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*—Abstract**
***Index Terms*—component, formatting, style, styling, insert**

I. Einführung

Im Jahr 2022 gaben bereits 84% der befragten 552 Unternehmen in Deutschland an, dass sie Cloud-Dienste in ihrem Unternehmen einsetzen [1]. Cloud Computing hat sich in der Zeit der digitalen Transformation zu einem wichtigen Bestandteil der Informationsverarbeitung entwickelt. Die Nutzung von Cloud-Diensten wird immer beliebter, da sie die Möglichkeit zur effizienten Speicherung großer Datenmengen, schnellen Zugang zu Ressourcen und nahtlosen Datenaustausch bietet. Durch den Verbreitung von digitaler Technologie in der Gesellschaft und in Unternehmen werden immer mehr Daten geteilt und gesammelt. Um diese großen Datenmengen sammeln und verarbeiten zu können, nutzen Unternehmen die Vorteile von Cloud-Diensten. Die Möglichkeit, Daten in Echtzeit zu teilen, verbessert Geschäftsprozesse und erleichtert die Zusammenarbeit im Unternehmen [2].

Da Informationen das wertvollste Gut eines Unternehmens sind, ist ihr Schutz von größter Bedeutung. Beim Sammeln von Daten ist ein Unternehmen zudem verpflichtet, sie vor Diebstahl, Verlust und Missbrauch zu schützen. Es gibt zahlreiche Datenschutzgesetze und -vorschriften, wie die EU-Datenschutz-Grundverordnung (DSGVO), um sensible Daten wie personenbezogene Daten zu schützen. Ziel dieser Vorschriften ist es, strengen Regeln für das Sammeln von Daten vorzugeben und der Einzelperson eine vergleichsweise hohe Kontrolle über ihre personenbezogenen Daten zu geben [3]. Unabhängig des Speicherorts besteht also das Risiko, dass die Datensicherheit verletzt wird.

Eines der Hauptziele der Informationssicherheit ist die Verhinderung der Offenlegung von Daten gegenüber Unbefugten. Datenlecks können jedoch aufgrund der Notwendigkeit, auf Informationen zuzugreifen, diese zu teilen und zu nutzen, nicht immer verhindert werden. Diese Bedrohung kann von böswilligen Außenstehenden ausgehen, die versuchen, sensible Daten zu erhalten. Umgekehrt können auch interne Mitarbeiter eine Gefahr darstellen, wenn sie beabsichtigt oder unbeabsichtigt Informationen preisgeben [4]. Bereits im Jahr 2018 haben Studien gezeigt, dass 53% der befragten Unternehmen Insider-Angriffe in den letzten 12 Monaten bestätigten. Dabei sind Bedrohungen von innen häufig schwerwiegender als von außen, da sie meist 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 immer häufiger 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 häufig große Datenmengen, ohne zu wissen, was erfasst wird oder wie sie nach personenbezogenen Daten suchen oder diese abrufen können. Das erschwert den Schutz der Privatsphäre. DLP-Systeme benötigen die Information, 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 künstliche 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 mit dem Einsatz von KI im Bereich Datenschutz liegt der Fokus in dieser Arbeit auf der Anwendung von Methoden der automatischen Datenklassifizierung zur Erkennung sensibler Informationen, um den Schutz sensibler Daten zu gewährleisten. Diese Arbeit beschäftigt sich mit der Frage, wie sensible Daten in große Datenmengen am besten erkannt werden können. Dabei 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. Accidental Cloud Data Disclosure

Die Cloud Security Alliance (CSA) veröffentlicht jährlich einen Bericht über die größten Bedrohungen der Cloud Security. Dabei werden über 700 Experten zu verschiedenen

Themen in der Cloud Security befragt. Ziel dieses Berichts ist es, auf Bedrohungen, Risiken und Schwachstellen in der Cloud aufmerksam zu machen. Der aktuellste Bericht von 2022 zeigt, dass sich die Verantwortung bezüglich der Sicherheit in der Cloud weg vom Cloud Service Provider und hin zum Cloud-Kunden bewegt [7]. Durch die Verlagerung der Verantwortung und die Komplexität der Cloud steigt das Risiko von Fehlern durch Unwissen. Deshalb ist das achte Sicherheitsproblem des Berichts 'Accidental Cloud Data Disclosure'.

Die versehentliche Offenlegung von Cloud Daten ist eine Sicherheitsbedrohung, die auftritt, wenn sensible Informationen unbeabsichtigt öffentlich zugänglich gemacht werden. Dies kann durch menschliches Versagen, Konfigurationsfehler oder unzureichende Sicherheitsmaßnahmen verursacht werden [7]. So wurde beispielsweise 2023 bekannt, dass bei Toyota Motor die persönlichen Daten von Kunden über mehrere Jahre offengelegt wurden. Der Grund war eine Fehlkonfiguration, wodurch die Datenbank in der Cloud öffentlich zugänglich war [8]. Im August 2023 wurden personenbezogene Daten der derzeitigen Beamten des nordirischen Polizeidienstes versehentlich von einem internen Mitarbeiter auf einer Online-Plattform veröffentlicht, der die Datei verwechselt hatte [9].

