Technische Hochschule Ingolstadt

Fakultät Informatik

Studiengang Wirtschaftsinformatik

Bachelorarbeit

Automatisierte Extraktion strukturierter Daten aus Informationssicherheits-Audits

vorgelegt von

Cedric Friedrichs
Matrikel-Nr: 00097252

Ausgegeben am:	17.02.2022
Abgegeben am:	07.07.2022

Erstprüfer:	Prof. Dr. Munir Georges
Zweitprüfer:	Prof. Dr.-Ing. Hans-Joachim Hof

Abstract

Die Automobilindustrie digitalisiert die eigene Produktion zunehmend und sammelt hierfür Daten im Shopfloor. Zur Gewährleistung der Informationssicherheit wurden eigene Standards wie die TISAX Zertifizierung entwickelt, sodass Daten entlang der Wertschöpfungskette vertraulich, verfügbar und integer sind. Im Rahmen der TISAX-Zertifizierung findet alle spätestens drei Jahre ein Audit statt. Eine kontinuierliche, externe Prüfung ist innerhalb des Zeitraums zwischen zwei Audits nicht gegeben.

Ziel dieser Arbeit ist eine automatisierte Extraktion strukturierter Daten aus Informationssicherheits-Audits. Dafür wird das in natürlicher Sprache verfasste Information Security Assessment in ein maschinenlesbares Dokument in einem standardisierten Datenformat transformiert. Dieses kann von Anwendungssystemen genutzt und weiterverarbeitet werden. Zusätzlich kann aus dem Dokument mit Hilfe von Natural Language Processing eine ausführbare Spezifikation erzeugt werden. Mit dieser Spezifikation kann schließlich aus jedem Szenario ein Test generiert werden, der in einer Testumgebung ausgeführt werden kann, womit ein kontinuierliches Audit und damit eine Überprüfung der Informationssicherheit im Unternehmen in Echtzeit ermöglicht wird.

Inhaltsverzeichnis

Abkürzungsverzeichnis

BDD	Behaviour-Driven Development	21
BSI	Bundesamt für Sicherheit in der Informationstechnik	3
CSV	comma-separated values	11
ISA	Information Security Assessment	1
ISO	International Organization for Standardization	6
IEC	International Electrotechnical Commission	6
ISMS	Informationssicherheitsmanagementsystem	6
ML	Machine Learning	11
NER	Named Entity Recognition	15
NLP	Natural Language Processing	15
POS	Part-of-Speech	17
TISAX	Trusted Information Security Assessment Exchange	1

1 Einführung

Im Folgenden werden Motivation sowie Zielsetzung und Vorgehen dargelegt, um einen Rahmen für die Bachelorarbeit zu setzen. Dabei wird in Kapitel 1.1 das Thema eingeführt und die grundlegende Problemstellung vorgestellt. Anschließend wird in Kapitel 1.2 das Ziel der Arbeit gesetzt und ein Überblick über die durchzuführenden Schritte wird geschaffen.

1.1 Motivation und Problemstellung

Die digitale Transformation ist derzeit das wichtigste Thema für produzierende Unternehmen und die digitale Vernetzung von Produktionsanlagen ist eines der definierenden Merkmale der digitalen Transformation in der Industrie. Zum Zeitpunkt der Erstellung der Arbeit befinden wir uns im Wandel, durch den Einsatz von digitalen Technologien und Prozessen die existierenden Wertschöpfungsketten zu transformieren und mit Hilfe von Daten diese zu optimieren. Ziel ist, dass unter dem Schlagwort Industrie 4.0, "[h]eutige starre und fest definierte Wertschöpfungsketten [...] abgelöst [werden] durch flexible, hochdynamische und weltweit vernetzte Wertschöpfungsnetzwerke mit neuen Arten der Kooperation" [Plattform Industrie 4.0 2019].

Die Automobilindustrie digitalisiert die eigene Produktion zunehmend und sammelt hierfür Daten im Shopfloor, dem Produktionsbereich des Unternehmens, sowie im restlichen Unternehmen. Zur Gewährleistung der Informationssicherheit wurden hier neben gesetzlichen Regelungen eigene Standards wie die Trusted Information Security Assessment Exchange (TISAX) Zertifizierung entwickelt, sodass Daten entlang der Wertschöpfungskette vertraulich, verfügbar und integer sind. Im Rahmen der TISAX-Zertifizierung findet eine manuelle Prüfung durch einen externen Auditor statt, wobei das TISAX-Label nach Bestehen des Prüfprozesses drei Jahre lang gültig ist und anschließend erneuert werden muss. Eine kontinuierliche, externe Prüfung innerhalb dieses Zeitraums zwischen zwei Audits ist nicht gegeben [Gleich 2022].

Im Rahmen des sAInd (Security Auditor für Industrienetze) Projektes des bayerischen Verbundforschungsprogramms des Freistaates Bayern soll mit dieser Arbeit durch eine automatisierte Extraktion strukturierter Daten aus der TISAX Zertifizierung ein Grundstein für eine Formalisierung und schließlich Automatisierung des Audits gelegt werden.

Das Information Security Assessment (ISA) für die Prüfung nach TISAX ist in natürlicher Sprache verfasst und wird als Tabellenkalkulations-Datei zur Verfügung gestellt. Eine Formalisierung ist Stand heute nicht gegeben, sie ist aber notwendig für eine automatisierte Verarbeitung und Prüfung. Nur durch die Formalisierung kann ein maschinelles ISA und damit eine automatisierte und kontinuierliche Prüfung nach TISAX erfolgen.

1.2 Vorgehen und Zielsetzung

Ziel dieser Arbeit ist eine automatisierte Extraktion strukturierter Daten aus Informationssicherheits-Audits. Da das ISA in natürlicher Sprache verfasst ist, ergibt sich die Notwendigkeit, das Dokument über die einzelnen Spalten hinaus in ein von Maschinen verarbeitbares Format zu transformieren und dabei explizit vorhandene Strukturen zu bewahren.

Zu Beginn wird für die Dokumente der TISAX, ISO/IEC 27000-Reihe und IT-Grundschutz in Kapitel 2 ein Verständnis über deren Inhalt und Aufbau, sowie den Bezug der Dokumente zueinander entwickelt. Anschließend werden die vorhandenen Dokumente in Kapitel 3 vorbereitet und transformiert, um in Kapitel 4 mit Hilfe von Natural Language Processing verarbeitet zu werden. Im nächsten Schritt werden in Kapitel 5 dann die neu erkannten Strukturen überführt und in das transformierte Dokument eingepflegt, sodass aus den Audit-Anforderungen ausführbare Software werden kann. Zur Bewertung der Ergebnisse erfolgt in Kapitel 6 eine Evaluation. Abschließend wird ein Fazit gezogen und ein Ausblick auf die Verwendung im Rahmen des sAInd Projektes und darüber hinaus gegeben.

2 Datenverständnis

Zu Beginn soll ein Datenverständnis über den TISAX-Fragenkatalog selbst, sowie die ISO/IEC 27000-Reihe und den IT-Grundschutz des Bundesamtes für Sicherheit in der Informationstechnik (BSI) erhalten werden. Die weiteren Datenquellen wurden ausgewählt, da die TISAX auf der ISO/IEC 27000-Reihe basiert und die ENX Association die Nähe zu ihr schätzt und beibehalten möchte. Der IT-Grundschutz als deutsches Pendant, das zur ISO/IEC 27000-Reihe kompatibel ist, ergänzt hierbei die Dokumente durch ausführliche Erklärungen [Friedrichs and Aufderheide 2022].

Für das Datenverständnis wird die TISAX auf Format und Inhalt analysiert, um mit der gewonnenen Erkenntnis anschließend die Datenvorbereitung durchzuführen und eine bessere Auswahl der Machine Learning Methodik zu treffen. Zusätzlich erfolgt eine statische Analyse der weiteren Datenquellen, bestehend aus den Standards der ISO/IEC 27000-Reihe sowie den Standards des BSI 200-1, 200-2 und 200-3 (IT-Grundschutz).

Die Dokumente werden jeweils in ihrer englischen Fassung verwendet, da die ISO/IEC 27000-Reihe als Ursprungsdatei im Original auf Englisch veröffentlicht wird. Zusätzlich sind weitaus mehr Sprachmodelle auf Englisch als auf Deutsch verfügbar, sodass es leichter fällt, ein geeignetes Modell für die Verarbeitung der Dokumente zu finden. Im weiteren Verlauf können Abbildungen zur Verdeutlichung und für einen leichteren Lesefluss in ihrer deutschen Version dargestellt sein, die Verarbeitung selbst erfolgt jedoch immer zuerst auf Englisch.

2.1 Untersuchung der TISAX

TISAX ist ein Mechanismus, Sicherheitsinformationen zwischen Organisationen auszutauschen und zu prüfen. Über eine Plattform wird ein Austausch mit anderen Teilnehmern ermöglicht und Prüfergebnisse von Geschäftspartnern zur gemeinsamen Anerkennung weitergegeben [ENX Association 2022b].

Der TISAX-Fragenkatalog[1] wird als Arbeitsmappe eines Tabellenkalkulationsprogramms auf der Website der ENX Association veröffentlicht. In dieser Arbeit wird die englische Version 5.0.4 vom 16. April 2021, die durch eine Agentur vom deutschen Original übersetzt wird, als Quelle betrachtet [Verband der Automobilindustrie e.V. 2021].

