücken in genutzten Cloud-Services eine Bedrohung dar [10] [11].

Um das Risiko einer versehentlichen Datenpreisgabe zu minimieren, können verschiedene Schutzmaßnahmen ergriffen werden. Ein kontrolliertes Identity Access Management (IAM) ermöglicht die Regulierung und Kontrolle des internen und externen Datenzugriffs. Die Einführung strenger Passwortrichtlinien und die Nutzung von Passwort-Manager-Software reduzieren das Risiko unbefugten Zugriffs auf Geräte, Benutzerkonten oder Cloud-Ressourcen. Das Prinzip des geringsten Privilegs gewährleistet, dass Benutzer nur die notwendigen Berechtigungen für ihre Aufgaben erhalten, was das Risiko von Fehlkonfigurationen oder missbräuchlichem Zugriff minimiert. Neben der Kontrolle der Zugriffe ist auch die Überwachung möglicher Schwachstellen entscheidend. Regelmäßige Schwachstellen-Scans helfen, Sicherheitslücken in der Cloud-Infrastruktur zu identifizieren und zu beheben, bevor sie ausgenutzt werden können. Die Überprüfung und Optimierung von Cloud-Konfigurationen gewährleistet korrekte Sicherheitseinstellungen. Eine zentrale Verwaltung aller in der Cloud vorhandenen Assets ermöglicht eine bessere Kontrolle und Überwachung von Daten, Diensten und Einstellungen. Regelmäßige Softwareaktualisierungen sind wichtig, um bekannte Sicherheitslücken zu schließen. Mitarbeiter sollten zudem durch Schulungen für sicherheitsrelevante Themen sensibilisiert werden, um menschliche Fehler zu minimieren [11].

Im Kontext dieser Bedrohung werden die Begriffe Data Loss (Datenverlust) und Data Leakage (Datenleck) häufig als Synonyme verwendet, weisen jedoch einige Unterschiede auf. Datenverlust bezieht sich auf den unwiederbringlichen Verlust von Daten, beispielsweise durch Schäden an Speichermedien, unbeabsichtigtes Löschen oder Hardwarefehler. Im Gegensatz dazu bezeichnet Datenleck die unbeabsichtigte oder absichtliche Übertragung von Daten aus einem gesicherten Bereich. Datenlecks können auftreten, wenn unbefugte Personen sensible oder vertrauliche Informationen erhalten [12]. Daher wird in dieser Arbeit der Begriff Datenleck oder Data Leakage verwendet, um die unbeabsichtigte Offenlegung von Daten zu beschreiben.

Die zunehmende Komplexität der Cloud-Infrastrukturen und die steigende Verantwortung der Kunden für die Sicherheit haben das Risiko der versehentlichen Offenlegung sensibler Daten erheblich erhöht. In Anbetracht dieser Herausforderun-

gen und um das Risiko von Datenlecks zu minimieren, werden DLP-Systeme zunehmend als entscheidende Sicherheitsmaßnahme eingesetzt.

III. Cloud Data Leakage Prevention

Im vorherigen Kapitel II wurde die Gefahr der versehentlichen Datenoffenlegung in Unternehmen verdeutlicht und die Notwendigkeit effektiver Maßnahmen zur Vermeidung von Datenlecks hervorgehoben. Angesichts der wachsenden Datenmengen und des damit verbundenen Risikos von Datenschutzverletzungen hat sich dieser Bedarf verstärkt. Dies wurde im Jahr 2022 erkannt, als die neueste Version der Norm ISO 27001:2022 die Data Leakage Prevention eingeführt hat. Die internationale Norm ISO 27001 legt die Bedingungen für die Einrichtung, Umsetzung und kontinuierliche Verbesserung eines dokumentierten Informationssicherheitsmanagementsystems fest. Zudem bietet sie Vorschriften für die Beurteilung und Behandlung von Informationssicherheitsrisiken, die an die spezifischen Bedürfnisse jedes Unternehmens angepasst werden müssen [13].

Ein Datenleck kann auf verschiedene Arten auftreten. Obwohl es nicht immer möglich ist, das Auftreten vollständig zu verhindern, können Maßnahmen ergriffen werden, um die Wahrscheinlichkeit zu verringern. Diese Maßnahmen werden als Data Leakage Prevention bezeichnet [13]. DLP umfasst eine Vielzahl von Technologien, Produkten und Methoden, die darauf abzielen, zu verhindern, dass vertrauliche Informationen ein Unternehmen unautorisiert verlassen. In den letzten Jahrzehnten wurden verschiedene Sicherheitssysteme wie Firewalls, Intrusion-Detection-Systeme (Einbrucherkennung) und virtuelle private Netzwerke (VPN) eingeführt, um das Risiko von Datenlecks zu reduzieren. Diese Systeme erfüllen ihren Zweck gut, wenn die zu schützenden Daten klar definiert, strukturiert und konstant sind. Jedoch sind sie weniger zuverlässig für Daten, die sich ändern oder unstrukturiert sind. Eine Firewall kann bspw. den Zugriff auf ein sensibles Datenobjekt durch einfache Regeln verhindern. Allerdings erkennt die Firewall nicht zwangsläufig, wenn dasselbe Datenobjekt über einen E-Mail-Anhang gesendet wird. DLP-Systeme hingegen sind darauf spezialisiert, vertrauliche Daten zu identifizieren, zu überwachen und zu schützen. Sie können so unerwünschte Datenbewegungen effektiver verhindern [4].

Ein DLP-System setzt sich aus einer Vielzahl von Regeln und Richtlinien zusammen, die Daten nach Typ klassifizieren und sicherstellen, dass sie nicht böswillig oder versehentlich verbreitet werden. Das System überwacht die Aktivitäten der Endnutzer, den Datenfluss und die über das Netzwerk übertragenen Informationen. Sobald es verdächtige Aktivitäten entdeckt, löst es eine Systemwarnung aus. DLP-Lösungen identifizieren sensible Inhalte mithilfe von Datenklassifizierungslabels, Inhaltsprüfungsverfahren und Kontextanalysen. Sie überwachen und kontrollieren die Datenaktivitäten nach vordefinierter DLP-Richtlinien. Diese Richtlinien legen fest, ob die Verwendung bestimmter Inhalte oder Daten in bestimmten Situationen zulässig ist [6].

Gartner klassifiziert DLP-Lösungen in drei Kategorien: Enterprise-DLP-System, integriertes DLP-System, Cloudnatives DLP-System. Eine Enterprise-DLP-Lösung ist ein zentrales System, das darauf ausgelegt ist, komplexe Anforderungen und Strukturen großer Unternehmen zu bewältigen. Sie verfügt über fortschrittliche Technologien zur Identifikation, Klassifizierung und Markierung sensibler Daten und kann verschiedene Datenquellen integrieren. Diese Lösung deckt den gesamten Lebenszyklus von Daten in einem Unternehmen ab, wobei DLP-Richtlinien zentral verwaltet und durchgesetzt werden. Integrierte DLP-Lösungen hingegen werden direkt in einen Dienst, wie in ein E-Mail-Gateway, integriert und verfügen daher nur über begrenzte Richtlinienfunktionen. Das Management von mehreren integrierten DLP-Systemen erfordert manuellen Aufwand, aber diese Systeme sind speziell an die Anforderungen des jeweiligen Dienstes angepasst und können Inhaltsüberprüfungen effektiver durchführen. Die dritte Kategorie steht für Cloud-native DLP-Lösungen, die sowohl Software-as-a-Service-Lösungen (SaaS) als auch Cloud-Anbieter mit integrierten DLP-Funktionen umfassen. Sie sind speziell für den Einsatz in Cloud-Umgebungen entwickelt und darauf ausgerichtet, sensible Daten in Cloud-Diensten zu schützen. Diese Lösungen verfügen über Mechanismen zur automatischen Erkennung von sensiblen Daten, die in Cloud-Anwendungen und -Speicherplätzen gespeichert sind, einschließlich der Identifikation von Daten in Form von Dokumenten, E-Mails, Datenbanken und anderen Formaten [6]. Für die Verhinderung der versehentlichen Offenlegung von sensiblen Daten in der Cloud ist dementsprechend ein Cloudnatives DLP-System die beste Wahl für ein Unternehmen. Im weiteren Verlauf der Arbeit wird der Begriff DLP-System für alle drei Kategorien verwendet.

Das Cybersecurity Framework des National Institute of Standards and Technology (NIST CSF) stellt freiwillige Standards und Best Practices bereit, um Unternehmen bei der Verwaltung und Reduzierung von Cybersecurity-Risiken zu unterstützen. Das CSF besteht aus fünf Kernfunktionen, die in Abbildung 1 dargestellt sind: Identifizieren, Schützen, Erkennen, Reagieren und Wiederherstellen. DLP-Systeme konzentrieren sich hauptsächlich auf die Identifizierung, die Erkennung und den Schutz und ergänzen diese Funktionen durch den Bereich der Überwachung. Die spezifischen Funktionen eines DLP-Systems können je nach Hersteller variieren [14].