Allein die beiden Beispiele zeigen, wie schnell Fehler passieren und so großer Schaden angerichtet werden kann. Durch die Verlagerung der Sicherheits-Verantwortung auf die Cloud-Kunden steigt das Risiko von menschlichem Versagen. Durch Social Engineering und Phishing-Attacken können Mitarbeiter eines Unternehmens unbeabsichtigt sensible Daten wie Zugangsdaten offenlegen. Wie im Beispiel der Polizei in Nordirland kann es Mitarbeitern auch passieren, Daten unwissentlich zu veröffentlichen. Auch der Verlust von unzureichend geschützten Geräten wie Laptop oder Smartphone kann zu einer Daten Offenlegung führen. Der einfache Zugang zu Cloud-Ressourcen kann außerdem dazu verleiten, neue Ressourcen anzulegen oder Services zu nutzen, ohne sich über die nötigen Möglichkeiten der Absicherung zu informieren, wodurch Daten durch Fehlkonfigurationen offengelegt werden können. Doch nicht nur der Mensch ist ein Risiko, sondern auch das Zielsystem. Durch schwache Passwörter, fehlende Authentifizierung bei sicherheitsrelevanten Systemen und weitere Fehlkonfigurationen können Daten in der Cloud unwissentlich offengelegt werden. Aber auch ungeschlossene Sicherheitslücken in verwendeten Cloud-Services sind ein Sicherheitsrisiko [10] [11].

Um die Risiken für eine versehentliche Offenlegung von Daten zu minimieren, gibt es verschiedene Schutzmaßnahmen. Mit einem kontrollierten Identity Access Management (IAM) kann der interne und externe Zugriff auf die Daten geregelt und kontrolliert werden. Mit einer genauen Kontrolle der Zugriffsrechte auf die Cloud-Ressourcen können unbefugte Benutzer daran gehindert werden, auf sensible Daten zuzugreifen. Durch die Einführung von strengen Passwortrichtlinien und dem Einsatz von Passwort-Manager-Software kann das Risiko des unbefugten Zugriffs auf Geräte, Benutzerkonten oder Cloud-Ressourcen minimiert werden. Durch den Einsatz des Prinzips des geringsten Privilegs erhalten Benutzer zudem nur die Berechtigungen, die sie unmittelbar für ihre Aufgaben benötigen. Das minimiert das Risiko von Fehlkonfigurationen oder missbräuchlichem Zugriff. Neben der Kontrolle der Zugriffe sollten auch die möglichen Schwachstellen überwacht werden. Regelmäßige Schwachstellen-Scans helfen dabei, Sicherheitslücken in der Cloud-Infrastruktur zu identifizieren und zu beheben, bevor diese ausgenutzt werden können. Die Überprüfung und Optimierung von Cloud-Konfigurationen gewährleistet, dass Sicherheitseinstellungen korrekt konfiguriert sind. Zudem ermöglicht eine zentrale Aufstellung und Management aller in der Cloud vorhandenen Assets eine bessere Kontrolle und Überwachung der Daten, Dienste und Einstellungen. Um bekannte Sicherheitslücken zu schließen, sollte eingesetzte Software regelmäßig aktualisiert werden. Um menschliche Fehler zu minimieren, sollten außerdem die Mitarbeiter mit Schulungen für sicherheitsrelevante Themen sensibilisiert werden [11].

Im Bezug auf diese Bedrohung werden die Begriffe Data Loss (Datenverlust) und Data Leakage (Datenleck) häufig als Synonym verwendet, aber sie haben einige Unterschiede. Datenverlust ist der Verlust von Daten, der nicht wiederherstellbar ist, wie z.B. durch Schäden an Speichermedien, unbeabsichtigtes Löschen oder Hardwarefehler. Datenlecks hingegen beziehen sich auf die unbeabsichtigte oder absichtliche Übertragung von Daten aus einem gesicherten Bereich. Daher können Datenlecks auftreten, wenn unbefugte Personen sensible oder vertrauliche Informationen erhalten [12]. Aus diesem Grund wird in dieser Arbeit der Ausdruck Datenleck oder Data Leakage verwendet, um die unbeabsichtigte Offenlegung von Daten zu beschreiben.

III. Cloud Data Leakage Prevention System

Im vorherigen Kapitel II über die Bedrohung durch versehentliche Datenoffenlegung wurden deutlich, wie schnell ein Datenleck in Unternehmen auftreten kann. Die Bedeutung effektiver Maßnahmen zur Vermeidung von Datenlecks hat sich aufgrund der wachsenden Datenmengen und des damit verbundenen Risikos einer Datenschutzverletzung erhöht. Dieser Bedarf wurde 2022 erkannt, als in der neuesten Version der Norm ISO 27001:2022 die Data Leakage Prevention eingeführt wurde. Die internationale Norm ISO 27001 definiert die Bedingungen für die Einrichtung, Umsetzung und kontinuierliche Verbesserung eines dokumentierten Informationssicherheits-Managementsystems. Die Norm gibt außerdem Vorschriften für die Beurteilung und Behandlung von Informationssicherheitsrisiken, die an die spezifischen Bedürfnisse jedes Unternehmens angepasst werden müssen [13]. Ein Datenleck kann auf verschiedene Weise auftreten. Trotz der Tatsache, dass es nicht immer möglich ist, das Auftreten vollständig zu verhindern, können Maßnahmen ergriffen werden, um die Wahrscheinlichkeit eines Auftretens zu verringern. Diese Maßnahmen werden als Data Leakage Prevention (DLP) bezeichnet [13]. Dabei handelt es sich um eine Reihe von Technologien, Produkten und Methoden, die dazu dienen, zu verhindern, dass vertrauliche Informationen ein Unternehmen verlassen. In den letzten Jahrzehnten wur-

den verschiedene Sicherheitssysteme wie Firewalls, Intrusion-Detection-Systeme (Einbrucherkennung) und virtuelle private Netzwerke (VPN) eingeführt, um das Risiko von Datenlecks zu reduzieren. Wenn die zu schützenden Daten klar definiert, strukturiert und konstant sind, erfüllen diese Systeme ihren Zweck. Jedoch sind sie unzuverlässig für Daten, die sich ändern oder unstrukturiert sind. Durch einfache Regeln kann beispielsweise eine Firewall den Zugriff auf ein sensibles Datenobjekt verhindern. Die Firewall erkennt jedoch nicht, wenn das Datenobjekt über einen E-Mail-Anhang gesendet wird. DLP-Systeme hingegen sind darauf spezialisiert, vertrauliche Daten zu identifizieren, zu überwachen und zu schützen und unerwünschte Datenbewegungen zu verhindern. [4].