Die ENX Association ist ein Verein nach französischem Recht mit Sitz der Geschäftsleitung in Deutschland. Die Mitglieder des im Jahr 2000 gegründeten Vereins sind Automobilhersteller, Zulieferer und Automobilverbände. "Ziel [des Vereins] ist es, sichere und vertrauenswürdige Zusammenarbeit in industriellen Wertschöpfungsnetzwerken zu ermöglichen und zu vereinfachen." [ENX Association 2022a]

Zur Veröffentlichung der TISAX wurde das Dateiformat eines proprietären Tabellenkalkulationsprogramms gewählt. Der Fragenkatalog beinhaltet neben dem Tabellenblatt "Welcome", das die Arbeitsmappe erklärt, die Tabellenblätter "Maturity levels" und "Definitions" als Auslegungshilfe der Anforderungen, die in den Tabellenblättern "Information Security", "Prototype Protection" und "Data Protection" enthalten sind. Die übrigen Tabellenblätter (Deckblatt, Ergebnisse, Beispiele) werden in

1 https://www.enx.com/de-de/TISAX/downloads/

der Arbeit nicht betrachtet. Der besondere Fokus liegt auf dem Blatt "Information Security", da die Anforderungen an Informationssicherheit das Kernthema der Arbeit sind.

Die Abbildung 1 zeigt eine Auswahl der Spalten sowie beispielhaft die ersten Fragen des Blattes "Informationssicherheit". Die Spalte "ISA New" indexiert die einzelnen Kontrollfragen nach Über- und Unterkapiteln sowie einer Nummer je Kontrollfrage. Nach dieser Spalte erfolgt auch die Standardsortierung. Die Spalte Kontrollfrage ("Control question") enthält die eigentliche Kontrollfrage, nach der geprüft wird. Jede Kontrollfrage hat weiterhin ein Ziel ("Objective"), das erreicht werden soll, wenn die Kontrollfrage ausreichend beantwortet wird. Das Ziel gibt für das zu prüfende Unternehmen auch Kontext und Richtung für die Beantwortung der Frage vor.

Die darauffolgenden Spalten enthalten die eigentlichen Anforderungen unterschiedlicher Kategorien, wobei nicht jede Kategorie enthalten sein muss:

- Anforderung (muss)
  ("Requirements (must)")
- Anforderung (sollte)
  ("Requirements (should)")
- Zusatzanforderungen bei hohem Schutzbedarf
  ("Additional requirements for high protection needs")
- Zusatzanforderungen bei sehr hohem Schutzbedarf
  ("Additional requirements for very high protection needs")

Die Spalten "Anforderung (muss)" und "Anforderung (sollte)" sind ebenfalls in Abbildung 1 sichtbar.

Weiterhin werden in "Addressed protection objectives" ("confidentiality", "integrity", "availability") ein bis drei der klassischen Schutzziele der Informationssicherheit, die erreicht werden sollen, genannt [Bundesamt für Sicherheit in der Informationstechnik 2017]. Optional wird auch eine "Usual person responsible for process implementation", sowie eine "Reference to other standards" (aus der ISO/IEC 27000-Reihe) aufgeführt.

Die einzelnen Zellen sind in natürlicher Sprache ausgefüllt, neben der expliziten Struktur, die durch die Bedienoberfläche der proprietäreren Software unterstützt wird, wurde zusätzlich innerhalb der Zellen eine implizite Strukturierung beispielsweise durch ungeordnete Listen mittels dem ’+’ Zeichen für Überpunkte und dem ’-’ Zeichen für Unterpunkte der Anforderungen gewählt. In Abbildung 1 ist dies in der Spalte "Anforderungen (muss)" zu erkennen. Auch andere, innerhalb der Zellen konsistente und für den Menschen leicht erkennbare innere Strukturierungen wurden verwendet, für eine maschinelle Verarbeitung ist jedoch eine explizite Struktur in den Daten erforderlich.

Information Security Assessment

Fragebogen

ISA New	Kontrollfrage	Ziel	Anforderungen (muss)	Anforderungen (sollte)
1	**IS Policies and Organization**			
1.1	**Information Security Policies**			
1.1.1	**Inwieweit sind Richtlinien zur Informationssicherheit vorhanden?**	Die Organisation benötigt mindestens eine Richtlinie für Informationssicherheit. Diese spiegelt die Wichtigkeit und Bedeutung der Informationssicherheit wider und ist an die Organisation angepasst. Weitere Richtlinien können je nach Organisationsgröße und -struktur sinnvoll sein.	+ Die Anforderungen an die Informationssicherheit sind ermittelt und dokumentiert. - Die Anforderungen sind an die Ziele der Organisation angepasst. - Eine Richtlinie ist erstellt und von der Organisationsleitung freigegeben. + Die Richtlinie enthält Ziele und den Stellenwert der Informationssicherheit in der Organisation.	+ Die Anforderungen an die Informationssicherheit auf der Grundlage der Organisationsstrategie, Gesetzen und Verträgen sind in der Richtlinie berücksichtigt. + Verantwortlichkeiten für die Durchführung sind definiert. + Die Richtlinie weist auf Konsequenzen bei Nichtbeachtung hin. + Weitere relevante Richtlinien zur Informationssicherheit sind erstellt. + Eine regelmäßige Prüfung und - falls notwendig - Überarbeitung der Richtlinien sind etabliert. + Die Richtlinien werden Mitarbeitern in geeigneter Form (z. B. Intranet) zur Verfügung gestellt. + Die Richtlinien werden fallbezogen (ggf. auch in Auszügen) an externe Geschäftspartner weitergegeben. + Mitarbeiter und externe Geschäftspartner werden über für sie relevante Änderungen informiert.
1.2	**Organization of Information Security**			
1.2.1	**Inwieweit wird in der Organisation Informationssicherheit gemanagt?**	Nur wenn Informationssicherheit in den strategischen Zielen einer Organisation verankert ist, kann Informationssicherheit nachhaltig in einer Organisation umgesetzt werden. Das Informationssicherheitsmanagementsystem (ISMS) ist ein Steuerungsinstrument für die Organisationsleitung, mit dem sie sicherstellt, dass Informationssicherheit nicht nur ein Ergebnis von Zufällen und individuellem Engagement, sondern von nachhaltigem Management ist.	+ Der Geltungsbereich (Scope) des ISMS (die vom ISMS gemanagte Organisation) ist festgelegt. + Die Anforderungen der Organisation an das ISMS sind ermittelt. + Die Organisationsleitung hat das ISMS beauftragt und freigegeben. + Das ISMS stellt der Organisationsleitung geeignete Kontroll- und Steuerungsmittel zur Verfügung (z. B. Management-Review). + Anwendbare Kontrollen wurden ermittelt (z. B. ISO 27001 Statement of Applicability, ausgefüllter ISA Katalog). + Die Wirksamkeit des ISMS wird regelmäßig durch das Management überprüft.	Keine
1.2.2	**Inwieweit sind die Verantwortlichkeiten für Informationssicherheit organisiert?**	Ein erfolgreiches ISMS benötigt klare Verantwortlichkeiten in der Organisation.	+ Verantwortlichkeiten für die Informationssicherheit in der Organisation sind definiert, dokumentiert und zugewiesen. + Die verantwortlichen Mitarbeiter sind definiert und für ihre Aufgabe qualifiziert. + Die notwendigen Ressourcen stehen zur Verfügung. + Die Ansprechpartner sind innerhalb der Organisation und relevanten Geschäftspartnern bekannt.	+ Es existiert eine Definition und Dokumentation einer geeigneten Informationssicherheitsstruktur in der Organisation.
1.2.3	**Inwieweit werden Informationssicherheitsanforderungen in Projekten berücksichtigt?**	Bei der Durchführung von Projekten ist es wichtig, dass die Informationssicherheitsanforderungen berücksichtigt. werden. Dies gilt für Projekte der Organisation, unabhängig von der Art des Projekts. Durch eine geeignete Verankerung des Informationssicherheitsprozesses in den Projektmanagementmethoden der Organisation wird sichergestellt, dass keine Anforderungen übersehen werden.	+ Projekte sind unter Berücksichtigung ihrer Anforderungen an die Informationssicherheit klassifiziert.	+ die Vorgehensweise und Kriterien zur Klassifizierung von Projekten sind dokumentiert. + in einer frühen Phase des Projektes wird eine Risikobewertung auf Basis der definierten Vorgehensweise durchgeführt und bei Änderungen des Projektes wiederholt. + Für identifizierte Informationssicherheitsrisiken werden Maßnahmen abgeleitet und im Projekt berücksichtigt.

Abbildung 1 Auswahl von Spalten aus dem TISAX Fragenkatalog [Verband der Automobilindustrie e.V. 2021]

2.2 Statische Analyse weiterer Datenquellen

Gleichermaßen soll mit einer statischen Analyse der Datensatz ein Verständnis über die zugrundeliegenden Normen gewonnen werden. Hierfür wird das Tool KH Coder[2] verwendet, das quantitative Textanalyse und Text Mining ermöglicht und die Ergebnisse visuell aufbereitet (siehe Abbildungen 2-6). Dafür ist eine unformatierte Bereitstellung des Textes notwendig. KH Coder befreit den Input anschließend von Stop Words, also häufig auftretenden Wörter, die keinen Informationsgewinn bringen, und eine Stammformreduktion wird durchgeführt. In den interaktiven Graphiken ist dann eine Ansicht der jeweiligen Schlüsselworte im Kontext möglich, so ist in Abbildung 2 zu erkennen, dass beispielsweise mit morphologischen Varianten des Schlüsselworts "ensure" im lachsfarbenen Cluster des IT-Grundschutz notwendige Maßnahmen zur Sicherstellung der Informationssicherheit, beispielsweise durch ein Informationssicherheitsmanagementsystem (ISMS), verdeutlicht werden.

L	C	R
A management system encompasses all the provisions	ensuring	the supervision and management so that the organisation can achieve its objectives.
However , experience has shown that it is more difficult to	ensure	that organisational safeguards are implemented consistently.
This is the only manner of	ensuring	that the various process steps and decisions remain comprehensible.
All of the above is based on additional ISO standards in order to	ensure	the high quality and comprehensibility of certificates.
The core objective of an ISMS is to	ensure	the maintenance of information security and to continuously improve it.