Die Literaturrecherche hat eine Auswahl an bewährten Methoden ergeben, die in DLP-Systemen vorhanden sein sollten. Um sensible Daten wirksam schützen zu können, ist zunächst die Identifikation erforderlich. Hierbei besteht die Aufgabe darin, ein Dateninventar zu erstellen, die Daten nach ihrer Sensibilität zu klassifizieren und entsprechend zu kennzeichnen. Zum Schutz sensibler Daten sollten Maßnahmen ergriffen werden, die den Zugriff auf die Daten einschränken. Dies beinhaltet die Einführung von Richtlinien wie minimale Zugriffsrechte, starke Authentifizierungsmethoden und strenge Zugriffskontrolllisten. Darüber hinaus sollten Daten sowohl im Ruhezustand als auch während der Übertragung verschlüsselt werden, um sicherzustellen, dass sie für unbefugte Benutzer

Abbildung 1: NIST Cybersecurity Framework Version 1.1. Quelle: [14].

selbst dann unlesbar bleiben, wenn sie abgefangen werden. Ein DLP-System sollte auch die Datenströme innerhalb und außerhalb des Unternehmens überwachen, um potenzielle Datenschutzverletzungen oder Richtlinienverstöße in Echtzeit zu erkennen. Dies ermöglicht eine schnelle Reaktion auf potenzielle Probleme und begrenzt den daraus resultierenden Schaden [16] [17].

Die Funktionen eines DLP-Systems basieren darauf, dass sensible Daten erkannt und in irgendeiner Art und Weise markiert sind. Der erste Schritt bei DLP-Systemen ist daher die Identifizierung sensibler Daten. Es gibt verschiedene Strategien und Methoden zur Klassifizierung dieser Daten, die durch den Einsatz von KI weiter verbessert wurden.

IV. Methoden der automatischen Datenklassifizierung

Unternehmen setzen immer mehr auf cloudbasierte SaaS-Produkte, anstatt diese selbst installieren und verwalten zu müssen [18]. In ihrem Tagesgeschäft verlassen sich Unternehmen oft auf mehrere Softwareprodukte, um unterschiedliche Anforderungen zu erfüllen. Das führt dazu, dass Unternehmensdaten über verschiedene Applikationen und Cloud-Plattformen verteilt sind. Die Herausforderung besteht darin, den Überblick zu behalten und zu wissen, wo sich die sensiblen Daten befinden. Das Sammeln und Identifizieren von Daten in DLP-Systemen stellt aufgrund von Verschlüsselung, verborgenen Kanälen, nicht unterstützten Datenformaten und großer Mengen an Daten eine große Herausforderung dar [19].

A. Erkennung von sensiblen Daten

Die Methoden zur Erkennung und Klassifizierung von sensiblen Daten unterscheiden sich je nach Art und Format der Daten, sowie deren Zustand. Außerdem gibt es die Möglichkeit, Daten manuell oder automatisiert zu klassifizieren.

1) Eigenschaften von Daten: Sensible Daten sind nahezu in jedem Aspekt unseres persönlichen und beruflichen Lebens präsent. Das Spektrum dieser sensiblen Informationen reicht von persönlichen Daten über Finanzinformationen und Geschäftsgeheimnissen bis hin zu biometrischen Merkmalen. Guo, Liu et al. [20] kategorisieren beispielsweise Daten in vier Bereiche:

- Persönliche Informationen (z.B. Name, Geburtsdatum oder Gesundheitsinformationen)
- Informationen zur Netzwerkidentität (z.B. IP-Adresse, MAC-Adresse oder E-Mail)
- Vertrauliche und Anmeldeinformationen (z.B. Login-Passwort-Kombinationen, API-Token oder digitale Zertifikate)
- Finanzinformationen (z.B. Bankkontodaten, Kreditkarteninformationen oder Verbrauchsdaten)

Die Kategorisierung und der Detailgrad können je nach Unternehmen variieren.

Zusätzlich zu den Kategorien muss auch der Kontext berücksichtigt werden, in dem eine Information verwendet wird. Der Kontext hat direkten Einfluss auf die Sensibilität. Pogiatzis und Samakovitis [21] leiten vier verschiedene Kontextklassen ab, die auf der Bedeutung, der Interaktion, der Priorität und der Präferenz basieren, die mit jeder Information verbunden sind.

- Der semantische Kontext entsteht auf Grundlage der semantischen Bedeutung eines Begriffs. Die semantische Bedeutung einer Sequenz wirkt sich also auch auf ihre Sensibilität aus.
- Im Kontext der Akteure hängt die Sensibilität von Daten von den Akteuren ab, die an der Informationsübermittlung beteiligt sind. Die Sensibilität wird durch die Beziehung zwischen den beteiligten Akteuren bestimmt. Zum Beispiel ist der Austausch von Gesundheitsinformationen zwischen Patient und Arzt nicht sensibel, außerhalb dieser Gruppe von Akteuren jedoch schon.
- Der zeitliche Kontext bezieht sich auf die Priorität der Informationen, die die Bedeutung des Begriffs beeinflussen. Eine Zeichenfolge, die als Passwort eingegeben wird, gilt bspw. als vertraulicher als wenn sie als Benutzername eingegeben wird.
- Der Selbstkontext wird durch die persönlichen Präferenzen des Nutzers in Bezug auf seine Privatsphäre bestimmt. Zum Beispiel kann eine Person ihre ethnische Herkunft als vertrauliche Information betrachten, eine andere nicht.

Auch hier können verschiedene kontextuelle Kategorien unterschieden oder definieren werden. Zudem sind sie nicht immer klar trennbar und schließen sich nicht gegenseitig aus. Ein oder mehrere Kontexte können gleichzeitig unterschiedlich auf die Sensibilität wirken. Manche Daten können jedoch auch unabhängig vom Kontext vertraulich sein, wie z.B. Passwörter oder Kreditkartennummern.

Die Klassifizierung von Daten in verschiedene Geheimhaltungsklassen ist ein häufig verwendetes Verfahren in militärischen und behördlichen Organisationen. Militärische Anwendungen verwenden dabei Begriffe wie 'eingeschränkt', 'vertraulich', 'geheim' und 'streng geheim' [22]. So ist es möglich, sensible Daten noch präziser in Vertraulichkeitsstufen zu unterteilen.

Die Klassifizierung wird auch von der Struktur der Daten beeinflusst. Über 80% der Daten im Internet bestehen aus unstrukturierten Daten [23]. Unstrukturierte Daten beziehen sich in der Regel auf Informationen, die nicht in einer relationalen Datenbank gespeichert sind. Folglich gibt es kein vordefiniertes Datenmodell und die Struktur ist unregelmäßig oder unvollständig. Selbst Datenformate wie CSV, JSON oder XML, die einige organisatorische Eigenschaften haben, verfügen in der Regel nicht über ein klar definiertes Datenmodell. Im Vergleich zu strukturierten Daten ist es schwieriger, unstrukturierte Daten abzurufen, zu analysieren und zu speichern. Während Menschen unstrukturierte Daten leicht verarbeiten können, haben Maschinen oft Schwierigkeiten damit [20].

Die Herausforderung bei der Datenklassifizierung besteht daher darin, die Kategorien, den Kontext, die Vertraulichkeitsstufen und die Struktur der Daten zu berücksichtigen.

2) Datenzustand: Im Bereich der Informationssicherheit werden Daten je nach ihrem Zustand unterschiedlich betrachtet. Die verschiedenen Datenzustände helfen dabei, die geeigneten Sicherheitsmaßnahmen zu bestimmen. Im Rahmen der Data Leakage Prevention können sich Daten in einem der drei Zustände befinden: Daten im Ruhezustand, Daten in Bewegung und Daten in Verwendung [24].

Daten im Ruhezustand sind solche, die auf einem physischen oder digitalen Speichermedium gespeichert sind, wie in einer Datenbank, auf einer Festplatte, im Cloud-Speicher oder einem externen Datenspeicher. Ruhende Daten sind in der Regel inaktiv, werden nicht gelesen oder übertragen. DLP-Systeme setzen für ruhende Daten Sicherheitsmaßnahmen wie Verschlüsselung, Authentifizierung und Zugriffskontrollregeln zu Erkennungs- und Überwachungszwecken ein [17] [24].

Daten in Bewegung beziehen sich hingegen auf den Zustand von Daten, wenn sie aktiv über Netzwerke oder andere Kommunikationskanäle übertragen werden oder sich im Speicher eines Computers befinden und zum Lesen, Aktualisieren und Verarbeiten bereit sind. Beispiele für Daten in dieser Kategorie sind Daten, die über das Internet, soziale Medien, E-Mail oder FTP/SSH übertragen werden. Daten, die über das Netzwerk übertragen werden, sollten mit Verschlüsselungsmethoden wie HTTPS, SSL oder TLS geschützt werden. DLP-Systeme nutzen Erkennungs- und Überwachungsfunktionen, um den Datenfluss durch das Netzwerk zu identifizieren und zu überwachen [17] [24].