A. Cloud Data Leakage Prevention System

Ein DLP-System umfasst eine Reihe von Regeln und Richtlinien, die Daten nach ihrem Typ klassifizieren, um sicherzustellen, dass sie nicht böswillig oder versehentlich weitergegeben werden. Das System überwacht Endbenutzeraktivitäten, den Datenfluss sowie die über das Netzwerk gesendeten Daten. Wenn verdächtige Aktivitäten erkannt werden, wird eine Systemwarnung ausgelöst. DLP-Lösungen identifizieren sensible Inhalte mithilfe von Datenklassifizierung-Label, Techniken zur Inspektion von Inhalten und Kontextanalysen. Sie überwachen die Datenaktivität und kontrollieren sie anhand vordefinierter DLP-Richtlinien. Die Richtlinien definieren, ob die Verwendung bestimmter Inhalte oder Daten in bestimmten Situationen erlaubt sind [6].

Gartner klassifiziert DLP-Lösungen in drei Kategorien. 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 ist in der Lage, verschiedene Datenquellen zu integrieren. So kann diese Lösung den gesamten Lebenszyklus von Daten in einem Unternehmen abdecken. DLP-Richtlinien werden dabei an zentraler Stelle verwaltet und durchgesetzt. Dagegen werden integrierte DLP-Lösungen direkt in einen Dienst, wie bspw. ein E-Mail-Gateway, integriert und verfügen deshalb nur über begrenzte Richtlinienfunktionen. Das Management von mehreren integrierten DLP-Systemen ist ein manueller Aufwand, jedoch werden diese Systeme im jeweiligen Dienst speziell an die Anforderungen angepasst und können Inhaltsüberprüfungen besser durchführen. Cloud-native DLP-Lösungen sind die dritte Kategorie, zu der sowohl SaaS-Lösungen als auch Cloud-Anbieter mit integrierten DLP-Funktionen gehören. 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. Dies umfasst die Identifikation von Daten in Form von Dokumenten, E-Mails, Datenbanken und anderen Formaten [6]. 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) bietet freiwillige Standards und Best Practices, die Unternehmen dabei helfen, Cybersecurity-Risiken zu managen und zu reduzieren. Es gibt Unternehmen eine Struktur, um ihre aktuelle Cybersicherheitssituation zu bewerten, verbesserungsbedürftige Bereiche zu identifizieren, Maßnahmen zu priorisieren, Fortschritte zu bewerten und mit den Stakeholdern zu kommunizieren. Die CSF besteht aus fünf Kernfunktionen: 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 Literatur-Recherche ergab die folgende Auswahl an Best-Practises, die in DLP-Systemen eingesetzt werden sollten. Um sensible Daten schützen zu können, müssen diese zuerst identifiziert werden. Die Aufgabe besteht darin, ein Dateninventar zu erstellen, die Daten nach ihrer Sensibilität zu klassifizieren und sie entsprechend zu kennzeichnen. Zum Schutz der sensiblen Daten sollten Maßnahmen ergriffen werden, die den Zugriff auf die Daten einschränken. Das bedeutet, dass Richtlinien wie minimale Zugriffsrechte, starke Authentifizierungsmethoden und strenge Zugriffskontrolllisten eingeführt werden sollten. Außerdem sollten Daten sowohl im Ruhezustand als auch während der Übertragung verschlüsselt werden. So wird sichergestellt, dass die Daten selbst dann, wenn sie abgefangen werden, für unbefugte Benutzer unlesbar bleiben. Zusätzlich sollte ein DLP-System die Datenströme innerhalb und nach außen überwachen, um potenzielle Datenschutzverletzungen oder Richtlinienverstöße in Echtzeit erkennen zu können. Dies ermöglicht eine schnelle Reaktion auf potenzielle Probleme und begrenzt den daraus resultierenden Schaden [15] [16] [17].

Die Funktionen eines DLP-Systems basieren alle darauf, dass sensible Daten erkannt und in irgendeiner Art 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.

B. Erkennung von sensiblen Daten

Unternehmen setzen immer mehr auf SaaS-Produkte, anstatt sie als Produkt zu kaufen [18]. In ihrem Tagesgeschäft verlassen sich Unternehmen oft auf mehrere Softwareprodukte, um verschiedene Anforderungen zu erfüllen. Das hat zur Folge, dass die Daten des Unternehmens über verschiedene Apps und Cloud-Plattformen verstreut 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].

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 durchdringen fast jeden Aspekt unseres persönlichen und beruflichen Lebens. Das Spektrum dieser sensiblen Informationen reicht von persönlichen Daten über Finanzinformationen und Geschäftsgeheimnisse bis hin zu biometrischen Merkmalen und umfasst eine Vielzahl von Kategorien. Guo, Liu et al. [20] kategorisiert beispielsweise Daten in die 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.

Neben den Kategorien muss auch der Kontext beachtet werden, in dem eine Information verwendet wird. Denn 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 Präferenz basieren, die mit jeder Information verbunden sind.

- Der semantische Kontext wird auf Grundlage der semantischen Bedeutung eines Begriffs gebildet. Die semantische Bedeutung einer Sequenz wirkt sich zum Beispiel auf ihre Sensibilität aus.
- Im Kontext der Akteure wird die Sensibilität von Daten abhängig von den Akteuren, die an der Informationsübermittlung beteiligt sind, betrachtet. 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. Sie sind zudem nicht immer klar trennbar und schließen sich nicht gegenseitig aus. Ein oder mehrere Kontexte können sich gleichzeitig unterschiedlich auf die Sensibilität auswirken. Manche Daten können jedoch auch unabhängig vom Kontext vertraulich sein, wie z.B. Passwörter oder Kreditkartennummern.