Abbildung 2 Beispielsätze, die den Wortstamm "ensure" als Schlüsselwort im Kontext zeigen [Bundesamt für Sicherheit in der Informationstechnik 2017] [Higuchi 2016][eigene Abbildung]

Im Folgenden wird der Begriff "Token" als Einheit von Zeichen im Text, die Wörter, Zahlen und/oder Interpunktion ergeben, verstanden [Martín Abadi et al. 2015].

2.2.1 ISO/IEC 27000-Reihe

Die ISO/IEC 27000-Reihe wird von der International Organization for Standardization (ISO) und der International Electrotechnical Commission (IEC) veröffentlicht und umfasst eine Reihe von Normen zur Informationssicherheit. In der ISO/IEC 27000 Norm wird in Kapitel 0.2 Zweck dieses Dokuments die ISO/IEC 27000-Reihe dabei wie folgt eingeführt:

> "Die ISMS-Normenfamilie beinhaltet Normen, die:
>
> a) Anforderungen an ein ISMS und an diejenigen, die solche Systeme zertifizieren, festlegen;
>
> b) direkte Unterstützung, detaillierte Anleitung und/oder Interpretationen für den Gesamtprozess zur Errichtung, Umsetzung, Aufrechterhaltung und Verbesserung eines ISMS anbieten;
>
> c) sektorspezifische Anleitungen für ISMS adressieren; und
>
> d) Konformitätsbewertung für ISMS adressieren."

[International Organization for Standardization 2020a]

2 https://khcoder.net/en/

Die Texte der Standards wurden jeweils extrahiert und alle Dokumente wurden zu einem unformatierten Text zusammengefügt, der von KH Coder verarbeitet werden kann.

Zur quantitativen Betrachtung wurden in englischer Sprache hier herangezogen:

- ISO/IEC 27000:2018 – ISMS – Overview and vocabular: Enthält die nötigen Definitionen zu den jeweiligen Begriffen
- ISO/IEC 27001:2013 – ISMS – Requirements: Wird Hauptsächlich als Referenz in der TISAX genannt
- ISO/IEC 27002:2005 – Code of practice for information security management: Ist ebenfalls als Referenz aufgeführt
- ISO/IEC 27004:2016 – Information security management measurements: Unterstützt durch Richtlinien und Prozesse die Messbarkeit von Informationssicherheit und ergänzt damit die Datengrundlage um Best Practices
- ISO/IEC 27017:2015 – Code of practice for information security controls based on ISO/IEC 27002 for cloud services: Ergänzt die ISO/IEC 27002 um Security in der Cloud

Insgesamt wurden in 411 Kapiteln 1929 Sätze erkannt. Aus dem Gesamtdokument von knapp 70.000 Wörtern wurden nach der Bereinigung 30.000 Token verwendet. Damit liegt die durchschnittliche Zahl an Sätzen je Kapitel bei 4,69; die durchschnittliche Zahl an Wörtern je Satz bei 36,17.

2.2.2 IT-Grundschutz

Das zweite zu untersuchende Dokument ist der IT-Grundschutz, herausgegeben vom BSI. Die BSI-Standards 200-1, 200-2 und 200-3 enthalten Empfehlungen zu Methoden, Prozessen und Verfahren sowie Vorgehensweisen und Maßnahmen zu unterschiedlichen Aspekten der Informationssicherheit. Dabei werden im BSI-Standard 200-1 die allgemeinen Anforderungen an ein ISMS definiert und vom BSI-Standard 200-2 mit Methodik zum Aufbau eines ISMS ergänzt. Zusätzlich beinhaltet der BSI-Standard 200-3 die risikobezogenen Arbeitsschritte für die Umsetzung des IT-Grundschutz [Bundesamt für Sicherheit in der Informationstechnik 2022]. Für die Analysen lag jeweils die Version 1.0 von Oktober 2017 in englischer Sprache vor.

Wie in Kapitel 2.2.1 wurde auch hier der Inhalt als unformatierter Text extrahiert und mit KH Coder aufbereitet, um ein Datenverständnis durch Text Mining und die visuelle Repräsentation der Zusammenhänge zu erhalten. Die Dokumente bestehen gesamt aus 171 Kapiteln mit insgesamt 3.008 Sätzen. Von der Dokumentenbasis von ca. 90.000 Wörtern wurden 34.512 Token zur statischen Analyse einbezogen. Dies ergibt 17,54 Sätze je Kapitel mit durchschnittlich 29,29 Wörtern je Satz.

2.2.3 Vergleich der statischen Analysen

Die hier genutzte ISO/IEC 27000-Reihe besteht aus ca. 70.000 Wörtern, der IT-Grundschutz aus ca. 90.000 Wörtern. Die beiden verglichenen Dokumentenreihen spielen damit in etwa in der gleichen Größenordnung, wobei die Datenmenge beim IT-Grundschutz etwas höher ist. Damit sind die verwendeten Dokumentenreihen beinahe so umfangreich wie ein Roman, so hat beispielsweise J. K. Rowlings "Harry Potter and the Philosopher's Stone" in KH Coder etwa 99.000 Wörter in 12.856 Sätzen.

Die ISO/IEC 27000-Reihe hat mit 411 Kapiteln 2,4-mal so viele Kapitel wie der IT-Grundschutz, jedoch liegt die durchschnittliche Zahl an Sätzen je Kapiteln mit 4,69 bei etwa einem Viertel des IT-Grundschutz. Grund dafür sind viele Definitionen und Aufzählungen mit Listen, gegenüber einem von Fließtext geprägten IT-Grundschutz.

In den Abbildungen 3 und 4 ist ein Co-Occurrence Network der häufigsten Wörter im Text der jeweiligen Dokumente dargestellt. Dabei sind die einzelnen Wörter als Knoten nach ihrer Nähe im Text zueinander in Subgraphen gruppiert und eingefärbt. Die Stärke der Kanten ist durch den Jaccard-Koeffizienten [Romesburg 2004], einem Wert zwischen 0 und 1, wobei 1 die maximale Co-Occurrence ist, angegeben. Beispielsweise ist in der ISO/IEC 27000-Reihe links zu erkennen, dass der Koeffizient der Nähe "access" ⇔ "user" stärker ist als der Koeffizient der Nähe "access" ⇔ "unauthorized", die beiden aber dennoch eine Co-Occurrence besitzen. Die Knoten und Kanten bilden damit ein Netzwerk, das sich besonders zur Analyse innerhalb einer Dokumentenreihe eignet. Zur Bestimmung der Platzierung der Token im [words-words] network verwendet KH Coder die Methode des "Graph Drawing by Force-directed Placement" von Fruchterman und Reingold. Diese Methode hat einheitliche Kantenlängen zum Ziel, um ästhetisch ansprechende, zweidimensionale Bilder von Graphen zu erzeugen [Fruchterman and Reingold 1991]. Zusätzlich ist die absolute Häufigkeit der Wörter in den Dokumentenreihen durch die Größe der Kreise dargestellt [Higuchi 2016].

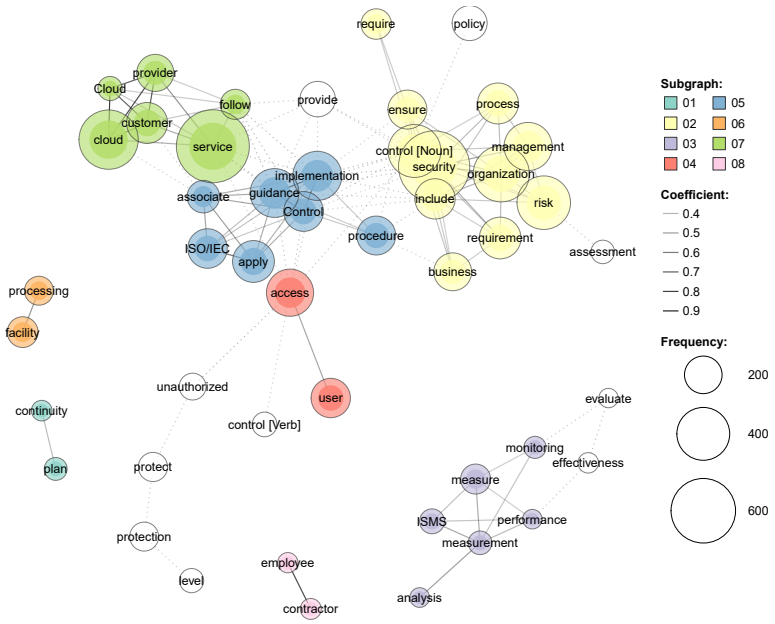


Abbildung 3 Co-Occurrence Network von ISO/IEC 27000-Reihe [Higuchi 2016][Eigene Abbildung]

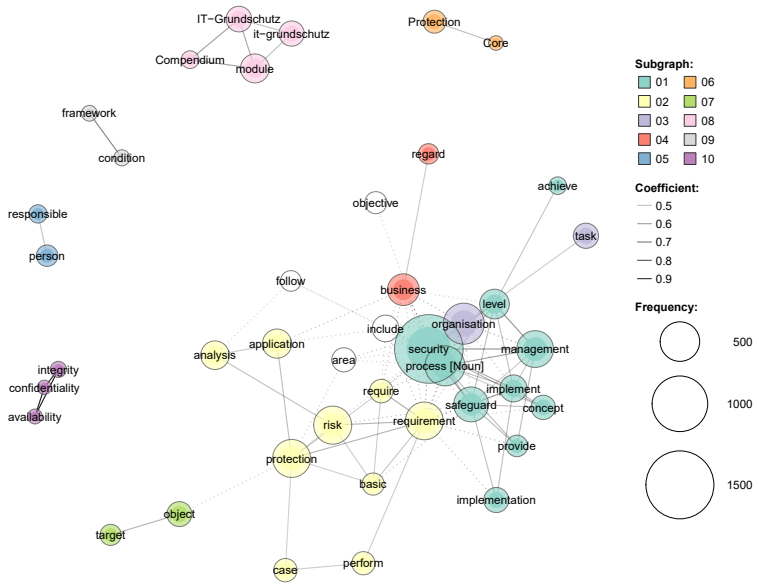


Abbildung 4 Co-Occurrence Network von IT-Grundschutz [Higuchi 2016][Eigene Abbildung]

Die Abbildungen 5 und 6 zeigen eine generierte self-organizing map of words der ISO/IEC 27000-Reihe links und des IT-Grundschutz rechts. KH Coder verwendet hier den euklidischen Abstand zur Berechnung der Distanz der häufigsten im Text vorkommenden Wörter. Die Knoten werden nach der Ward-Methode klassifiziert. Dabei werden n Ein-Element-Cluster gebildet und iterativ zu größeren Clustern zusammengeführt, bis eine gewünschte Zahl k von Clustern erreicht wird [Upton and Cook 2008]. Die entstandenen 8 Cluster (KH Coder verwendet als default k=8) werden schließlich für eine leichtere Interpretation farblich gekennzeichnet [Higuchi 2016].