Daten in Verwendung sind solche, die aktiv von einer Person oder einem System verarbeitet, aktualisiert, anhängt oder gelöscht werden. Diese Daten befinden sich auf dem Laptop, dem USB-Stick oder der externen Festplatte des Endnutzers sowie auf Netzlaufwerken oder Druckern. Da diese Daten aktiv manipuliert werden, sind sie besonders anfällig für Sicherheitsbedrohungen. Sie werden über verschiedene Endpunkte hinweg sichtbar, wenn darauf zugegriffen wird. Um diese Daten zu schützen, implementiert das DLP-System starke Benutzerauthentifizierung sowie ein Identitäts- und Profilmanagement. Je nach DLP-Technologie kann auch ein Endpunkt-Agent auf dem Gerät des Endnutzers installiert werden, um die Datennutzung und -übertragung zu überwachen [17] [24].

Die Unterscheidung zwischen den drei Zuständen von Daten - in Ruhe, in Bewegung und in Verwendung - ist entscheidend für die Identifizierung sensibler Informationen und die Klassifizierung von Daten. Durch die Berücksichtigung der Datenzustände können sensible Daten differenziert betrachtet und klassifiziert werden.

B. Herausforderungen in der Datenklassifizierung

In der Literatur wurden verschiedene Studien durchgeführt, um Klassifizierungsmethoden zur Gewährleistung der Datensicherheit in der Cloud zu untersuchen. Die vorgeschlagenen Lösungen lassen sich in manuelle und automatische Klassifizierung unterteilen. Bei der manuellen Klassifizierung werden die Klassen durch eine Person festgelegt und bei der automatische Klassifizierung werden dafür Algorithmen eingesetzt. Obwohl automatisierte Technologien Fortschritte machen, bleibt die manuelle Datenklassifizierung, insbesondere für sensible Informationen wie geistiges Eigentum, weit verbreitet. Trotz ihres Potenzials für umfassende Berücksichtigung von Datenkontexten und -zuständen birgt die manuelle Klassifizierung Risiken wie menschliche Fehler und Inkonsistenzen. Zudem kann sie bei großen Datenmengen zeitaufwändig werden und die Anpassungsfähigkeit beeinträchtigen [27]. Angesichts des wachsenden Datenvolumens erscheint die automatisierte Datenklassifizierung als effizientere Lösung für präzise und konsistente Identifizierung sensibler Daten.

Die automatische Datenklassifizierung bezeichnet den Prozess, bei dem maschinelle Algorithmen und Technologien verwendet werden, um Daten automatisch zu identifizieren, zu kategorisieren und entsprechend ihrer Sensitivität zu klassifizieren. Diese Methode nutzt maschinelles Lernen, Mustererkennung oder künstliche Intelligenz, um Daten zu analysieren und automatisch passende Klassifizierungen zuzuweisen.

Aktuelle Techniken in der Data Leakage Prevention lassen sich im Allgemeinen in zwei Kategorien unterteilen: inhaltsbasierte Analyse und kontextbasierte Analyse. Inhaltsbasierte Methoden untersuchen den Dateninhalt anhand von Merkmalen sensibler Informationen. Sie nutzen vorhersehbare Muster wie bspw. IP-Adressen, E-Mail-Adressen oder bestimmte Wörter, um sensible Daten zu erkennen. Kontextbasierte Techniken identifizieren vertrauliche Daten anhand von Merkmalen im Zusammenhang mit den überwachten Daten. Der kontextbasierte Ansatz ist somit effektiver für vertrauliche Daten ohne vorhersehbare Muster [3].

Um sensible Informationen umfassender und genauer zu extrahieren, sollten daher verschiedene Methoden angewendet werden. Tabelle I zeigt die am häufigsten verwendeten Methoden in der Literatur. Die meisten kontexbasierten Ansätze kombinieren inhaltsbasierte und kontextbasierte Methoden, um die Vorteile beider Kategorien zu nutzen und die Genauigkeit der Klassifizierung zu verbessern.

C. Klassifizierung mit manueller Definition

In diesem Abschnitt werden regelbasierte Methoden und Data Fingerprinting als zwei häufig verwendete Ansätze zur

Tabelle I: Methoden der automatischen Datenklassifizierung. Quelle: eigene Darstellung.

Kategorie	Methode
inhaltsbasiert	regelbasierte Methoden Data Fingerprinting kNN Boosting
inhalts- und kontextbasiert	Clusteranalyse CASSED BERT-BiLSTM

automatischen Datenklassifizierung vorgestellt. Eine gemeinsame Eigenschaft dieser Methoden ist, dass sie voraussetzen, dass eine Person zunächst Regeln oder ein Wörterbuch definiert, die die Methode dann zur automatischen Klassifizierung der Daten verwendet.

a) regelbasierte Methoden: Regelbasierte Methoden zur automatischen Datenklassifizierung verwenden im Wesentlichen ein Wörterbuch oder eine Regel, um den gegebenen Text mit einer Liste regulärer Ausdrücke und Schlüsselwörter abzugleichen. Diese Methode setzt vordefinierte Regeln und Bedingungen ein, um bestimmte Datentypen oder Muster automatisch zu identifizieren und zu klassifizieren. Diese Regeln können auf verschiedenen Merkmalen wie Schlüsselwörtern, Mustern, Dateiformaten oder spezifischen Attributen wie Datumsangaben basieren [29]. Ein Beispiel für die Anwendung regelbasierter Methoden ist die Identifikation von Kreditkartennummern in Textdokumenten. Hier kann ein regulärer Ausdruck verwendet werden, der die Zeichenfolge nach dem Muster einer Kreditkartennummer definiert. Muster in regulären Ausdrücken umfassen normale Zeichen mit wörtlicher Bedeutung sowie Metazeichen, um ein Erkennungsmuster zu bilden [4]. Der reguläre Ausdruck für die MasterCard Kreditkartennummer wäre zum Beispiel `^5[1-5][0-9]{14}$`. Dieser Ausdruck definiert, dass alle MasterCard-Nummern mit einer 5 beginnen, dann eine Zahl zwischen 1 und 5 enthalten und anschließend eine beliebige Reihenfolge von 14 Zahlen haben. Regeln können sich auch auf bestimmte Schlüsselwörter oder Phrasen beziehen, wie bspw. 'Sozialversicherungsnummer' oder 'vertraulich', die auf personenbezogene Informationen hinweisen können. Der Vorteil regelbasierter Methoden liegt in ihrer klaren Struktur und der Möglichkeit, spezifische Anforderungen und Richtlinien der Organisation abzubilden. Sie ermöglichen eine präzise und konsistente Klassifizierung von Daten gemäß vordefinierten Sicherheitsstandards. Da die Klassifizierung auf klaren, vorher festgelegten Regeln basiert, ermöglichen regelbasierte Ansätze auch eine gewisse Transparenz und Nachvollziehbarkeit. Jedoch können regelbasierte Methoden bei der Verarbeitung komplexer und sich verändernder Datenmuster weniger nützlich sein. Die Regeln werden schnell unpraktisch, wenn verschiedene Datenformate, Kontexte, Wortvariationen und Abkürzungen kombiniert werden müssen. Der effektive Einsatz dieser Methode erfordert zudem ein gut definiertes und gepflegtes Wörterbuch und Regelwerk. Außerdem wird der semantische Kontext der Wörter bei einem reinen Textabgleich nicht berücksichtigt, was zu einer geringen Genauigkeit der Klassifizierung führen kann [29]. Trotz dieser Einschränkungen bieten regelbasierte Ansätze eine grundlegende und robuste Methode zur Sicherstellung einer konsistenten Datenklassifizierung.

b) Data Fingerprinting: Data Fingerprinting oder auch Document Fingerprinting erstellt eindeutige Fingerabdrücke für bestimmte Datenfragmente oder ganze Dateien. Diese Fingerabdrücke dienen als eindeutige Identifikatoren für die entsprechenden Daten und werden genutzt, um sensible Daten zu identifizieren und automatisch zu klassifizieren. Anhand von Wortmustern aus regulären Ausdrücken oder vordefinierten Wörterbüchern werden benüötigte Wörter, Sätze oder ganze Dateien identifiziert und eine Auswahl an eindeutigen Fingerabdrücke dafür erstellt, die als Vorlage dienen. Diese Fingerabdrücke können dann verwendet werden, um Fingerabdrücke von nicht-klassifizierten Daten zu vergleichen und zu klassifizieren. Häufig werden Hash-Funktionen wie MD5 oder SHA1 verwendet, um Datenfingerprints zu erstellen, die eine algorithmisch generierte Zeichenfolge fester Größe für die Daten darstellen. Die Hashes von zwei Dateien unterscheiden sich jedoch bereits, sobald nur ein Zeichen verändert wurde [4]. Ein weiterer Ansatz ist deshalb das sogenannte 'Fuzzy-Hashing'. Dabei werden die Daten in Blöcken verarbeitet, wodurch die Hash-Ausgabe bei ähnlichen Daten größtenteils übereinstimmende Blöcke enthält. So kann die prozentuale Ähnlichkeit mithilfe einer mathematischen Vergleichsfunktion bestimmt werden [30].