Die Einteilung von Daten in verschiedene Geheimhaltungsklassen ist ein häufig verwendetes Verfahren in militärischen und behördlichen Anwendungen. 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.

Außerdem wird die Klassifizierung auch von der Struktur der Daten beeinflusst. Mehr als 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 Daten, die auf einem physischen oder digitalen Speichermedium gespeichert sind, wie bspw. einer Datenbank, auf einer Festplatte, im Cloud-Speicher oder einem externen Datenspeicher. Ruhende Daten sind in der Regel inaktiv und werden nicht aktiv gelesen oder gerade übertragen. Sicherheitsmaßnahmen wie Verschlüsselung, Authentifizierung und Zugriffskontrollregeln werden von DLP-Systemen zu Erkennungs- und Überwachungszwecken eingesetzt [24] [17].

Daten in Bewegung hingegen beziehen sich auf den Zustand von Daten, wenn sie gerade 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 überprüfen [24] [17].

Daten in Verwendung sind Daten, die eine Person oder ein System aktiv verarbeiten, aktualisieren, anhängen oder löschen. Dabei befinden sich die Daten auf der Workstation, dem Laptop, dem USB-Stick oder der externen Festplatte des Endnutzers sowie auf Netzlaufwerken oder Druckern. Diese Art von Daten ist besonders anfällig für Sicherheitsbedro-

hungen, da sie aktiv manipuliert werden. Diese Daten sind über verschiedene Endpunkte hinweg sichtbar, wenn auf sie zugegriffen wird. Um diese Daten zu schützen, wird über das DLP-System eine starke Benutzerauthentifizierung sowie ein Identitäts- und Profilmanagement eingeführt. Außerdem kann je nach DLP-Technologie ein Endpunkt-Agent auf dem Gerät des Endnutzers installiert werden, um die Datennutzung und -übertragung zu überwachen [24] [17].

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

IV. Methoden der automatischen Datenklassifizierung

Viele Studien in der Literatur haben Klassifizierungsmethoden verwendet, um die Datensicherheit in der Cloud zu gewährleisten. Die vorgeschlagenen Lösungen lassen sich in zwei Klassen einteilen: die manuelle Klassifizierung, die vom Nutzer festgelegt wird, und die automatische Klassifizierung, bei der ein Algorithmus zum Einsatz kommt.

Die manuelle Datenklassifizierung ist trotz Fortschritte bei automatisierten Technologien eine gängige Methode. In einigen Fällen, wie bei der Klassifizierung von Informationen wie geistigem Eigentum oder Geschäftsgeheimnissen, bleibt die manuelle Klassifizierung erforderlich. Menschen sind in der Lage, leicht die Kategorien, Kontexte, Strukturen und Zustände der Daten ganzheitlich zu berücksichtigen. Außerdem kann die manuelle Klassifizierung für kleinere Unternehmen kostengünstig sein [25] [26].

Die manuelle Klassifizierung von Daten ist jedoch sehr anfällig für menschliche Fehler und Inkonsistenzen. Die Subjektivität der Menschen kann zu inkonsistenten Klassifizierungen führen, was die Genauigkeit und Zuverlässigkeit der Sicherheitsmaßnahmen beeinträchtigen kann. Zudem kann eine unzureichende Genauigkeit zu unvollständigen oder falschen Klassifizierungen führen. Im Normalfall werden Daten bei ihrer Erstellung klassifiziert, doch sie können sich im Laufe der Zeit ändern, wodurch die ursprüngliche Klassifizierung veraltet und nicht mehr richtig sein kann. Die manuelle Einordnung von großen Datenmengen kann sehr zeitaufwändig und arbeitsintensiv sein und ab einer bestimmten Größe nicht mehr manuell verarbeitet werden. Manuelle Prozesse können die Anpassungsfähigkeit und schnelle Reaktion auf sich ändernde Geschäftsanforderungen reduzieren [27].

Aufgrund der wachsenden Datenmengen und der zunehmenden Komplexität der Informationssicherheitsanforderungen erscheint die automatisierte Datenklassifizierung häufig als effizientere Lösung, die eine genauere und konsistentere Identifizierung sensibler Daten ermöglicht.

Der Prozess, bei dem maschinelle Algorithmen und Technologien verwendet werden, um Daten automatisch zu identifizieren, zu kategorisieren und entsprechend ihres Sensitivität zu klassifizieren, wird als automatische Datenklassifizierung bezeichnet. Diese Methode verwendet maschinelles Lernen, Mustererkennung oder künstliche Intelligenz, um Daten zu analysieren und automatisch geeignete Klassifizierungen zuzuweisen. Die aktuellen Techniken in der Data Leakage Prevention können allgemein in zwei Kategorien eingeteilt werden: inhaltsbasierte Analyse und kontextbasierte Analyse. Methoden, die auf dem Inhalt basieren, untersuchen den Inhalt von Daten anhand von Merkmalen sensibler Informationen wie regulären Ausdrücken und Datenfingerabdrücken. Inhaltsbasierte Methoden verwenden vorhersehbare Muster wie z.B. IP-Adressen oder E-Mail-Adressen, um sensible Daten zu erkennen. Kontextbasierte Techniken identifizieren vertrauliche Daten anhand von Merkmalen im Zusammenhanf mit den überwachten Daten. Der kontextbasierte Ansatz ist damit effektiver für vertrauliche Daten ohne vorhersehbare Muster [20] [28] [3]. Um sensible Informationen umfassender und genauer zu extrahieren, sollten daher verschiedene Methoden angewendet werden. Die 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.