Der IT-Grundschutz des BSI ist kompatibel zum internationalen Standard ISO / IEC 27001, daher liegt es nahe, dass sich auch in den Clustern Ähnlichkeiten zwischen den beiden Texten finden. So finden sich im grauen Cluster der ISO/IEC 27000-Reihe die gleichen Wörter "management", "resource", "organi[s/z]ation", "[S/s]ecurity" wie im lachsfarbenen Cluster des IT-Grundschutz. Die Schlussfolgerung ist, dass die Themen in beiden Texten ähnlich angesprochen werden.

Es finden sich auch Unterschiede: Die Themenfelder des grünen Clusters mit den Wörtern "[C/c]loud", "service", "provider" und "customer" der ISO/IEC 27000-Reihe kommen nicht oder nicht in ausreichender Stärke im IT-Grundschutz vor, sodass sie nicht auf der word map sichtbar sind. Als Schlussfolgerung wurde in der ISO/IEC 27000-Reihe mit der ISO/IEC 27017 ein Schwerpunkt auf Informationssicherheit in der Cloud gelegt, der im IT-Grundschutz nicht enthalten ist.

Abbildung 5 Self-organizing word map der ISO/IEC 27000-Reihe [Higuchi 2016][Eigene Abbildung]

Abbildung 6 Self-organizing word map des IT-Grundschutz [Higuchi 2016][Eigene Abbildung]

3 Datenvorbereitung

Bevor die Daten analysiert und in Machine Learning (ML) Modellen verwendet werden können, müssen diese in ein geeignetes, maschinenlesbares Format gebracht und für die weitere Verarbeitung vorbereitet werden. Dieser Vorgang wird Data Preparation genannt und muss auf die individuellen Anforderungen des Zielsystems angepasst werden, um eine qualitativ hochwertige Datengrundlage zu erhalten. Im Folgenden werden die Texte der Datenquellen TISAX, ISO/IEC 27000-Reihe und IT-Grundschutz transformiert, sodass eine maschinelle Weiterverarbeitung möglich ist [Webb 2010].

In Kapitel 3.1 werden die Kontrollfragen des Blattes "Informationssicherheit" des TISAX-Fragenkatalogs in eine YAML transformiert. Anschließend wird in Kapitel 3.2 der Text der ISO/IEC 27000-Reihe und des IT-Grundschutz extrahiert.

3.1 Überführung des TISAX-Fragenkatalogs in ein maschinenlesbares Format

Bei der maschinellen Verarbeitung des TISAX Fragenkatalogs treten aufgrund der Veröffentlichung als Arbeitsmappe eines Tabellenkalkulationsprogramms einige Probleme auf. Wie in Kapitel 2.1 erkannt, geben die einzelnen Spalten eine explizite Struktur der Informationen vor, jedoch findet sich innerhalb der Zellen eine spaltenspezifische implizite Struktur, die vor der maschinellen Verarbeitung vereinheitlicht und aufbereitet werden muss.

Die Spalten lassen sich als comma-separated values (CSV) exportieren, das anschließend mit Skripten weiter verarbeitet werden kann. Menschliche Fehler in Formatierung und Rechtschreibung mussten manuell nachgebessert werden, beispielsweise fehlende Leerzeichen oder Buchstaben. Das Zielformat soll auf Schlüssel-Wert Paaren aufbauen und Verschachtelung und Listen unterstützen, um sowohl die Kapitelstruktur der TISAX als auch die über- und untergeordneten Listen der Anforderungen wiedergeben zu können.

Für den Austausch und die Persistenz der Schlüssel-Wert Paare fällt die Wahl zunächst auf das Datenformat YAML [döt Net et al. 2021], da es gegenüber JSON [T. Bray 2017] aufgrund der kompakteren und übersichtlicheren Syntax leichter von Menschen lesbar ist. Dies ist eine leicht revidierbare Entscheidung, denn YAML ist ein Superset von JSON. Auch eine Überführung in XML [Thompson and Lilley 2014] ist möglich und langfristig sinnvoll, wenn zusätzlich zum Datenformat selbst ein Schema zur Validierung der Struktur explizit gewünscht ist.

Die Abbildung 7 zeigt einen Vorschlag, wie der TISAX Fragenkatalog als maschinenlesbares Format in Form einer YAML aussehen könnte. Dabei werden die Spaltennamen als Key und die Zelle als Value festgelegt. In den Anforderungsspalten ist das ’+’ Zeichen zu einer "Hauptanforderung", die Unterpunkte sind eine Liste von "Zusatzanforderungen" zu der jeweiligen Hauptanforderung. Bei der Spalte "Reference to other standards" werden die Referenzen als Liste von Kapiteln unter dem Dokument als Key dargestellt. Die Sortierung erfolgt dabei nach dem Key "ISA New", der als Identifier der jeweiligen Kontrollfrage dient. Der Vorschlag ist hierbei keine endgültige Version, sondern ein Proof of Concept, wie die ENX Association ein mensch- und maschinenlesbares Format des TISAX-

Fragenkatalogs herausbringen könnte. Verbesserungen sind insbesondere im Bereich der Granularität der vier Anforderungsspalten möglich. Auch wurden die "Support" Spalten und "Weitere Informationen" bei der ersten Iteration nicht berücksichtigt, da sie als Ausfüll- und Auslegungshilfe nicht die essenziellen Informationen enthalten.

Die Überführung zum maschinenlesbaren Format erfolgte sowohl für die deutsche als auch die englische Sprache. Für die weitere Verarbeitung wird das englische Dokument verwendet, um quellenübergreifend eine einheitliche Sprache zu verwenden.

3.2 Extraktion der Daten aus dem Fließtext

Die ISO/IEC 27000-Reihe und der IT-Grundschutz werden jeweils als Fließtext als PDF veröffentlicht. Zu Beginn werden die Seiten der Dokumente mit Hilfe des Tools pdfplumber[3] auf den Haupttext zugeschnitten und der unformatierte Text extrahiert. Anschließend werden mit Hilfe von regular expressions die Kapitel und Unterkapitel des Textes rekursiv in ein Dictionary, ein assoziatives Array mit Schlüssel-Wert Paaren [van Rossum 2022], transformiert, das dann in ein strukturiertes Datenformat, beispielsweise YAML, geschrieben werden kann. Somit liegen auch die ISO/IEC 27000-Reihe und der IT-Grundschutz als maschinenlesbare, nach Kapiteln strukturierte Daten vor und können weiterverarbeitet werden. Für eine einheitliche Datengrundlage wurden jeweils die englischen Varianten der Texte gewählt.

Aufzählungen innerhalb des Textes (Abbildung 8a) werden mit dem bisherigen Stand nicht erkannt und entsprechend strukturiert (Abbildung 8b), dies ist für den Einsatz in Produktivsystemen wünschenswert, um eine höhere Granularität der Struktur der Daten innerhalb einzelner Kapitel zu erzielen. Für die Verarbeitung im Rahmen dieser Arbeit ist die durchgeführte Vorbereitung dennoch ausreichend.