Das Data Fingerprinting kann sowohl im Speicher als auch bei Netzwerkübertragungen und Verwendung von Daten eingesetzt werden. Es ist jedoch wichtig zu beachten, dass der Fingerabdruck in manchen Fällen keine zuverlässige Methode ist. Die Klassifizierung funktioniert zun Beispiel nicht, wenn die Daten verschlüsselt oder passwortgeschützt sind oder wenn der Inhalt nicht eindeutig mit dem Fingerabdruck übereinstimmt. Zudem ist es notwendig, dass die Vorlagen kontinuierlich aktualisiert werden. Außerdem kann der Ansatz kann bei großen Datenmengen ressourcenintensiv sein, da die Fingerabdrücke ständig gespeichert und berechnet werden müssen.

D. Klassifizierung mit künstlicher Intelligenz

Die Datenklassifizierung ist eine Technik aus dem Bereich der KI, mit der die Klassen von nicht klassifizierten Daten vorhergesagt wird. KI-Methoden lassen sich in zwei Kategorien einteilen: überwachtes Lernen und unüberwachtes Lernen. Beim überwachten Lernen sind Testdaten mit bereits zugeteilten Klassen vordefiniert. Das Modell kann seine Ergebnisse mit den Zielklassen vergleichen und von den Fehlern lernen. Im Gegensatz dazu sind beim unüberwachten Lernen keine Klassen definiert, sondern die Klassifizierung der Daten erfolgt automatisch. Ein unüberwachter Algorithmus sucht nach Mustern und Ähnlichkeiten zwischen den Elementen [33].

Die meisten Methoden erfordern eine vorherige Bereinigung oder Vorbereitung der Daten, wobei die genauen Schritte je nach Methode variieren können. Der Datensatz wird meistens von Stop-Wörtern bereinigt, die keinen oder nur geringen Kontext liefern, wie bspw. 'und' oder 'der'. In der Textverarbeitung sind Stemming und Lemmatization klassische Bereinigungsmethoden. Beim Stemming werden die Wörter der Eingabe auf ihren Wortstamm zurückgeführt, während bei der Lemmatization ähnliche oder inhaltlich gleiche Begriffe vereinheitlicht werden. Diese Schritte dienen dazu, die grammatikalische Komplexität des Inputs zu reduzieren und die Datenqualität zu optimieren. Für Methoden wie neuronale Netze müssen die Eingabedaten zusätzlich in Token und Vektoren umgewandelt werden. Bei der Tokenization wird der Text in einzelne Bestandteile, sogenannte Tokens, aufgeteilt. In der Regel sind das einzelne Worte. Die Tokens werden dann in Vektoren umgewandelt, die das jeweilige Token im Modell repräsentieren. Für die Vektorisierung gibt es verschiedene Algorithmen, um Informationen wie inhaltlich ähnliche oder zusammenhängende Tokens zu behalten [31].

Im Folgenden werden Klassifizierungsmethoden aus dem Bereich der künstlichen Intelligenz analysiert. Dabei wurden Methoden ausgewählt, die sich in Studien als effektiv für die automatischen Klassifizierung im Kontext der Informationssicherheit bewährt haben.

a) k-NN: Die Methode von Zardari, Jung et al. [32] markiert einen Meilenstein, indem sie laut eigener Aussage als erste maschinelles Lernen zur Datenklassifizierung im Kontext des Cloud Computing einsetzen. Der Bereich des maschinellen Lernens ist eine Unterkategorie der künstlichen Intelligenz, dessen Ziel es ist, präzise Ergebnisse durch Identifizieren von Mustern zu liefern. Sie verwendeten die k-Nearest-Neighbor-Methode (k-NN), um Daten in der Cloud als sensibel und nicht-sensibel zu klassifizieren. K-NN ist ein überwachter maschineller Lernalgorithmus, der für Klassifizierung, Mustererkennung und Schätzungen verwendet wird. Diese Methode geht davon aus, dass ähnliche Dinge normalerweise nahe beieinander liegen. Die k-NN-Methode verwendet die nächsten Nachbarn eines Datenpunkts, um Vorhersagen oder Klassenzuweisungen zu treffen. Der Algorithmus beginnt mit einem Datensatz, der Beispiele mit bekannten Klassen oder Werten enthält. Die Wahl des 'k' bestimmt die Anzahl der nächsten Nachbarn. Um die Ähnlichkeit zwischen den Datenpunkten zu bestimmen, wird ein Distanzmaß wie der euklidische Abstand oder die Manhattandistanz verwendet. Sie messen den Abstand zwischen den Merkmalsvektoren der Datenpunkte. Für einen gegebenen Datenpunkt werden die k nächsten Nachbarn basierend auf dem berechneten Distanzmaß aus dem Datensatz ausgewählt. Bei der Klassifikation werden die Klassen der ausgewählten k Nachbarn betrachtet und die Mehrheitsklasse wird für den gegebenen Datenpunkt verwendet [33].

In [32] hat die Einteilung der Daten in zwei Klassen gut funktioniert. Doch die Wahl des passenden 'k' ist bei diesem Algorithmus entscheidend. Ein zu kleines 'k' kann dazu führen, dass potenziell passende Datenpunkte ausgeschlossen werden und ein zu großes 'k' führt zu einer zu groben Klassifizierung. Abbildung 2 veranschaulicht die Auswirkungen der Wahl des 'k'. Je nach Größe des k wird der unbekannte Punkt (roter Stern) entweder der Klasse A oder der Klasse B zugeordnet. Aufgrund der Verwendung aller Trainingsdaten ist zudem die Rechenkomplexität bei diesem Algorithmus hoch, da bei jeder Vorhersage der Abstand berechnet und die gesamten Trainingsdatendistanzen sortiert werden müssen. Außerdem ist die Mehrheitsentscheidung bei der Klassifizierung nicht immer die optimale Methode, da je nach Anzahl der Nachbarn die Abstände stark variieren können und trotzdem immer alle gewählten Nachbarn berücksichtigt werden.

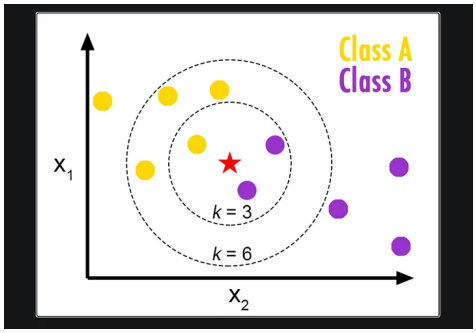


Abbildung 2: Einfluss des gewählten 'k'. Quelle: [34].

b) Boosting: Ensemble Learning ist ein Begriff aus dem maschinellen Lernen und beschreibt das Zusammenschalten von mehreren Methoden, um das Modellergebnis zu verbessern. Boosting ist ein Ensemble-Learning-Ansatz, der mehrere schwache Modelle zu einem starken Modell zusammenzusetzt [35]. Kaur, Zandu [36] schlagen für die Klassifizierung von sensiblen Daten eine neue Boosting-Architektur vor. Als Klassifikator wird eine Kombination aus dem Naive Bayes Klassifikator und AdaBoost verwendet. Der Naive Bayes Algorithmus ist eine Klassifizierungsmethode, die auf dem Bayes'schen Theorem beruht. Die Grundlage ist die Annahme, dass das Auftreten eines Merkmals unabhängig vom Auftreten eines anderen Merkmals innerhalb der Klasse ist. Das Bayes'sche Theorem ist eine Formel zur Berechnung der bedingten Wahrscheinlichkeit $P(A|B)$, also der Wahrscheinlichkeit von A unter der Bedingung von B. Diese bedingte Wahrscheinlichkeit lässt sich mit der Formel 1 berechnen. Diese Methode ermöglicht die Klassifizierung eines Merkmals, indem für jede Klasse berechnet wird, mit welcher Wahrscheinlichkeit das Merkmal zu dieser Klasse gehört. Die Zuordnung erfolgt dann basierend auf der höchsten Wahrscheinlichkeit. [37].

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{1}$$

Im Ansatz von [36] wird der Klassifikator mit AdaBoost optimiert, einer Abkürzung für adaptive Boosting. AdaBoost kombiniert mehrere schwache Klassifikatoren zu einem starken Klassifikator. Dabei werden iterativ mehrere Klassifikatoren hinzugefügt und der Datensatz stetig neu gewichtet, damit sich der nächste Klassifikator auf die Fehler des vorherigen Klassifikators konzentriert [35].

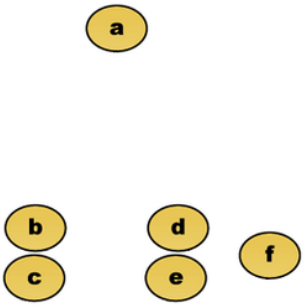
In der neuen Klassifikationsmethode von [36] wird der Trainingsdatensatz zuerst in eine bestimmte Anzahl von Teildatensätzen aufgeteilt. Anschließend werden die einzelnen Teildatensätze schrittweise verarbeitet. Jeder Teildatensatz enthält eine Menge an Datentupeln, also eine Menge an Wörtern oder Sätzen. Zu Beginn erhält jedes Tupel die gleiche Gewichtung. Dann wird das hybride Klassifizierungsmodell aus Naive Bayes und AdaBoost mit dem Teildatensatz trainiert. Anschließend werden die Gewichte der Tupel aktualisiert, je nachdem ob sie richtig oder falsch klassifiziert wurden. Nach den Durchgängen aller Teildatensätze ergibt sich ein Set an Klassifikator-Modellen mit jeweiligen Gewichtungen, die in Kombination genaue Vorhersagen für Klassen treffen.