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

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

A. Klassifizierung mit manueller Definition

a) regelbasierte Methoden: Eine einfache und häufig angewendete Technik zur automatischen Datenklassifizierung basiert auf einem Wörterbuch oder einer Regel, die im Wesentlichen den gegebenen Text mit einer Liste vordefinierter regulärer Ausdrücke und Schlüsselwörter abgleicht. Diese Methode verwendet vordefinierte Regeln und Bedingungen, 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]. Die Identifikation von Kreditkartennummern in Textdokumenten ist ein Beispiel für die Verwendung regelbasierter Methoden. Zur Erkennung kann ein regulärer Ausdruck verwendet werden, der die Zeichenfolge nach dem Muster einer Kreditkartennummer definiert. Muster in regulären Ausdrücken umfassen meistens normale Zeichen mit wörtlicher Bedeutung und Metazeichen, um ein Erkennungsmuster zu

bilden [4]. Regeln können sich auch auf bestimmte Schlüsselwörter oder Phrasen beziehen, die auf personenbezogene Informationen wie „Sozialversicherungsnummer“ oder „vertraulich“ 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, da die Regeln schnell unpraktisch werden, wenn Datenformate, Kontext, Wortvariationen und Abkürzungen kombiniert werden müssen. Der Einsatz dieser Methode erfordert auch ein gut definiertes und gepflegtes Wörterbuch und Regelsatz. AAußerdem wird der semantische Kontext der Wörter bei einem reinen Textabgleich nicht berücksichtigt, was zu einer geringen Genauigkeit der Klassifizierung führen kann [29]. Trotzdem 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 sind eindeutige Identifikatoren für die entsprechenden Daten und werden genutzt, um sensible Daten zu identifizieren und automatisch zu klassifizieren. Eindeutige Fingerabdrücke für Wörter, Sätze oder ganze Dateien werden mithilfe von Wortmustern aus regulären Ausdrücken oder vordefinierten Wörterbüchern erstellt und als Vorlage für sensible Daten verwendet. 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, sobald nur ein Zeichen verändert wurde [4]. Ein weiterer Ansatz ist deshalb das „Fuzzy-Hashing“. Hierbei 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].

Sowohl im Speicher als auch bei Netzwerkübertragungen oder bei Verwendung kann das Data Fingerprinting sensible Daten erkennen. Der Fingerabdruck ist in manchen Fällen jedoch keine zuverlässige Methode. Die Klassifizierung funktioniert nicht, wenn die Daten verschlüsselt oder passwortgeschützt sind oder der Inhalt nicht eindeutig mit dem Fingerabdruck übereinstimmt. Außerdem ist es notwendig, dass die Vorlagen kontinuierlich aktualisiert werden und der Ansatz kann bei großen Datenmengen ressourcenintensiv sein.

B. Klassifizierung mit maschinellem Lernen

Die Datenklassifizierung ist eine Technik des maschinellen Lernens, mit der die Klasse der nicht klassifizierten Daten vorhergesagt wird. Dabei werden die Methoden in zwei Kategorien eingeteilt: überwachtes Lernen und unüberwachtes Lernen. Für überwachtes Lernen sind Testdaten mit bereits zugeteilten Klassen vordefiniert. So kann das Modell sein Ergebnis mit der Ziel-Klasse vergleichen und entsprechend von den Fehlern lernen. Beim unüberwachten Lernen sind keine Klassen definiert, sondern die Klassifizierung der Daten erfolgt automatisch. Ein unüberwachter Algorithmus sucht nach Mustern und Ähnlichkeiten zwischen den Elementen citeFrochte.2018b.

Für die meisten Methoden des maschinellen Lernens ist es erforderlich, dass die Daten vorab bereinigt oder vorbereitet werden. Die einzelnen Schritte können dabei je nach Methode variieren. Der Datensatz wird meistens von Stop-Wörtern bereinigt, die keinen oder nur wenig Kontext liefern wie bspw. ’und’ oder ’der’. Klassische Methoden sind in der Textverarbeitung Stemming und Lemmatization. Beim Stemming werden die Worte der Eingabe auf ihren Wortstamm zurückgeführt, bei der Lemmatization werden ähnliche oder inhaltlich gleiche Begriffe vereinheitlicht. Dadurch wird die grammatikalische Komplexität des Inputs reduziert und die Datenqualität optimiert. Für Methoden wie neuronale Netze müssen Eingabedaten zusätzlich noch in Token und Vektoren umgewandelt werden. Bei der Tokenization wird der Text ist einzelne Bestandteile, sogenannte Tokens aufgeteilt werden. Meistens sind das einzelne Worte. Die Tokens werden anschließend in Vektoren umgewandelt, die den jeweiligen Token im Modell repräsentiert. Für die Vektorisierung gibt es verschiedene Algorithmen, um Informationen wie inhaltlich ähnliche oder zusammenhängende Tokens zu behalten [31].

a) k-NN: Zardari, Jung et al. [32] waren die ersten, die eine Methode aus dem maschinellen Lernen zur Datenklassifizierung verwendet haben, um im Cloud Computing die Datensicherheit zu verbessern. Sie verwendeten die k-Nearest Neighbor-Methode (k-NN), um Daten in der Cloud als sensibel und nicht-sensibel zu klassifizieren. k-NN ist eine Methode aus dem überwachten maschinellen Lernen und wird häufig zur Klassifizierung, Mustererkennung und Schätzung verwendet. Es handelt sich um einen sogenannten Instanzbasierten Algorithmus, der darauf basiert, dass ähnliche Dinge in der Regel nah beieinander liegen. Die k-NN-Methode verwendet die Nachbarn eines Datenpunkts, um Vorhersagen zu treffen oder Klassenzuweisungen zu machen. Der Algorithmus beginnt mit einem Datensatz, der aus Beispielen mit bekannten Klassen oder Werten besteht. 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 einer Klassifikation

werden die Klassen der ausgewählten k Nachbarn betrachtet und die Mehrheitsklasse 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. 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.