3 https://github.com/jsvine/pdfplumber

```yaml
- ISA New: '1'
  Kapitel: IS Policies and Organization
  Unterkapitel:
  - ISA New: '1.1'
    Unterkapitel: Information Security Policies
    Fragen:
    - ISA Classic: '05.1'
      ISA New: '1.1.1'
      Kontrollfrage: Inwieweit sind Richtlinien zur Informationssicherheit
        vorhanden?
      Ziel: Die Organisation benötigt mindestens eine Richtlinie für
        Informationssicherheit. Diese spiegelt die Wichtigkeit und Bedeutung der
        Informationssicherheit wider und ist an die Organisation angepasst. Weitere
        Richtlinien können je nach Organisationsgröße und -struktur sinnvoll sein.
      Anforderungen  (muss):
      - Hauptanforderung: Die Anforderungen an die Informationssicherheit
          sind ermittelt und dokumentiert.
        Zusatzanforderungen:
        - Die Anforderungen sind an die Ziele der Organisation angepasst.
        - Eine Richtlinie ist erstellt und von der Organisationsleitung
          freigegeben.
      - Hauptanforderung: Die Richtlinie enthält Ziele und den Stellenwert
          der Informationssicherheit in der Organisation.
      Anforderungen  (sollte):
      - Hauptanforderung: Die Anforderungen an die Informationssicherheit
          auf der Grundlage der Organisationsstrategie, Gesetzen und Verträgen
          sind in der Richtlinie berücksichtigt.
      - Hauptanforderung: Verantwortlichkeiten für die Durchführung sind
          definiert.
      - Hauptanforderung: Die Richtlinie weist auf Konsequenzen bei Nichtbeachtung
          hin.
      - Hauptanforderung: Weitere relevante Richtlinien zur Informationssicherheit
          sind erstellt.
      - Hauptanforderung: Eine regelmäßige Prüfung und - falls notwendig
          - Überarbeitung der Richtlinien sind etabliert.
      - Hauptanforderung: Die Richtlinien werden Mitarbeitern in geeigneter
          Form (z. B. Intranet) zur Verfügung gestellt.
      - Hauptanforderung: Die Richtlinien werden fallbezogen (ggf. auch
          in Auszügen) an externe Geschäftspartner weitergegeben.
      - Hauptanforderung: Mitarbeiter und externe Geschäftspartner werden
          über für sie relevante Änderungen informiert.
      Adressierte Schutzziele:
      - Vertraulichkeit
      - Integrität
      - Verfügbarkeit
      Referenz zu anderen Standards:
      - ISO 27001: ['A.5.1.1', 'A.5.1.2']
  - ISA New: '1.2'
    Unterkapitel: Organization of Information Security
    Fragen:
    - ISA Classic: '01.1'
      ISA New: '1.2.1'
      Kontrollfrage: Inwieweit wird in der Organisation Informationssicherheit
        gemanagt?
      Ziel: Nur wenn Informationssicherheit in den strategischen Zielen
        einer Organisation verankert ist, kann Informationssicherheit nachhaltig
        in einer Organisation umgesetzt werden. Das Informationssicherheits-
        managementsystem (ISMS) ist ein Steuerungsinstrument für die Organisations-
        leitung, mit dem sie sicherstellt, dass Informationssicherheit nicht nur
        ein Ergebnis von Zufällen und individuellem Engagement, sondern
        von nachhaltigem Management ist.
      Anforderungen  (muss):
      - Hauptanforderung: Der Geltungsbereich (Scope) des ISMS (die vom
          ISMS gemanagte Organisation) ist festgelegt.
      - Hauptanforderung: Die Anforderungen der Organisation an das ISMS
          sind ermittelt.
      - Hauptanforderung: Die Organisationsleitung hat das ISMS beauftragt
          und freigegeben.
      - Hauptanforderung: Das ISMS stellt der Organisationsleitung geeignete
          Kontroll- und Steuerungsmittel zur Verfügung (z. B. Management-Review).
      - Hauptanforderung: Anwendbare Kontrollen wurden ermittelt (z. B.
          ISO 27001 Statement of Applicability, ausgefüllter ISA Katalog).
      - Hauptanforderung: Die Wirksamkeit des ISMS wird regelmäßig durch
          das Management überprüft.
      Adressierte Schutzziele:
      - Vertraulichkeit
      - Integrität
      - Verfügbarkeit
      Referenz zu anderen Standards:
      - ISO 27001: ['4']
```

Abbildung 7 TISAX als YAML [Verband der Automobilindustrie e.V. 2021][eigene Abbildung]

4.2.5 Management system

A management system uses a framework of resources to achieve an organization's objectives. The management system includes organizational structure, policies, planning activities, responsibilities, practices, procedures, processes and resources.

In terms of information security, a management system allows an organization to:

a) satisfy the information security requirements of customers and other stakeholders;

b) improve an organization's plans and activities;

c) meet the organization's information security objectives;

d) comply with regulations, legislation and industry mandates; and

e) manage information assets in an organized way that facilitates continual improvement and adjustment to current organizational goals.

(a) Kapitel 4.2.5 "Management system" als PDF [International Organization for Standardization 2018]

```yaml
- section: '4'
  title: Information security management systems
  subsections:
  - section: '4.1'
    # ...
  - section: '4.2'
    title: What is an ISMS?
    subsections:
      # ...
    - section: 4.2.5
      title: Management system
      text: >
        A management system uses a framework of resources to achieve an
        organization's objectives. The management system includes organizational
        structure, policies, planning activities, responsibilities, practices,
        procedures, processes and resources. In terms of information security, a
        management system allows an organization to: a) satisfy the information
        security requirements of customers and other stakeholders; b) improve an
        organization's plans and activities; c) meet the organization's
        information security objectives; d) comply with regulations, legislation
        and industry mandates; and e) manage information assets in an organized
        way that facilitates continual improvement and adjustment to current
        organizational goals
```

(b) Kapitel 4.2.5 "Management system" als YAML [International Organization for Standardization 2018][eigene Abbildung]

4 Vorgehen zur Extraktion von Daten mit Hilfe von NLP

Im Rahmen dieser Arbeit wird auf bereits existierende ML Frameworks zugegriffen, um die Flexibilität der Frameworks zu nutzen und schneller Ergebnisse vorweisen zu können. Zusätzlich sind erprobte Vorgehensweisen und ML Algorithmen bereits implementiert und breit getestet. Die selbst gewählten Anforderungen an das Framework sind hierbei die bereits vorhandene Nutzung im wissenschaftlichen Kontext, das Vorhandensein vortrainierter Modelle, sowie die Möglichkeit des fine trainings dieser mit entsprechender Flexibilität bei Hyperparametern, ein Interface zum Trainieren eigener Modelle und schließlich die Nachvollziehbarkeit der Ergebnisse, soweit dies im ML Kontext möglich ist. Das Framework soll auf Natural Language Processing (NLP) zugeschnitten sein und die gewünschten NLP-Teildisziplinen unterstützen. Im Folgenden wird auf die Auswahl der Teilgebiete eingegangen und die verwendete Methodik und das gewählte Framework vorgestellt.

Für die Extraktion strukturierter Daten aus natürlichsprachlichen Texten wurde sich für die Machine Learning-Disziplin des Natural Language Processing entschieden. Im Folgenden werden die Dokumente in Kapitel 4.1 mittels Named Entity Recognition mit einem pre-trained model sowie mit einem fine-trained model untersucht. In Kapitel 4.2 erfolgt die Erkennung der lexikalischen Kategorie und der Beziehungen in einem Satz. In Kapitel 4.3 soll geprüft werden, ob eine Zusammenfassung der TISAX mittels NLP möglich ist. Nach einer Untersuchung der inhaltlichen Ähnlichkeit innerhalb der TISAX sowie der Ähnlichkeit zwischen der TISAX, der ISO/IEC 27000-Reihe und dem IT-Grundschutz in Kapitel 4.4 wird dann in Kapitel 4.5 ein Fazit zu den jeweiligen NLP-Teildisziplinen gezogen.

4.1 Named Entity Recognition

Named Entity Recognition (NER) ist ein Teilgebiet der Token Classification. Ziel ist die Erkennung von named entities (deutsch: Eigennamen), dies können Personen, Orte, Zahlen und weitere sein. Als Ergebnis wird der annotierte Text zurückgegeben. Aufgrund der guten Evaluationsergebnisse und der einfachen Verwendung wird die Library spaCy von Explosion[4] und das englische Transformer-Model `en_core_web_trf-3.2.0`[5] verwendet. Ziel ist die Erkennung und Extrahierung bedeutsamer Informationen im Kontext der Informationssicherheit durch die Hervorhebung von Eigennamen. In den Unterkapiteln wird dazu jeweils die verwendete Methodik dargelegt und die Eignung im Kontext von Informationssicherheits-Audits bewertet.

4.1.1 Verwendung eines pre-trained models

Zu Beginn wird das frei verfügbare, pre-trained Model `en_core_web_trf-3.2.0` eingesetzt. Ein pre-trained Model ist ein Machine Learning Model, das als fertig trainiertes Model zur Verfügung steht. Die Datengrundlage des gewählten Modells ist OntoNotes 5 [Weischedel, Ralph et al. 2013], ClearNLP Constituent-to-Dependency Conversion [Jinho D. Choi 2012], WordNet 3.0 [Fellbaum 2000] und RoBERTa Base [Liu et al. 2019]. Als Ergebnis werden bei der Betrachtung der Spalten 'Control

4 https://explosion.ai
5 https://spacy.io/models/en#en_core_web_trf

question', 'Objective', 'Requirements (must)', 'Requirements (should)', 'Additional requirements for high protection needs', 'Additional requirements for very high protection needs' aus insgesamt etwa 8000 Token nur 31 named entities der Kategorien LAW (9), CARDINAL (7), ORG (7), ORDINAL (5) und DATE (3) erkannt.

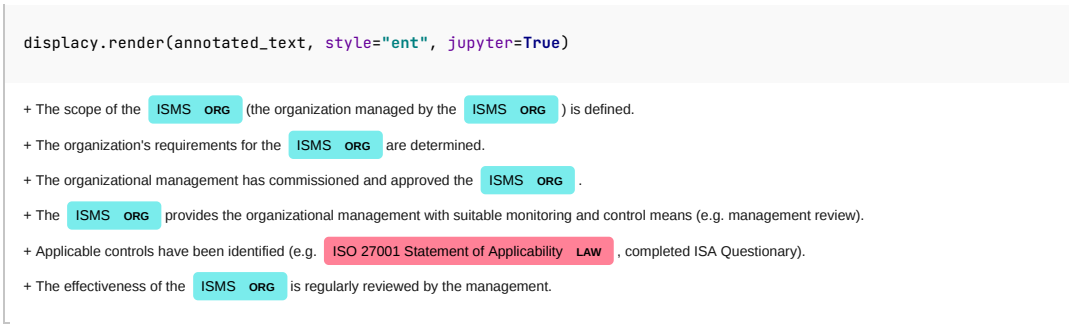


Abbildung 9 Visualisierung der named entities mit displaCy
[Verband der Automobilindustrie e.V. 2021][Explosion 2021][eigene Abbildung]

Die Abbildung 9 zeigt eine Visualisierung von erkannten named entities der 'Requirements (must)' der Kontrollfrage 1.2.1 mit displaCy. Hier ist zu sehen, dass die sieben Erkennungen von ORG (Erklärung: Companies, agencies, institutions, etc.) auf das Token "ISMS" zurückzuführen sind, wobei Informationssicherheitsmanagementsystem im Kontext der Informationssicherheit keine Organisation ist, sondern Politik, Verfahren, Richtlinien, Ressourcen und Tätigkeiten zum Schutz der Informationswerte einer Organisation umfasst [International Organization for Standardization 2020b]. Positiv ist jedoch die Erkennung der Token 'ISO 27001 Statement of Applicability' als LAW. Bei der weiteren Untersuchung zeigt sich auch, dass beispielsweise das Token "third" als ORDINAL markiert wird, im Text jedoch meist in Kombination mit dem Token "party" auftritt.