Kaur, Zandu [36] zeigten, dass ihre vorgeschlagene Methode mit 94,2529% Genauigkeit deutlich besser klassifiziert als der k-NN-Algorithmus, der nur eine Genauigkeit von 51,7241% hatte.

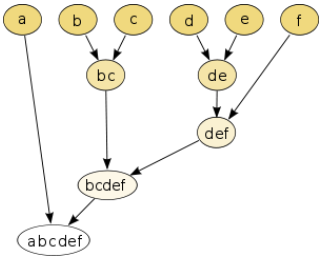
c) Clusteranalyse: Die Clusteranalyse ist eine Datenanalysetechnik aus dem Bereich des maschinellen Lernens. Clustering ist eine unüberwachte Lernmethode, die Muster in Eingabedaten ohne vordefinierte Zielwerte erkennt. Bei der Clusteranalyse werden unsortierte Informationen auf der Grundlage von Ähnlichkeiten, Mustern und Unterschieden gruppiert, ohne dass die Daten zuvor trainiert wurden. Im Kontext der Klassifizierung sensibler Daten wird die Clusteranalyse auf die Daten in einem Unternehmen angewendet. Die resultierenden Cluster enthalten dann ähnliche Dokumente nach einer Ähnlichkeitsmetrik wie dem euklidischen Abstand oder der Kosinusähnlichkeit [38]. Um die Clusteranalyse als Klassifikator zu nutzen, müssen vor der Analyse relevante Merkmale oder Attribute definiert werden, um sensible Daten zu identifizieren. Anschließend können die Cluster anhand der Merkmale klassifiziert werden.

Zwei gängige Clustering-Methoden sind das hierarchische Clustering und das partitionierende Clustering. Hierarchisches Clustering wird in der Regel für das Clustering von Texten verwendet, wobei jedes Dokument auf der Grundlage seiner Ähnlichkeit schrittweise in einen vordefinierten Cluster zusammengeführt wird. Durch diesen Prozess entsteht eine Clusterhierarchie, die als Baumstruktur, das sogenannte Dendrogramm, dargestellt werden kann. Abbildung 3a zeigt einen Beispieldatensatz mit den Objekten 'a' bis 'f'. In diesem Datensatz liegen die Objekte 'b' und 'c' sowie 'd' und 'e' sehr nahe beieinander. Der Clustering-Algorithmus fasst nach und nach die Objekte mit dem geringsten Abstand zusammen, gefolgt von den nächstgelegenen Objekten oder Clustern, bis der gesamte Datensatz gruppiert ist. So entsteht ein Baum wie in Abbildung 3b, bei dem die Blätter Cluster darstellen, die nur ein einzelnes Objekt aus dem Datensatz enthalten, und die Wurzel einen einzelnen Cluster, der alle Objekte enthält. Die Kanten zwischen den Knoten haben außerdem ein Attribut, das den Abstand zwischen den beiden Clustern angibt. Je nach gewünschter Anzahl von Clustern können die Cluster auf einer bestimmten Ebene des Baumes verwendet werden [38].

Wegen seiner Einfachheit und Flexibilität wird hierarchisches Clustering häufig eingesetzt und bietet den Vorteil, dass



(a) Datensatz. Quelle: In Anlehnung an [39].



(b) Dendrogramm. Quelle: In Anlehnung an [39].

jede Art von Ähnlichkeitsmessung durchgeführt werden kann. Dieses Verfahren bietet außerdem eine detaillierte Darstellung der Clusterstruktur, wodurch unterschiedliche Granularitätsstufen von Clustern untersucht werden können.

Beim partitionierenden Clustering wird ein Datensatz in eine bestimmte Anzahl von Clustern eingeteilt. Jeder Datenpunkt gehört zu einem bestimmten Cluster und das Ziel besteht darin, möglichst viele Datenpunkte in die Cluster zu verteilen und dabei die Ähnlichkeit zwischen den Clustern zu minimieren. Das am weitesten verbreitete Verfahren ist der k-means-Algorithmus. Das 'k' steht für die Anzahl an zu definierenden Clustern und 'means' für den Mittelwert bzw. das Zentrum des Clusters. Zu Beginn muss die Anzahl der Cluster bestimmt werden. Dies kann sich zum Beispiel daran orientieren, in welche Vertraulichkeitsstufen die Daten eingeteilt werden sollen. Anschließend werden für die Cluster initial jeweils zufällig Cluster-Mittelpunkte, auch Centroids genannt, gewählt. Danach wird für jeden Datenpunkt der Abstand zwischen dem Punkt und den Cluster-Centroids berechnet. Der Punkt wird dem jeweiligen Cluster zugeordnet, welcher am nächsten ist und die Cluster sind initial befüllt. Die nächsten Schritte wiederholen sich, bis sich die Cluster nicht mehr ändern. Zuerst wird für jedes Cluster aus den Datenpunkten ein neuer Mittelwert bestimmt, der den neuen Centroid darstellt. Dann werden alle Datenpunkte anhand ihrer Distanzen zu den neuen Zentren neu zugeordnet [38].

Der k-mean Algorithmus ist beliebt, da er einfach ist, nur eine kleine Anzahl an Iterationen benötigt und parallel berechnet werden kann. Allerdings ist das Ergebnis des Algorithmus stark von der Wahl des 'k' und der initialen Cluster abhängig [38]. Durch die automatische Erkennung von Mustern und Ähnlichkeiten in den Daten, konzentriert sich die Clusteranalyse nicht nur auf den Inhalt, sondern berücksichtigt auch den Kontext.

d) CASSED: Kužina, Petric et al. [3] sahen eine große Herausforderung darin, sensible Daten in strukturierten Datenbanken zu klassifizieren. Das Problem besteht darin, die einzelnen Spalten einer Datenbanktabelle zu durchsuchen und zu bestimmten, ob sie sensible Daten enthalten und welche Arten sensibler Daten vorhanden sind. Hierbei ist es erforderlich, den Inhalt der Tabellenzelle zu interpretieren und den Kontext der umgebenden Zellen zu berücksichtigen. Frühere

Ansätze nutzten dafür stark regelbasierte Methoden, deren Grenzen bei einer Vielzahl verschiedener Datentypen schnell erreicht wurden. Zudem konnten sie nur begrenzt Kontext und Semantik einbeziehen. Um dieses Problem zu bewältigen, entwickelten Kužina, Petric et al. [3] eine neue Methode namens 'Context-based Approach for Structured Sensitive Data Detection' (CASSED). Dabei wird durch die Kombination von Spaltenmetadaten und Zellwerten ein Spaltenkontext hergestellt, der dann in einen einzelnen Input-Vektor umgewandelt wird. Dieser Input-Vektor wird anschließend zur Klassifizierung durch das BERT-Modell verwendet.

BERT, eine Abkürzung für 'Bidirectional Encoder Representations from Transformers', ist ein Open-Source Framework, das von Google entwickelt wurde, um Transformer-basierte Natural-Language-Processing-Modelle zu erstellen. Diese Modelle sind darauf spezialisiert, kontextuelle Zusammenhänge und Beziehungen zwischen Wörtern in Texten zu erfassen. Transformer-basierte Modelle verwenden einen Selbstaufmerksamkeitsmechanismus, der die Beziehung jedes Worts zu allen anderen Wörtern in einem Satz bestimmt. Sie bestehen aus mehreren Encoder- und Decoder-Schichten, die einen Text lesen und versuchen, das nächste Wort vorherzusagen. Außerdem enthalten sie vollständig vernetzte neuronale Netze. BERT nutzt den Transformer-Mechanismus mit ausschließlich Encoder-Schichten. Ein charakteristisches Merkmal von BERT ist seine bidirektionale Verarbeitung von Texteingaben. Das bedeutet, dass Sequenzen von Wörtern sowohl von Anfang als auch von Ende her analysiert werden, um ein verbessertes Verständnis für die kontextuelle Beziehung zwischen den Wörtern zu gewinnen.

Im ersten Schritt werden die Spalten in Input-Vektoren umgewandelt. Hierbei repräsentiert jeder Input-Vektor die Spaltenüberschrift zusammen mit mehreren Zellwerten derselben Spalte als Tokens, die durch Trennzeichen getrennt sind. In Abbildung 4 ist eine Umwandlung einer Spalte in einen Input-Vektor veranschaulicht. Die Spaltenüberschrift wird dabei mit einem Punkt von der ersten Zelle getrennt, während die Zellen untereinander durch Kommata separiert sind. Diese Darstellung liefert dem Modell zusätzliche Informationen darüber, dass diese Werte unterschiedlich behandelt werden sollten. Es ist zu beachten, dass BERT maximal 512 Tokens als Input-Vektor verarbeiten kann. Daher müssen größere Input-Vektoren aufgeteilt werden.