b) Boosting: Ensemble Learning ist ein Begriff aus dem maschinellen Lernen und beschreibt das Zusammenschalten von mehreren Methoden, um das Modellergebnis zu verbessern. Der Ansatz beim Boosting ist, mehrere schwache Modelle zu einem starken Modell zusammenzusetzen [34]. Kaur, Zandu [35] 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 Theorem beruht. Die Grundlage ist die Annahme, dass das Auftreten eines Merkmals unabhängig mit dem Auftreten eines anderen Merkmals innerhalb der Klasse ist. Das Bayes Theorem ist eine Formel zur Berechnung der bedingten Wahrscheinlichkeit $P(A|B)$. Diese bedingte Wahrscheinlichkeit lässt sich mit der Formel 1 berechnen. Zur Klassifizierung eines Merkmals wird mit der Formel für jede Klasse berechnet, mit welcher Wahrscheinlichkeit das Merkmal zu dieser Klasse gehört. Gewählt wird schließlich die Zuordnung mit der höchsten Wahrscheinlichkeit [36].

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

Im Ansatz von [35] wird dieser Klassifikator mit AdaBoost, Abkürzung für adaptive Boosting, optimiert. 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 [34].

In der neuen Klassifikationsmethode von [35] wird zuerst der Trainingsdatensatz in eine bestimmte Anzahl an Datensätzen aufgeteilt. Anschließend werden die einzelnen Teildatensätze schrittweise verarbeitet. Jeder Teildatensatz enthält eine Menge an Daten-Tupeln. Zu Beginn erhält jedes Tupel die gleiche Gewichtung. Dann wird das hybride Klassifizierungsmodell aus Naive Bayes und AdaBoost mit dem Teildatensatz trainiert. Anschließend werden bei den Tupeln die Gewichte aktualisiert, je nachdem ob sie richtig oder falsch klassifiziert wurden. Nach den Durchgängen aller Teildatensätze ergibt sich dadurch ein Set an Klassifikator-Modellen mit jeweiligen Gewichtungen, die in Kombination eine genaue Vorhersage für eine Klasse treffen.

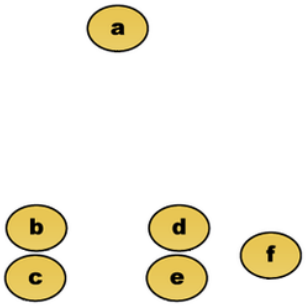
Kaur, Zandu [35] zeigten, dass ihre vorgeschlagene Methode mit 94,2529% Genauigkeit deutlich besser klassifiziert als der k-NN-Algorithmus, der nur eine Genauigkeit von 51,7241% hatte.

c) Clusteranalyse: Die Clusteranalyse ist eine Datenanalysetechnik aus dem Bereich maschinellen Lernens. Clustering ist eine unüberwachte Lernmethode, die Muster in Eingabedaten ohne vordefinierte Zielwerte erkennt. Beim Clustering werden unsortierte Informationen auf der Grundlage von Ähnlichkeiten, Mustern und Unterschieden gruppiert, ohne dass die Daten zuvor trainiert wurden. Im Zusammenhang mit der Klassifizierung sensibler Daten wird die Clusteranalyse auf die Daten in einem Unternehmen angewendet. Die daraus resultierenden Cluster enthalten dann ähnliche Dokumente nach einer Ähnlichkeitsmetrik wie dem euklidischen Abstand oder der Kosinusähnlichkeit [37]. 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.

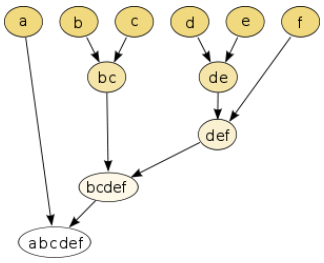
Zwei gängige Clustering-Methoden sind das hierarchische Clustering und das partitionierende Clustering. Hierarchisches Clustering wird in der Regel für dsa Clustering von Texten verwendet, wobei jedes Dokument auf der Grundlage seiner Ähnlichkeit schrittweise in einen vordefinierten Cluster zusammengeführt wird. Durch diesen Prozess entsteht eine Clusterhierarchie, die als Baumstruktur, das sogenannte Dendrogramm, dargestellt werden kann. Abbildung 1a 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 dann 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 1b, 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 enthalten 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 [37].

Aufgrund seiner Einfachheit und Flexibilität wird hierarchisches Clustering häufig verwendet und bietet den Vorteil, dass jede Art von Ähnlichkeitsmessung durchgeführt werden kann. Außerdem bietet dieses Verfahren eine detaillierte Darstellung der Clusterstruktur, wodurch unterschiedliche Granularitätsstufen von Clustern untersucht werden können.

Beim partitionierenden Clustering wird ein Datensatz in eine vorab definierte 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



(a) Datensatz. Quelle: In Anlehnung an [38].



(b) Dendrogramm. Quelle: In Anlehnung an [38].

der k-means-Algorithmus. Das 'k' steht für die Anzahl an zu definierenden Clustern und 'means' für den Mittelwert, also das Zentrum des Clusters. Zu Beginn muss die Anzahl der Cluster bestimmt werden. Dies kann sich z.B. daran orientieren, in welche Vertraulichkeitsstufen die Daten eingeteilt werden sollen. Anschließend werden für die Cluster initial jeweils zufällig ein Cluster-Mittelpunkt, auch Centroid genannt, gewählt. Dann 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. Nun folgen Schritte, die sich solange wiederholen, 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 [37].