Da das Modell auf generischen Daten trainiert wurde, ist es für Texte gut geeignet, die kein Nischenthema belegen, nicht jedoch für die hier vorliegende Schnittmenge aus Normen und Texten über Informationssicherheit. Somit erfolgt nur eine unzureichende Erkennung von named entities im Kontext Informationssicherheit, die named entities sind nicht auf die TISAX angepasst.

4.1.2 Fine training eines language models

Alternativ zum Einsatz eines fertigen Models kann auch auf einem bestehenden language model aufgebaut werden, sodass dieses mit eigenen Daten weiter trainiert wird. Eigene Untersuchungen zeigen auf, dass ein tieferes Verständnis über die gewählte Domäne notwendig ist, um Klassen von Entitäten zu definieren und anschließend die Daten zu annotieren. Dazu wurde ein Experteninterview mit Dennis Aufderheide geführt, der als Mitglied einer Arbeitsgruppe der ENX Association den TISAX Fragenkatalog inhaltlich weiterentwickelt und beim Automobilzulieferer IAV GmbH[6] als Cyber Security Officer die externen Anforderungen sowie Best Practices umsetzt. Zu dieser Tätigkeit gehört auch die Begleitung des TISAX Audits.

Aus dem Experteninterview hat sich ergeben, dass ein Entwickeln eigener NER-Klassen nicht möglich

6 https://www.iav.com

sei, da die Fragen und die Inhalte sich stark voneinander unterscheiden und in der gemeinsamen Schnittmenge der Themenfelder Informationssicherheit und Normen die Unterthemen stark individuell sind [Friedrichs and Aufderheide 2022].

4.2 Part-of-Speech-Tagging

Beim Part-of-Speech (POS)-Tagging, ebenfalls ein Teilgebiet der Token Classification, erkennt das Sprachmodell die lexikalische Kategorie (Wortart) sowie die Beziehung zwischen Token in einem Satz. Das Ergebnis ist ein annotierter Satz, der als Hilfestellung für Analysen und die weitere Verarbeitung dienen kann. Abbildung 10 zeigt eine Tabelle von erkannten Tags der ersten Anforderung in 'Requirements (must)' der Kontrollfrage 1.1.1 des Fragenkatalogs. Die Spalte "UNIVERSAL_TAG" folgt den Universal POS tags [Universal Dependencies 2021], die spezifischen Tags ("POS_TAG") und Abhängigkeiten ("DEPENDENCY") werden durch den Aufruf der Methode `spacy.explain(term)` erklärt.

TEXT	LEMMA	UNIVERSAL_TAG	POS_TAG	DEPENDENCY	IS_ALPHA	IS_STOP
The	the	DET	DT	det	TRUE	TRUE
requirements	requirement	NOUN	NNS	nsubjpass	TRUE	FALSE
for	for	ADP	IN	prep	TRUE	TRUE
information	information	NOUN	NN	compound	TRUE	FALSE
security	security	NOUN	NN	pobj	TRUE	FALSE
have	have	AUX	VBP	aux	TRUE	TRUE
been	be	AUX	VBN	auxpass	TRUE	TRUE
determined	determine	VERB	VBN	ROOT	TRUE	FALSE
and	and	CCONJ	CC	cc	TRUE	TRUE
documented	document	VERB	VBN	conj	TRUE	FALSE
.	.	PUNCT	.	punct	FALSE	FALSE

Abbildung 10 Tabelle der POS-Tags und Abhängigkeiten einer Anforderung
[Verband der Automobilindustrie e.V. 2021][Explosion 2021][eigene Abbildung]

Bei den annotierten Zuordnungen von Wörtern und Abhängigkeiten innerhalb der Sätze wurden bei einer Stichprobe von 10 zufällig gewählten Anforderungen keine Fehler erkannt, somit ist das POS-Tagging für eine weitere Verarbeitung gut geeignet.

4.3 Summarization

Summarization ist eine Methode, um ein Dokument zu verkürzen, während die Informationen im Text erhalten bleiben, sodass die Informationsdichte erhöht wird. Nach dem Experteninterview (siehe Kapitel 4.1.2) hat sich ergeben, dass hinter den Fragen Formalismen stehen würden, darunter auch, die Kontrollfragen und die Anforderungen so kurz wie möglich zu halten, ohne Beispiele über technische Implementierungen oder Ähnliches zu nennen. Somit würden die jeweiligen Kontrollfragen bereits überwiegend die maximale Informationsdichte enthalten und eine weitere Zusammenfassung würde zu Informationsverlust führen [Friedrichs and Aufderheide 2022].

Ziel	Zusammenfassung des Ziels
The organization needs at least one information security policy. This reflects the importance and significance of information security and is adapted to the organization. Additional policies may be appropriate depending on the size and structure of the organization.	An information security policy is a set of rules and regulations designed to protect the confidentiality, integrity and availability of information.

Abbildung 11 Gegenüberstellung von Ziel (links) und Zusammenfassung des Ziels (rechts) durch das model `google/pegasus-xsum` [Verband der Automobilindustrie e.V. 2021][Zhang et al. 2019][eigene Abbildung]

Dies deckt sich mit den stichprobenartigen Versuchen einer Zusammenfassung von einzelnen Kontrollfragen. Abbildung 11 zeigt die Summarization des Ziels der ersten Kontrollfrage, wobei links das Ziel und rechts die Zusammenfassung des `google/pegasus-xsum` models [Zhang et al. 2019], dass beispielhaft aufgrund der hohen Verbreitung gewählt wurde, steht. Die Thematik des Ziels der Kontrollfrage wurde korrekt erfasst, jedoch wurde das Ziel so weit abstrahiert, dass die Information über die Notwendigkeit für Informationssicherheitsrichtlinien verloren gegangen ist. Als Ergebnis ist die Zusammenfassung damit nicht geeignet zur Erkennung strukturierter Daten in der TISAX.

4.4 Sentence Similarity

Eine weitere Teildisziplin des NLP ist Sentence Similarity, also die Bewertung der Ähnlichkeit zweier Texte. Dabei werden Wörter, Sätze, Absätze oder Texte in Vektoren konvertiert, um dann durch die Entfernung der Token den Koeffizienten der Ähnlichkeit zu berechnen. Im Folgenden wird zuerst die Ähnlichkeit der Kontrollfragen zueinander und anschließend die Ähnlichkeit der Kontrollfragen zur ISO/IEC 27000-Reihe und dem IT-Grundschutz mit Gensim[7] untersucht.

4.4.1 Analyse innerhalb der TISAX

Für eine Analyse der Ähnlichkeit der Kontrollfragen untereinander wurde durch Gensim auf das Lsi-Model [Deerwester et al. 1990] zugegriffen. Die Datengrundlage bilden die Spalten "Control question", "Objective", "Requirements (must)", "Requirements (should)", "Additional requirements for high protection needs" und "Additional requirements for very high protection needs" der TISAX. Den Corpus bilden die durch das Trennzeichen "." zusammengefügten Spalten und jedes Dokument wurde mittels `gensim.utils.simple_preprocess(doc)`[8] in eine Liste von lowercase Token konvertiert. Anschließend wurde für jedes Liste von Token das Cosinus-Maß {-1; 1} der Ähnlichkeit zu den anderen Kontrollfragen berechnet [Řehůřek and Sojka 2010].

Abbildung 12 zeigt einer Heatmap der Ähnlichkeit der Kontrollfragen zueinander, wobei sowohl auf der X- als auch auf der Y-Achse jeweils die Kontrollfragen mit dem Label "ISA New" stehen. In der Ähnlichkeitsmatrix ist sehr gut erkennbar, dass die einzelnen Unterpunkte meist Ähnlichkeitscluster bilden, beispielsweise 4.1.1, 4.1.2 und 4.1.3, und sich dabei von anderen Kontrollfragen, beispielsweise aus der Gruppe 2, absetzen. Ebenfalls auffällig ist, dass die Kontrollfragen 1.6.1 und 3.1.2 durch

7 https://radimrehurek.com/gensim/
8 https://radimrehurek.com/gensim/utils.html#gensim.utils.simple_preprocess

eine niedrige Ähnlichkeit zu anderen Kontrollfragen für sich stehen, jedoch gegenseitig eine sehr hohe Ähnlichkeit besitzen.