Für die Klassifizierung generiert der Decoder von BERT für jede potenzielle Klasse eine nicht normalisierte Vorhersage. Diese Vorhersagen werden dann über alle Teile der Spalte gemittelt, in die die Spalte aufgeteilt wurde. Anschließend wird eine Sigmoidfunktion auf die Vorhersagewerte angewendet, um normalisierte Wahrscheinlichkeiten für jede Klasse zu erzeugen.

Der CASSED-Ansatz enthält neben dem BERT-Modell auch eine regelbasierte Schicht, die strukturierte Formate wie E-Mails oder Sozialversicherungsnummern mittels regulären Ausdrücken klassifiziert. Zudem wird ein Wörterbuch für bekannte sensible Daten oder Merkmale von Geschäftsgeheimnissen verwendet. Auch diese Schicht ermittelt für jede

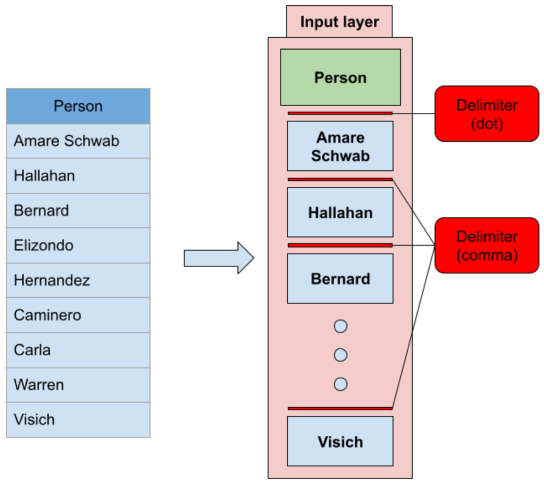


Abbildung 4: Beispiel einer Umwandlung einer Spalte in einen Input-Vektor. Quelle: [3].

Klasse eine mögliche Wahrscheinlichkeit. Die Wahrscheinlichkeiten der regelbasierten Schicht und der BERT-Schicht werden kombiniert, um eine Gesamtwahrscheinlichkeit pro Klasse zu erhalten. Anschließend werden den Daten jene Klassen zugeteilt, deren Wahrscheinlichkeit einen festgelegten Schwellwert überschreitet. Es besteht somit die Möglichkeit, dass einer Datenbankspalte mehrere Klassen zugeordnet wird. Die Architektur des CASSED-Ansatzes wird in 5 dargestellt.

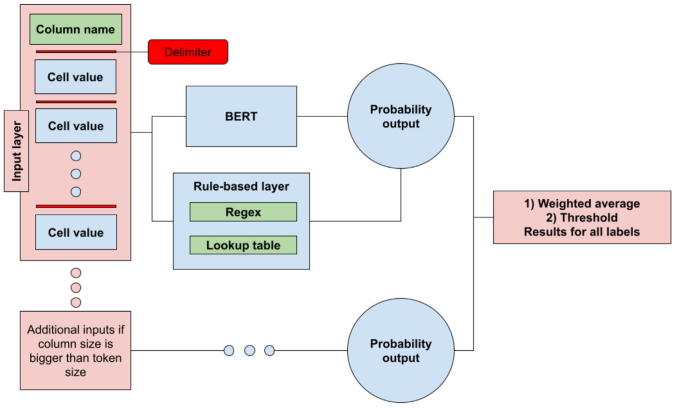


Abbildung 5: Überblick über die CASSED Methode. Quelle: [3].

Die CASSED-Methode zeigt im Vergleich zu anderen kontextbasierten Klassifizierungsmethoden deutlich bessere Ergebnisse. Das Modell erzielt nicht nur bessere Leistungen durch den alleinigen Einsatz von BERT, sondern wird durch die Integration der regelbasierten Schicht noch präziser.

Guo, Liu et al. [20] verfolgen einen ähnlichen Ansatz mit ihrem Modell 'Exsense'. Auch sie verwenden zwei Schichten: eine inhaltsbasierte Analyse mit regulären Ausdrücken und eine kontextbasierte Analyse mit einem BERT-BiLSTM-Attention-Modell. Dabei wird das BERT-Modell kombiniert mit einem Bidirectional Long Short-Term Memory Modell (BiLSTM). BiLSTM ist ein rekurrentes neuronales Netzwerk,

das für die Verarbeitung von sequenziellen Daten entwickelt wurde und sich gut zur Erfassung von Abhängigkeiten in langen Sequenzen eignet. Im BERT-BiLSTM-Modell dient die bidirektionale Kontextrepräsentation von BERT als Eingabe für das BiLSTM. Dieser Ansatz zeigt ebenfalls herausragende Ergebnisse bei der Klassifizierung sensibler Daten unter Berücksichtigung des Kontextes.

Insgesamt zeigen alle vorgestellten Klassifizierungsmethoden aus dem KI-Bereich sehr gute Ergebnisse bei der Klassifizierung von sensiblen Daten.

E. Anwendung in der Cloud Security

Wie in Kapitel III erläutert, spielt die Datenklassifikation eine entscheidende Rolle in DLP-Systemen. Das automatische Erkennen und Klassifizieren sensibler Daten ist ein zentraler Bestandteil dieser Systeme, da Sicherheitsmaßnahmen und -techniken auf der Kenntnis über die Sicherheitsstufen der Daten basieren. Ein Beispiel für ein solches Cloud-DLP-System ist der Google Cloud Data Loss Prevention Service. Dieser vollständig verwaltete Dienst ermöglicht Unternehmen die Erkennung, Klassifizierung und den Schutz sensibler Daten in Cloud- und lokalen Umgebungen. Der Dienst nutzt maschinelles Lernen und Mustervergleichstechniken, um bis zu 150 verschiedenen Datentypen zu identifizieren und entsprechend zu klassifizieren. Durch diese Klassifizierung können dann im Cloud-DLP-System unterschiedliche Schutzmaßnahmen für verschiedene Datentypen und Klassen festgelegt werden [40].

Ein weiteres Beispiel ist das Software-Produkt Digital Guardian Cloud Data Protection des Anbieters Fortra. Das Cloud-DLP-System umfasst verschiedene Module, wobei eines davon die Datenklassifizierung ist. Dieses Modul identifiziert und klassifiziert sensible Daten und hängt ihnen verschiedene Sicherheitsstufen-Labels an. Die Klassifizierung erfolgt dabei automatisch auf Grundlage von Inhalt und Kontext. Zudem besteht die Möglichkeit, eine manuelle Klassifizierung durch den Benutzer vorzunehmen. Das Cloud-DLP-System von Fortra ist dabei mit allen führenden Cloud-Plattform-Anbietern kompatibel [41].

Die Klassifizierung und die Markierung der Daten dient in Cloud-DLP-Systemen anschließend dazu, geeignete Schutzmaßnahmen zu definieren. Im Folgenden werden verschiedene Techniken beschrieben, die eine versehentliche Offenlegung von sensiblen Daten verhindern.

Eine der ersten Techniken zum Schutz von sensiblen Daten ist die Wasserzeichen-Methode. Dabei werden spezifische Identifikationsmerkmale oder Muster durch einen Algorithmus in Daten eingebettet, um die Quelle oder den Eigentümer zu kennzeichnen. Wasserzeichen dienen dazu, den Ursprung von Daten zu verfolgen und zu überwachen, um unbefugte Verbreitung oder Weitergabe zu erkennen. Systeme können anhand der eingebetteten Wasserzeichen erkennen, wenn Daten ohne Berechtigung weitergegeben oder veröffentlicht werden, und Maßnahmen ergreifen. Auf diese Weise können Datenlecks und -manipulationen mithilfe von Wasserzeichen erkannt und verhindert werden [42].

Eine weit verbreitete Methode in DLP-Systemen ist das Blacklisting oder Whitelisting. Das Modell besteht aus Regeln, die Muster böswilliger Aktivitäten oder markierte sensible Daten definieren, die das Unternehmensnetzwerk nicht verlassen dürfen. Beim Whitelisting wird definiert, welche Daten geteilt werden dürfen, beim Blacklisting werden die Daten und Muster definiert, die nicht geteilt werden dürfen. So kann ein DLP-System anhand dieser Regeln und Listen Transaktionen blockieren und somit eine versehentliche Offenlegung von sensiblen Daten verhindern. Allerdings ist es für böswillige Insider oft leicht, eine Blacklisting-basierte Erkennung zu umgehen, da je nach Berechtigungen die Regeln und Listen bekannt sind [43].

Die Kryptografie ist eine häufig verwendete Technik im Datenschutz. Sie wandelt Informationen von einem lesbaren Format in ein verschlüsseltes Format um. Je nach Sicherheitsstufe der Daten können unterschiedliche Verschlüsselungsalgorithmen und kryptografische Funktionen zum Einsatz kommen [16]. Dabei ist es nicht sinnvoll, einfach alle vorhandenen Daten verschlüsselt zu speichern oder zu übertragen. Insbesondere bei großen Datenmengen kann das schnell zu Performance-Problemen führen. Die Verschlüsselung von Daten erfordert Rechenressourcen und erhöhten Speicherbedarf. Deshalb wird meistens nur nach Bedarf verschlüsselt [44]. Zardari, Jung et al. [32] kategorisieren in ihrem vorgeschlagenen Modell Daten nach sensibel und nicht-sensibel. Nicht-sensible Daten werden direkt im Cloud-Speicher gespeichert, sensible Daten werden vorher mit dem RSA-Algorithmus verschlüsselt.