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 abhängig von der Wahl des 'k' und der initialen Cluster [37].

d) Context-based Approach for Structured Sensitive Data Detection: 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. Dafür muss der Inhalt der Tabellen-Zelle interpretiert werden und der Kontext der umgebenden Zellen berücksichtigt werden. Bisherige Ansätze verwendeten dafür stark regelbasierte Methoden, deren Grenzen bei einer großen Menge an verschiedenen Datentypen schnell erreicht wurden. Außerdem können sie nur begrenzt Kontext und Semantik miteinbeziehen. Um dieses Problem zu lösen, entwickelten Kužina, Petric et al. [3] eine neue Methode namens 'Context-based Approach for Structured Sensitive Data Detection' (CASSED). Dabei stellen sie einen Spaltenkontext durch die Kombination von Spaltenmetadaten und Zellwerten her, der in einen einzelnen Input-Vektor umgewandelt wird. Mit diesem Input-Vektor wird anschließend das BERT-Modell für die Klassifizierung verwendet.

BERT steht für 'Bidirectional Encoder Representations from Transformers' und ist ein von Google entwickeltes Open-Source Framework zur Erstellung von Transformer-basierten Natural-Language-Processing-Modellen. BERT ist darauf spezialisiert, kontextuelle Zusammenhänge und Beziehungen zwischen Wörtern zu erfassen. Transformer-basierte Modelle nutzen einen sogenannten Selbstaufmerksamkeitsmechanismus, indem die Beziehung eines Wortes mit jedem anderen Wort in einem Satz bestimmt wird. Außerdem enthalten sie mehrere Encoder- und Decoder Schichten, die einen Text lesen und versuchen, das nächste Wort vorherzusagen, sowie voll-vernetzte neuronale Netze. BERT nutzt Transformer, einen Aufmerksamkeitsmechanismus und nur Encoder-/Decoder-Schichten. BERT verarbeitet Texteingaben bidirektional, indem die Sequenzen sowohl von Anfang als auch vom Ende her analysiert werden, um ein besseres Verständnis für die kontextuellen Beziehungen der Wörter zu bekommen.

Im ersten Schritt erfolgt die Umwandlung der Spalten in Input-Vektoren. Dabei werden die Spaltenüberschrift zusammen mit mehreren Zellwerten derselben Spalte als Tokens dargestellt und durch Trennzeichen getrennt. In Abbildung 2 ist eine Umwandlung einer Spalte in einen Input-Vektor dargestellt. Dabei wird die Spaltenüberschrift mit einem Punkt zur ersten Zelle getrennt, während die Zellen zueinander mit einem Komma getrennt werden. So wird dem Modell zusätzliche Informationen geliefert, dass diese Werte unterschiedlich behandelt werden sollten. BERT kann nur eine Anzahl von maximal 512 Tokens als Input-Vektor verarbeiten. Deshalb müssen größere Input-Vektoren noch aufgeteilt werden.

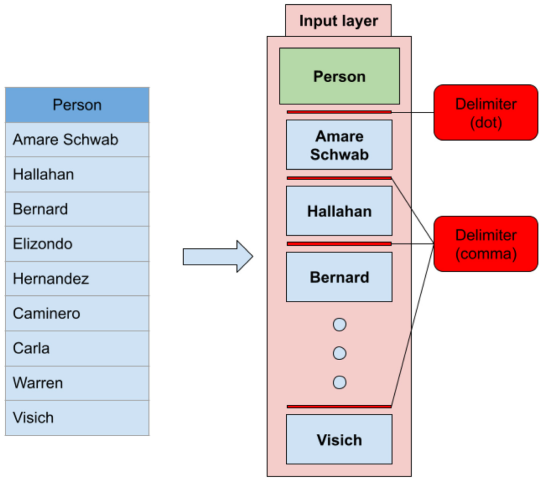


Abbildung 2: Beispiel einer Umwandlung einer Spalte in einen Input-Vektor. Quelle: [3].

Zur Klassifizierung erzeugt der Decoder von BERT für jede mögliche Klasse eine nicht normalisierte Vorhersage, die anschließend über alle Spaltenteile gemittelt wird, in die die Spalte aufgeteilt wurde. Auf diese Vorhersage-Werte wird eine Sigmoidfunktion angewendet, um normalisierte Wahrscheinlichkeiten für jede Klasse zu erzeugen.

Zusätzlich zum BERT-Modell enthält der CASSED-Ansatz eine regelbasierte Schicht, die strukturierte Formate wie E-Mails oder Sozialversicherungsnummern mittels regulären Au-

drücken klassifiziert. Außerdem wird ein Wörterbuch für bekannte sensible Daten oder Merkmale von Geschäftsgeheimnissen eingesetzt. Auch diese Schicht ermittelt für jede Klasse eine mögliche Wahrscheinlichkeit. Diese und die Wahrscheinlichkeiten der BERT-Schicht werden kombiniert zu einer Gesamt-Wahrscheinlichkeit pro Klasse. Anschließend werden die Klassen zugeteilt, deren Wahrscheinlichkeit einen bestimmten Schwellwert überschreiten. So ist es auch möglich, dass eine Spalte mehrere Klassen erhält. Abbildung 3 stellt die Architektur des CASSED-Ansatzes dar.

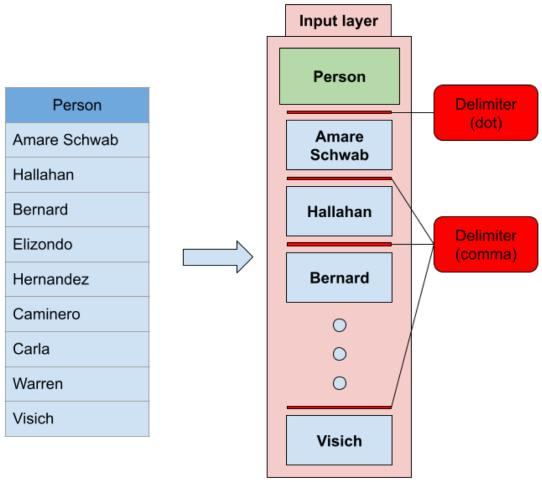


Abbildung 3: Überblick über die CASSED Methode. Quelle: [3].