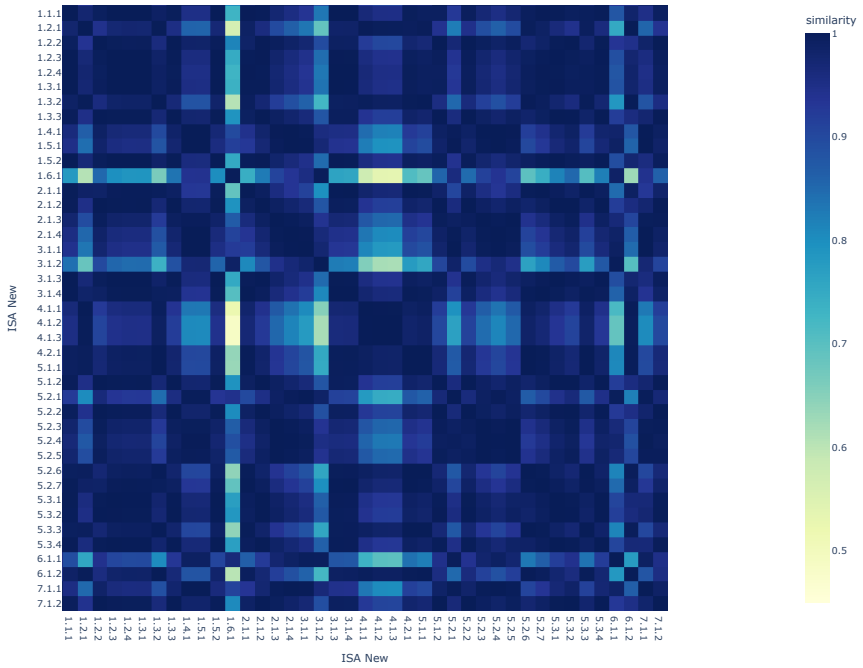


Abbildung 12 Heatmap der Ähnlichkeit der Kontrollfragen
[Verband der Automobilindustrie e.V. 2021][Řehůřek and Sojka 2010][eigene Abbildung]

Sentence Similarity ist damit gut geeignet, um innerhalb des TISAX Fragenkatalogs Ähnlichkeiten zu identifizieren und diese für weitere Analysen zu nutzen. Für die automatisierte Extraktion strukturierter Daten hilft Sentence Similarity innerhalb des Dokuments jedoch nicht, da bereits eine Struktur in Form von Indizes vorliegt und damit die Granularität der obigen Analyse erreicht ist.

4.4.2 Ähnlichkeit zwischen TISAX und ISO/IEC 27000-Reihe und IT-Grundschutz

Zusätzlich zur Ähnlichkeit der Kontrollfragen untereinander soll auch die Ähnlichkeit der TISAX Kontrollfragen zur ISO/IEC 27000-Reihe sowie zum IT-Grundschutz untersucht werden. Da in der TISAX die Nähe zur ISO/IEC 27000-Reihe sehr geschätzt wird, können so zusätzlich zur Spalte "Reference to other standards" weitere Textpassagen mit Bezug zur jeweiligen Kontrollfrage identifiziert werden [Friedrichs and Aufderheide 2022].

Dafür werden die Texte aus Kapitel 3.2 mit `gensim.utils.simple_preprocess(doc)` vorbereitet. Für ein besseres Ergebnis werden außerdem jeweils die Kapitel "Scope", "Normative References" und "Terms and Definitions" aus dem Corpus herausgenommen, da diese bei der Suche nach inhaltlichen Übereinstimmungen keinen Mehrwert bieten, sondern Begriffe innerhalb des Textes definieren oder erklären, wodurch jedoch keine Information über die Definition oder Erklärung hinaus erlangt wird.

Als Model wird Doc2Vec [Le and Mikolov 2014] durch Gensim geladen. Anschließend wird die Spalte "Objective" als Input-Query verwendet, um mit `model.dv.most_similar(keys)`[9] die 10 ähnlich-

9 https://radimrehurek.com/gensim/models/keyedvectors.html#gensim.models.keyedvectors.KeyedVectors.most_similar

sten Elemente aus dem Corpus zu erhalten. Diese werden dann der TISAX YAML Datei (siehe Kapitel 3.1) angehängt und können durch nachfolgende Schritte weiterverarbeitet werden. Die Abbildung 13 zeigt die Liste der identifizierten Kapitel sowie die Bewertung der Genauigkeit, gerundet auf fünf Nachkommastellen, die Berechnung erfolgte dabei in der englischen Sprache.

```yaml
ISA New: '1.1.1'
Kontrollfrage: Inwieweit sind Richtlinien zur Informationssicherheit
  vorhanden?
Ziel: Die Organisation benötigt mindestens eine Richtlinie für
  Informationssicherheit. Diese spiegelt die Wichtigkeit und Bedeutung der
  Informationssicherheit wider und ist an die Organisation angepasst. Weitere
  Richtlinien können je nach Organisationsgröße und -struktur sinnvoll sein.
Anforderungen:
# ...
Adressierte Schutzziele:
# ...
Referenz zu anderen Standards:
- ISO 27001:
  - A.5.1.1
  - A.5.1.2
Ähnlichkeiten:
- 'ISO 27002: 15.1': 0.38716
- 'ISO 27002: 10.2.3': 0.38081
- 'ISO 27002: 15.1.1': 0.37917
- 'ISO 27002: 8.1.3': 0.37761
- 'ISO 27001: 5.2': 0.35993
- 'ISO 27001: 8.3': 0.35869
- 'ISO 27004: 8.7': 0.35579
- 'ISO 27002: 10.2': 0.35033
- 'ISO 27002: 7.1': 0.34831
- 'ISO 27002: 9.1': 0.34808
```

Abbildung 13 Ergänzung von Ähnlichkeiten zur TISAX YAML
[Verband der Automobilindustrie e.V. 2021][eigene Abbildung]

4.5 Ergebnisse der Datenextraktion

Aus den überprüften NLP Teildisziplinen eignen sich Named Entity Recognition und Summarization nicht, um strukturierte Daten aus dem Fragenkatalog zu extrahieren, da die TISAX für vortrainierte Modelle zu sehr in einer Nische ist, die Themenfelder innerhalb der Nische sehr breit sind und die untersuchten Spalten bereits präzise ausformuliert sind. Im Gegenzug ist Sentence Similarity grundsätzlich gut geeignet, um die TISAX in sich zu analysieren, aber auch die ISO/IEC 27000-Reihe als Quelle und den IT-Grundschutz als kompatibles und erklärendes Werk gegenüberzustellen und Anwendern und Auditoren weitere Informationen bereitzustellen.

Der vielversprechendste Ansatz unter den untersuchten NLP Teildisziplinen ist das POS-Tagging. Die durch Skripte aufbereiteten Daten aus Kapitel 5 und die verwendete Methodik bietet eine gute Grundlage für eine Überführung des Fragenkatalogs in strukturierte Daten.

5 Überführung in strukturierte Daten

In diesem Kapitel werden die bereinigten Daten schließlich in strukturierte Daten überführt. Zu Beginn wird in Kapitel 5.1 für jede Anforderung der Kontrollfragen der TISAX ein Graph erzeugt. Nach der Einführung von Behaviour-Driven Development in Kapitel 5.2 wird anschließend in Kapitel 5.3 der Graph in eine ausführbare Spezifikation übersetzt, um aus den Audit-Anforderungen automatisiert Testfälle und damit eine Vorgabe für die Umsetzung von Software zu erzeugen.

5.1 Erstellung von Graphen aus Anforderungen

Mit Hilfe der YAML aus Kapitel 3.1 können die einzelnen Anforderungen mittels POS-Tagging in einen gerichteten Graphen überführt werden. Für die Beschreibung der Graphen wird das Package Graphviz [Gansner and North 2000] eingesetzt.

Die Abbildung 14a zeigt den vollständigen DOT Quellcode der ersten Anforderung der Kontrollfrage 5.3.4, der von Graphviz wiedergegeben werden soll. Der Index des Wortes im Satz (beginnend bei 0) wird als Key festgelegt, das auf das Grundwort reduzierte Token ist das Label des Knoten. Als zusätzliche Attribute können den Knoten noch beispielsweise der POS-Tag und der ursprüngliche Text für die spätere Verarbeitung mitgegeben werden, weitere Attribute wie "is_stop" oder spezifische POS-Tags wären ebenso möglich. Die gerichtete Kante zeigt die Abhängigkeit "DEPENDENCY" innerhalb des Satzes vom Startknoten zum Endknoten.

Die Abbildung 14b zeigt den gerenderten Graphen des DOT Quellcodes (Abbildung 14a). Sichtbar sind hier das auf das Grundwort reduzierte Token als Label der Knoten sowie die Abhängigkeit als Kanten im gerichteten Graphen.

Sowohl die Erzeugung der annotierten Sätze als auch die Überführung in DOT Notation ist vollständig automatisiert möglich. Damit kann der Graph, wie in Abbildung 14c gezeigt, als mehrzeiliger String des Keys "Diagraph" an jede Kontrollfrage der TISAX-YAML angehängt werden.

5.2 Vorstellung von Behaviour-Driven Development

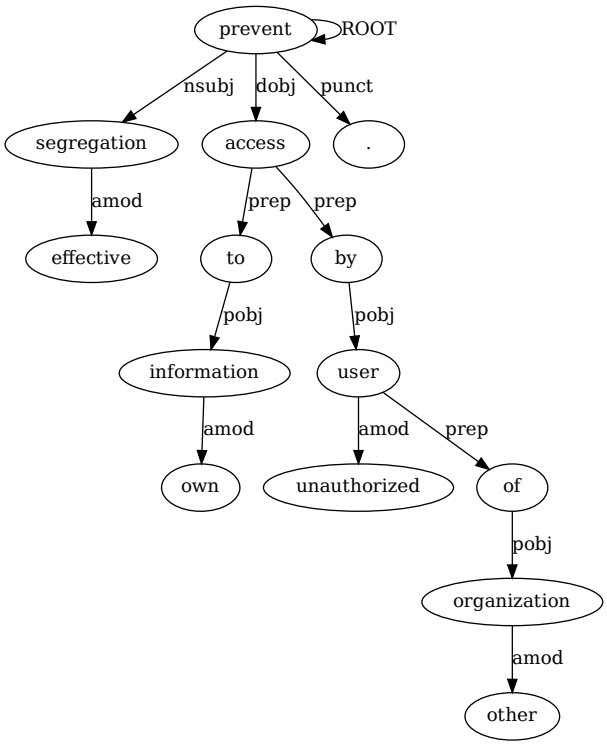
Behaviour-Driven Development (BDD) ist der Ansatz, die Lücke zwischen Softwareentwicklern und Experten einer Domäne zu schließen, in dem ein gemeinsames Verständnis über das Verhalten einer Anwendung geschaffen, Feedback durch kleine Entwicklungszyklen erhöht und Dokumentation als Grundlage für automatisierte Tests umgesetzt wird. Durch die erhöhte Kommunikation wird außerdem ein Rahmen für Konversationen gesetzt und als Ergebnis schließlich in kürzerer Zeit bessere Software entwickelt [Wynne 2019].