Das Fingerprinting-Verfahren, das im Kapitel IV-C0b zur automatischen Klassifizierung beschrieben wurde, kann auch als Schutzmaßnahme dienen. Mit einem Vergleich von Datei- oder Daten-Hashes kann der ausgehende Datenverkehr überwacht werden. Ein DLP-System gleicht dabei die Hashes der Daten im Datenverkehr mit denen der als sensibel markierten Daten ab und kann so bei Übereinstimmung ein Datenleck erkennen und verhindern. Auch hier ist die Wahl des Hashing-Verfahrens relevant, wie verlässlich auch veränderte Daten noch erkannt werden [16].

Zhang, Jing et al. [45] stellen eine Methode vor, um die versehentliche Offenlegung von sensiblen Daten zu verhindern, selbst wenn bspw. ein ungeschützter Laptop verloren geht. Der Ansatz namens 'Cloud Shredder' sieht vor, dass vertrauliche Dateien geschreddert, also in Stücke geteilt wird. Ein Teil wird auf dem physischen Gerät und ein Teil im Cloud-Speicher gespeichert. Nur wer Zugriff auf beide Bereiche hat, kann die sensiblen Dateien lesen oder verarbeiten.

Ein weiterer Ansatz namens 'DocGuard' wurde von Gui, Puzis et al. in [46] vorgestellt. Die Idee besteht darin, dass bereits vorhandene Antiviren-Software durchgesickerte Daten identifiziert und den Zugriff blockiert. Das funktioniert, indem DocGuard bei sensiblen Daten eine versteckte Signatur einer bekannten Schadsoftware einfügt. Wenn diese Daten das Unternehmensnetzwerk verlassen, erkennt eine Antiviren-Software die schadhafte Signatur als Bedrohung und greift ein.

In Unternehmen werden externe Cloud-Dienste wie Google Docs vermehrt genutzt. Dabei können Benutzer schnell ge-

gen unternehmensinterne Daten-Richtlinien verstoßen. Sensible Daten sind leicht aus einem geschützten Cloud-Dokument in ein ungeschütztes Cloud-Dokument kopiert und damit offengelegt. Papagiannis, Watcharapichat et al. [47] schlagen als Lösung für dieses Problem ihren Ansatz namens 'BrowserFlow' vor. BrowserFlow ist eine browserbasierte Middleware, die Datenflüsse verfolgt, indem sie die Ähnlichkeit zwischen Textfragmenten erkennt. Anhand von gekennzeichneten Daten können so nicht autorisierte Datenflüsse identifiziert und verhindert werden.

Es zeigt sich also, dass es viele verschiedene einfache und aufwändigere Methoden in der Cloud Security gibt, um sensible Daten zu schützen. Die Zuverlässigkeit dieser Methoden basieren auf der Qualität der Klassifizierung. Deshalb ist eine automatische Klassifizierung mit bewährten KI-Verfahren eine gute Wahl und immer häufiger in DLP-Systemen im Einsatz.

V. Ausblick

Diese Arbeit gibt einen Überblick über verschiedene Ansätze im Bereich der Cloud Data Leakage Prevention unter Verwendung von automatischen Datenklassifizierungsmethoden. Zu Beginn wurde die Problematik der versehentlichen Offenlegung von Cloud-Daten analysiert. Anschließend wurden verschiedene Ansätze zur Prävention von Datenlecks in der Cloud vorgestellt, wobei der Schwerpunkt auf der Nutzung von Cloud-DLP-Systemen lag. Dabei hat sich herausgestellt, dass die Erkennung sensibler Daten von verschiedenen Datenmerkmalen beeinflusst wird, darunter Kontext, Kategorien, Vertraulichkeitsstufen, Struktur und Zustände. Im Fokus stand die automatische Datenklassifizierung, die als entscheidendes Element für den Schutz sensibler Informationen betrachtet wurde. Hierbei wurden verschiedene Methoden wie die Klassifizierung mit manueller Definition betrachtet, die regelbasierte Methoden und Data Fingerprinting umfasst. Zudem wurde die Anwendung von KI-basierten Methoden und maschinellem Lernen in der automatischen Datenklassifizierung untersucht, wobei Algorithmen wie k-NN, Boosting, Clusteranalyse und ein kontextbasierter Ansatz für die Erkennung strukturierter sensibler Daten betrachtet wurden. Mit der reinen Klassifizierung von sensiblen Daten sind diese jedoch noch nicht geschützt. Deshalb wurden im letzten Kapitel noch einige Schutzmaßnahmen betrachtet, die auf klassifizierten Daten basieren.

Besonders die manuelle Datenklassifizierung und auch die Klassifizierung mit manueller Definition werden zunehmend risikobehaftet. Bei der Anwendung manueller Sicherheitsrichtlinien besteht die Gefahr, dass nicht alle relevanten und sensiblen Daten einbezogen werden, und die Richtlinien möglicherweise nicht kontinuierlich aktualisiert werden. Des Weiteren erfordert der manuelle Aufwand, dass geschultes Sicherheitspersonal diese Aufgaben übernimmt. Allerdings verfügen diese Mitarbeiter nicht über umfassendes Wissen über sämtliche Daten innerhalb des Unternehmens. Die Literaturrecherche verdeutlicht, dass die automatische Datenklassifizierung bereits erfolgreich angewendet wurde und aktuell erforscht wird, um die Klassifizierung zu verbessern und effizienter zu gestalten. Insbesondere der verstärkte Einsatz von Cloud-Technologien stellt die Data Leakage Prevention vor neue Herausforderungen. Die Verteilung von Daten auf verschiedene Medien, der schnelle Zugriff und die großen Datenmengen sind nur einige davon. Trotz der herausragenden Ergebnisse des Einsatzes von Methoden aus dem KI-Bereich stehen Unternehmen damit vor neue Herausforderungen. Die Entwicklung eines maschinellen Lernmodells oder eines neuronalen Netzes zur Erkennung sensibler Daten hängt stark von der Verfügbarkeit realer Datensätze mit vertraulichen Daten ab. Allerdings sind solche Datensätze nicht immer für die Öffentlichkeit zugänglich und könnten zudem nicht alle notwendigen Eigenschaften für ein Unternehmen enthalten. In datenschutzrelevanten Bereichen werden daher häufig synthetische Datensätze verwendet. Diese Datensätze können sowohl für das Training des Modells als auch zur Erweiterung eines realen Datensatzes verwendet werden [3].

Die fortlaufende Entwicklung im Bereich der Cloud Data Leakage Prevention und automatischen Datenklassifizierung bietet vielversprechende Perspektiven. Die erörterten Ansätze bieten einen umfassenden Überblick über Möglichkeiten zur Verhinderung von Datenlecks in der Cloud, wobei die automatische Datenklassifizierung als zentrales Element für den Schutz sensibler Informationen herausgestellt wurde. Es wird deutlich, dass die Herausforderungen im Zusammenhang mit der verstärkten Nutzung von Cloud-Technologien und der Datenverteilung auf verschiedene Medien weiterhin im Fokus stehen. Die Anwendung von KI und maschinellem Lernen zeigt vielversprechende Ergebnisse. Für die Zukunft der Forschung und Implementierung in diesem Bereich könnten weiterführende Studien den Fokus auf die Optimierung von automatischen Datenklassifizierungsmethoden legen, um die Effizienz und Genauigkeit zu steigern. Zudem könnten innovative Ansätze zur Integration von Schutzmaßnahmen auf Basis der klassifizierten Daten weiter erforscht werden, um einen umfassenden Schutz sensibler Informationen in der Cloud zu gewährleisten.

Literatur

[1] KPMG, "Nutzung von Cloud Computing in Unternehmen in Deutschland in den Jahren 2011 bis 2022." https://de.statista.com/statistik/daten/studie/177484/umfrage/einsatz-von-cloud-computing-in-deutschen-unternehmen-2011/, 2022. Letzter Zugriff: 07.12.2023.

[2] C. Surianarayanan and P. R. Chelliah, "Introduction to Cloud Computing," in *Essentials of Cloud Computing*, (Cham), pp. 1–38, Springer International Publishing and Imprint: Springer, 2023.

[3] V. Kužina, A.-M. Petric, M. Barišić, and A. Jović, "CASSED: Context-based Approach for Structured Sensitive Data Detection," *Expert Systems With Applications*, vol. 223, p. 119924, 2023.

[4] S. Alneyadi, E. Sithirasenan, and V. Muthukkumarasamy, "A survey on data leakage prevention systems," *Journal of Network and Computer Applications*, vol. 62, pp. 137–152, 2016.

[5] CA Technologies, "2018 Insider Threat Report." https://crowdresearchpartners.com/wp-content/uploads/2017/07/Insider-Threat-Report-2018.pdf, 2018. Letzer Zugriff: 07.12.2023.