Im Vergleich zu anderen kontextbasierten Klassifizierungsmethoden erzielt die CASSED Methode deutlich bessere Ergebnisse. Allein durch den reinen Einsatz von BERT schneidet das Modell besser ab und ist durch die regelbasierte Schicht sogar noch präziser.

Einen ähnlichen Ansatz verfolgen auch Guo, Liu et al. [20] mit ihrem Ansatz '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 gut darin, Abhängigkeiten in langen Sequenzen zu erfassen. Im BERT-BiLSTM-Modell wird die bidirektionale Kontextrepräsentation von BERT als Eingabe für das BiLSTM verwendet. Auch dieser Ansatz erzielte sehr gute Ergebnisse zur Klassifizierung von sensiblen Daten unter Berücksichtigung des Kontextes.

C. Anwendung in der Cloud Security

1) Labeling für andere Maßnahmen:

V. Ausblick

Literatur

[1] KPMG, "Nutzung von cloud computing in unternehmen in deutschland in den jahren 2011 bis 2022," 2022. [Online]. Available: https://de.statista.com/statistik/daten/studie/177484/umfrage/einsatz-von-cloud-computing-in-deutschen-unternehmen-2011/

[2] C. Surianarayanan and P. R. Chelliah, "Introduction to cloud computing," in *Essentials of Cloud Computing*, ser. Texts in Computer Science, C. Surianarayanan and P. R. Chelliah, Eds. Cham: Springer International Publishing and Imprint: Springer, 2023, pp. 1–38.

[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," 2018. [Online]. Available: https://crowdresearchpartners.com/wp-content/uploads/2017/07/Insider-Threat-Report-2018.pdf

[6] R. Chugh and A. Bales, "Marktführer für data loss prevention," 2023. [Online]. Available: https://www.gartner.com/doc/reprints?id=1-2EYVL2C2&ct=230912&st=sb

[7] Cloud Security Alliance, "Top threats to cloud computing pandemic eleven," 2022. [Online]. Available: https://cloudsecurityalliance.org/artifacts/top-threats-to-cloud-computing-pandemic-eleven

[8] Z. Whittaker, "Toyota confirms another years-long data leak, this time exposing at least 260,000 car owners," *TechCrunch*, 31.05.2023. [Online]. Available: https://techcrunch.com/2023/05/31/toyota-customer-data-leak-years/

[9] PSNI, "Police service of northern ireland statement on data breach | psni," 12.11.2023. [Online]. Available: https://www.psni.police.uk/latest-news/police-service-northern-ireland-statement-data-breach

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

[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)*, IEEE, Ed. IEEE, 2015, pp. 350–355.

[12] Proofpoint, "Dlp - data loss prevention: Schutz vor datenverlust | proofpoint de," 2021. [Online]. Available: https://www.proofpoint.com/de/threat-reference/dlp

[13] V. Monev, "Data leakage prevention in iso 27001: Compliance and implementation," in *2023 International Conference on Information Technologies (InfoTech)*, IEEE, Ed. IEEE, 2023, pp. 1–5.

[14] NIST, "Framework for improving critical infrastructure cybersecurity," 2014. [Online]. Available: https://www.nist.gov/system/files/documents/cyberframework/cybersecurity-framework-021214.pdf

[15] M. E. Hussain and R. Hussain, "Cloud security as a service using data loss prevention: Challenges and solution," in *Internet of Things and Connected Technologies*, ser. Lecture Notes in Networks and Systems, R. Misra, N. Kesswani, M. Rajarajan, B. Veeravalli, and A. Patel, Eds. Cham: Springer International Publishing and Imprint: Springer, 2022, vol. 340, pp. 98–106.

[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)*. IEEE, 2022, pp. 1–6.

[18] Gartner, "Umsatz mit software-as-a-service (saas) weltweit von 2010 bis 2022 und prognose bis 2024 (in milliarden us-dollar)," 2023. [Online]. Available: https://de.statista.com/statistik/daten/studie/194117/umfrage/umsatz-mit-software-as-a-service-weltweit-seit-2010/

[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." [Online]. Available: https://arxiv.org/pdf/1707.02919.pdf

[24] A. Shabtai, Y. Elovici, and L. Rokach, "A taxonomy of data leakage prevention solutions," in *A survey of data leakage detection and prevention solutions*, ser. SpringerBriefs in Computer Science, A. Shabtai, Y. Elovici, and L. Rokach, Eds. New York: Springer, 2012, pp. 11–15.

[25] S. Divadari, J. Surya Prasad, and P. Honnavalli, "Managing data protection and privacy on cloud," in *Proceedings of the 3rd international conference on recent trends in machine learning, IoT, smart cities and applications*, ser. Lecture Notes in Networks and Systems, J. M. Zurada and V. K. Gunjan, Eds. Singapore: Springer, 2023, vol. 540, pp. 383–396.

[26] M. A. Alsuwaie, B. Habibnia, and P. Gladyshev, "Data leakage prevention adoption model & dlp maturity level assessment," in *2021 International Symposium on Computer*, 2021, pp. 396–405.

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

[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 2017*, G. C. Fox, Ed. Piscataway, NJ: IEEE, 2017, pp. 399–406.

[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 *Universiti Teknologi PETRONAS, IEEE Computer Society et al.*, 2014, pp. 1–6.

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

[34] ——, "Entscheidungsbäume," in *Maschinelles Lernen*, J. Frochte, Ed. München: Carl Hanser Verlag GmbH & Co. KG, 2018, pp. 117–160.

[35] 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.

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

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

[38] H. Bonthu, "Single-link hierarchical clustering clearly explained!" 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/single-link-hierarchical-clustering-clearly-explained/