[6] R. Chugh and A. Bales, "Marktführer für Data Loss Prevention." https://www.gartner.com/doc/reprints?id=1-2EYVL2C2\&ct=230912\&st=sb, 2023. Letzter Zugriff: 07.12.2023.

[7] Cloud Security Alliance, "Top Threats to Cloud Computing Pandemic Eleven." https://cloudsecurityalliance.org/artifacts/top-threats-to-cloud-computing-pandemic-eleven, 2022. Letzter Zugriff: 07.12.2023.

[8] Z. Whittaker, "Toyota confirms another years-long data leak, this time exposing at least 260,000 car owners." https://techcrunch.com/2023/05/31/toyota-customer-data-leak-years/, 2023. Letzter Zugriff: 07.12.2023.

[9] PSNI, "Police Service of Northern Ireland statement on data breach | PSNI." https://www.psni.police.uk/latest-news/police-service-northern-ireland-statement-data-breach, 2023. Letzter Zugriff: 07.12.2023.

[10] S. Trabelsi, "Monitoring leaked confidential data," in *2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pp. 1–5, IEEE, 2019.

[11] T. Brindha and R. S. Shaji, "An analysis of data leakage and prevention techniques in cloud environment," in *2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 350–355, IEEE, 2015.

[12] Proofpoint, "DLP - Data Loss Prevention: Schutz vor Datenverlust | Proofpoint DE." https://www.proofpoint.com/de/threat-reference/dlp, 2021. Letzter Zugriff: 07.12.2023.

[13] V. Monev, "Data Leakage Prevention in ISO 27001: Compliance and Implementation," in *2023 International Conference on Information Technologies (InfoTech)*, pp. 1–5, IEEE, 2023.

[14] NIST, "Framework for Improving Critical Infrastructure Cybersecurity." https://www.nist.gov/system/files/documents/cyberframework/cybersecurity-framework-021214.pdf, 2014. Letzter Zugriff: 07.12.2023.

[15] M. E. Hussain and R. Hussain, "Cloud Security as a Service Using Data Loss Prevention: Challenges and Solution," in *International Conference on Internet of Things and Connected Technologies*, (Cham), pp. 98–106, Springer, 2021.

[16] I. Herrera Montano, J. J. García Aranda, J. Ramos Diaz, S. Molina Cardín, I. de La Torre Díez, and J. J. P. C. Rodrigues, "Survey of Techniques on Data Leakage Protection and Methods to address the Insider threat," *Cluster Computing*, vol. 25, no. 6, pp. 4289–4302, 2022.

[17] B. S. Shishodia and M. J. Nene, "Data Leakage Prevention System for Internal Security," in *2022 International Conference on Futuristic Technologies (INCOFT)*, pp. 1–6, IEEE, 2022.

[18] Gartner, "Umsatz mit Software-as-a-Service (SaaS) weltweit von 2010 bis 2022 und Prognose bis 2024 (in Milliarden US-Dollar)." https://de.statista.com/statistik/daten/studie/194117/umfrage/umsatz-mit-software-as-a-service-weltweit-seit-2010/, 2023. Letzter Zugriff: 08.12.2023.

[19] B. Hauer, "Data and information leakage prevention within the scope of information security," *IEEE Access*, vol. 3, pp. 2554–2565, 2015.

[20] Y. Guo, J. Liu, W. Tang, and C. Huang, "Exsense: Extract sensitive information from unstructured data," *Computers & Security*, vol. 102, p. 102156, 2021.

[21] A. Pogiatzis and G. Samakovitis, "Using BiLSTM Networks for Context-Aware Deep Sensitivity Labelling on Conversational Data," *Applied Sciences*, vol. 10, no. 24, p. 8924, 2020.

[22] C. E. Landwehr, C. L. Heitmeyer, and J. McLean, "A security model for military message systems," *ACM Transactions on Computer Systems*, vol. 2, no. 3, pp. 198–222, 1984.

[23] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919*, 2017.

[24] A. Shabtai, Y. Elovici, L. Rokach, A. Shabtai, Y. Elovici, and L. Rokach, "A taxonomy of data leakage prevention solutions," *A survey of data leakage detection and prevention solutions*, pp. 11–15, 2012.

[25] S. Divadari, J. Surya Prasad, and P. Honnavalli, "Managing Data Protection and Privacy on Cloud," in *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2022*, pp. 383–396, Springer, 2023.

[26] M. A. Alsuwaie, B. Habibnia, and P. Gladyshev, "Data Leakage Prevention Adoption Model & DLP Maturity Level Assessment," in *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, pp. 396–405, IEEE, 2021.

[27] J. Venhorst, "Warum eine manuelle Datenklassifizierung nicht sinnvoll ist." https://www.computerweekly.com/de/meinung/Warum-eine-manuelle-Datenklassifizierung-nicht-sinnvoll-ist, 2019. Letzer Zugriff: 08.12.2023.

[28] D. Gugelmann, P. Studerus, V. Lenders, and B. Ager, "Can content-based data loss prevention solutions prevent data leakage in web traffic?," *IEEE Security & Privacy*, vol. 13, no. 4, pp. 52–59, 2015.

[29] Y. J. Ong, M. Qiao, R. Routray, and R. Raphael, "Context-aware data loss prevention for cloud storage services," in *2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*, pp. 399–406, IEEE, 2017.

[30] X. Shu, D. Yao, and E. Bertino, "Privacy-preserving detection of sensitive data exposure," *IEEE transactions on information forensics and security*, vol. 10, no. 5, pp. 1092–1103, 2015.

[31] U. Kamath, J. Liu, and J. Whitaker, *Deep learning for NLP and speech recognition*. Cham: Springer, 2019.

[32] M. A. Zardari, L. T. Jung, and N. Zakaria, "K-nn classifier for data confidentiality in cloud computing," in *2014 International Conference on Computer and Information Sciences (ICCOINS)*, pp. 1–6, ieee, 2014.

[33] J. Frochte, "Maschinelles lernen – überblick und abgrenzung," in *Maschinelles Lernen* (J. Frochte, ed.), pp. 13–31, München: Carl Hanser Verlag GmbH & Co. KG, 2018.

[34] R. Shah, "Introduction to k-Nearest Neighbors (kNN) Algorithm: A Powerful Supervised Machine Learning Algorithm." https://ai.plainenglish.io/introduction-to-k-nearest-neighbors-knn-algorithm-e8617a448fa8, 2021. Letzter Zugriff: 08.12.2023.

[35] J. Frochte, "Entscheidungsbäume," in *Maschinelles Lernen* (J. Frochte, ed.), pp. 117–160, München: Carl Hanser Verlag GmbH & Co. KG, 2018.

[36] K. Kaur and V. Zandu, "A secure data classification model in cloud computing using machine learning approach," *International Journal of Grid and Distributed Computing*, vol. 9, no. 8, pp. 13–22, 2016.

[37] J. Frochte, "Statistische grundlagen und bayes-klassifikator," in *Maschinelles Lernen* (J. Frochte, ed.), pp. 68–87, München: Carl Hanser Verlag GmbH & Co. KG, 2018.

[38] H. Suyal, A. Panwar, and A. S. Negi, "Text clustering algorithms: a review," *International Journal of Computer Applications*, vol. 96, no. 24, pp. 36–40, 2014.

[39] H. Bonthu, "Single-Link Hierarchical Clustering Clearly Explained!." https://www.analyticsvidhya.com/blog/2021/06/single-link-hierarchical-clustering-clearly-explained/, 2023. Letzter Zugriff: 08.12.2023.

[40] Google Cloud, "Cloud Data Loss Prevention | Google Cloud." https://cloud.google.com/dlp?hl=de, 2023. Letzter Zugriff: 08.12.2023.

[41] Digital Guardian, "Enterprise IP & DLP Software." https://www.digitalguardian.com/, 2023. Letzter Zugriff: 08.12.2023.

[42] R. Naik and M. N. Gaonkar, "Data leakage detection in cloud using watermarking technique," in *2019 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6, IEEE, 2019.

[43] E. Costante, D. Fauri, S. Etalle, J. Den Hartog, and N. Zannone, "A hybrid framework for data loss prevention and detection," in *2016 IEEE security and privacy workshops (SPW)*, pp. 324–333, IEEE, 2016.

[44] O. Arki, A. Zitouni, and M. Djoudi, "A security method for cloud storage using data classification," *International Journal of Grid and High Performance Computing (IJGHPC)*, vol. 15, no. 1, pp. 1–17, 2023.

[45] N. Zhang, J. Jing, and P. Liu, "Cloud shredder: Removing the laptop on-road data disclosure threat in the cloud computing era," in *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 1592–1599, IEEE, 2011.

[46] M. Guri, R. Puzis, K.-K. R. Choo, S. Rubinshtein, G. Kedma, and Y. Elovici, "Using malware for the greater good: Mitigating data leakage," *Journal of Network and Computer Applications*, vol. 145, p. 102405, 2019.

[47] I. Papagiannis, P. Watcharapichat, D. Muthukumaran, and P. Pietzuch, "Browserflow: Imprecise data flow tracking to prevent accidental data disclosure," in *Proceedings of the 17th International Middleware Conference*, pp. 1–13, 2016.