Das Cucumber Framework[10] kann für die Implementierung von automatisierten Tests auf Basis von ausführbaren Spezifikationen verwendet werden, die der Gherkin Syntax[11] folgen.

10 https://cucumber.io
11 https://cucumber.io/docs/gherkin/reference/

```
digraph requirement {
  0 [label=effective is_stop=False pos=ADJ tag=JJ text=Effective]
  1 -> 0 [label=amod]
  1 [label=segregation is_stop=False pos=NOUN tag=NN text=segregation]
  2 -> 1 [label=nsubj]
  2 [label=prevent is_stop=False pos=VERB tag=VBZ text=prevents]
  2 -> 2 [label=ROOT]
  3 [label=access is_stop=False pos=NOUN tag=NN text=access]
  2 -> 3 [label=dobj]
  4 [label=to is_stop=True pos=ADP tag=IN text=to]
  3 -> 4 [label=prep]
  5 [label=own is_stop=True pos=ADJ tag=JJ text=own]
  6 -> 5 [label=amod]
  6 [label=information is_stop=False pos=NOUN tag=NN text=information]
  4 -> 6 [label=pobj]
  7 [label=by is_stop=True pos=ADP tag=IN text=by]
  3 -> 7 [label=prep]
  8 [label=unauthorized is_stop=False pos=ADJ tag=JJ text=unauthorized]
  9 -> 8 [label=amod]
  9 [label=user is_stop=False pos=NOUN tag=NNS text=users]
  7 -> 9 [label=pobj]
  10 [label=of is_stop=True pos=ADP tag=IN text=of]
  9 -> 10 [label=prep]
  11 [label=other is_stop=True pos=ADJ tag=JJ text=other]
  12 -> 11 [label=amod]
  12 [label=organization is_stop=False pos=NOUN tag=NNS text=organizations]
  10 -> 12 [label=pobj]
  13 [label="." is_stop=False pos=PUNCT tag="." text="."]
  2 -> 13 [label=punct]
}
```

(a) Anforderung als Diagraph mit Annotation [eigene Abbildung]



(b) Anforderung als Diagraph gerendert [Gansner and North 2000][eigene Abbildung]

```
ISA New: '5.3.4'
Control question: To what extent is information protected in shared external
↪  IT services?
Objective: Clear segregation between individual tenants must be ensured such
↪  as to protect own information in external IT services at all times and to
↪  prevent it from being accessed by other organizations (tenants).
Requirements:
# ...
Addressed protection objectives:
# ...
Reference to other standards:
# ...
Similarities:
# ...
Diagraph: |
  diagraph requirement {
    0 [label=effective is_stop=False pos=ADJ tag=JJ text=Effective]
    1 -> 0 [label=amod]
    # ...
  }
```

(c) Diagraph in der TISAX YAML [Verband der Automobilindustrie e.V. 2021][eigene Abbildung]

5.3 Erzeugung von ausführbaren Spezifikationen

Der Graph kann in eine ausführbare Spezifikation zur Umsetzung von BDD überführt werden. Die Abbildung 15 zeigt die Kontrollfrage 5.3.4 mit der Spalte "ISA New" als Feature, sowie die Spalte "Objective" als Beschreibung des Features, der Spalte "Control question" als Regel, sowie die Spalten "Requirement (must)" und "Requirement (should)" als ausführbare Spezifikationen dieser Regel. Dabei wird der Gherkin Syntax[12] gefolgt.

Zur Erzeugung der ausführbaren Spezifikation wurde das folgende Schema manuell ausgeführt:

1. "ROOT" des Satzes wird an die erste Stelle des Then-Keywords gesetzt.
2. Das Objekt "dobj", auf das sich root bezieht, hat zwei Präpositionen "prep":
   a) Der Satzteil hinter dem Wort "to" wird zum Then-Keyword ergänzt
   b) Der Satzteil hinter dem Wort "by" wird an die Stelle des When-Keywords gesetzt.
3. Das Subjekt "nsubj" wird an die Stelle des Given-Keywords gesetzt.

Im Scenario "Requirement (should) – 1" existieren einige Sonderfälle:

- Der Satz hat kein Objekt, daher wird an die Stelle des When-Keywords ein *exists* gesetzt.
- Die Konjunktion im Satz wird durch ein And-Keyword repräsentiert.
- Die "Sub requirements" werden ebenfalls durch das And-Keyword an das Then-Keyword des Szenarios angehängt.

Dieser manuelle Algorithmus soll zeigen, wie aus automatisiert erzeugten Graphen ausführbare Spezifikationen erstellt werden können. Mit einer weiterführenden Analyse und linguistischer Expertise kann

12 https://cucumber.io/docs/gherkin/reference/

```
Feature: 5.3.4
  Clear segregation between individual tenants must be ensured such as
  to protect own information in external IT services at all times and to prevent
  it from being accessed by other organizations (tenants).

  Rule: To what extent is information protected in shared external IT services?
    Scenario: Requirement (must) -- 1
      Given effective segregation
      When access by unauthorized users of other organization
      Then prevent access to own information

    Scenario: Requirement (should) -- 1
      Given the provider segregation concept
      When exists
      Then be document
      And adapt to any change
      And separation of datum, function, application, operating system, memory and network
      And risk assesment for the operation of third party software withing the share environment

    Scenario: Requirement (should) -- 2
      Given it system
      When in share use
      Then make resilient accordingly
```

Abbildung 15 Ausführbare Spezifikation einer Anforderung der TISAX [eigene Abbildung]

auch dieser Schritt unter Beachtung der Gherkin Best Practices[13] automatisiert werden, sodass der bestehende TISAX-Fragenkatalog in entsprechende ausführbare Spezifikationen übersetzt wird. Als Ergebnis kann die TISAX also automatisch in ihrer Umsetzung mit Hilfe von BDD verifiziert werden.

13 https://cucumber.io/docs/bdd/better-gherkin/

6 Evaluation

In diesem Kapitel sollen die Ergebnisse der beiden erfolgversprechendsten Ansätze bewertet werden. In Kapitel 6.1 wird die Sentence Similarity erst aus technischer Sicht, dann aus fachlicher Sicht betrachtet. Anschließend erfolgt in Kapitel 6.2 die Evaluation der ausführbaren Spezifikation.

6.1 Evaluation der Sentence Similarity

Die Überschneidung zwischen genannter Referenz in der TISAX und erkannter Ähnlichkeit in den Quellen ist sehr gering. Betrachtet werden die 10 häufigsten Ähnlichkeiten aller Kontrollfragen. Bei einem exakten Match zwischen Referenz und erkannter Ähnlichkeit gab es keine Treffer. Bei einem Match mit Abweichung um ein direktes Eltern-, Kind-, oder Geschwisterkapitel gab es einen Treffer. Um hier verwertbare Ergebnisse zu erhalten, ist ein Expertenwissen erforderlich, sodass die Hyperparameter entsprechend eingestellt werden können, um eine optimale Überschneidung zwischen erkannter Ähnlichkeit und genannter Referenz zu erhalten.

Fachlich kann dieser Ansatz bei erfolgreicher Durchführung Anwendern dabei zu helfen, den TISAX-Fragebogen durch mehr zur Verfügung stehende Informationen umfänglicher auszufüllen und ein Verständnis der Thematik zusätzlich zur Spalte "Objective" zu erhalten. Dies senkt die Einstiegshürde in das Themengebiet Auditierung von Informationssicherheit in Unternehmen und verringert die Einarbeitungszeit und damit die Kosten, die mit der Zertifizierung verbunden sind. Ebenfalls können Auditoren durch den zusätzlichen Kontext die Antworten zu einzelnen Kontrollfragen genauer bewerten und bei der Prüfung bei unzureichend beantworteten Kontrollfragen zusätzliche Informationsquellen zur Verbesserung der Informationssicherheit im Unternehmen bieten.

6.2 Evaluation der ausführbaren Spezifikation

Die manuelle Erstellung einer ausführbaren Spezifikation mit einem Szenario je Anforderung aus dem Graphen ist einfach, da Menschen den Kontext erkennen, die Grammatik verstehen und bei Sonderfällen leicht den geforderten Transfer leisten können. In Kapitel 5.3 wurde genau dies gemacht, was zu gut lesbarer Gherkin Syntax und damit zu Erreichung des Ziels führte. Die Maschine benötigt zusätzlich zum Verständnis des Textes auch die Fähigkeit zur Erzeugung neuer Texte, die der Gherkin Syntax folgen, um den Prozess zu automatisieren. Dies ist mit dem bestehenden System nicht möglich, jedoch erforderlich für eine automatisierte Transformation aller Anforderungen in ausführbare Spezifikationen.

Mit dieser Spezifikation kann schließlich aus jedem Gherkin Szenario ein Test generiert werden, der dann in einer Testumgebung ausgeführt werden kann, beispielsweise in Java, Kotlin oder JavaScript. Ein Entwickler kann dann gegen diesen Test programmieren und somit die geforderte Business Funktionalität umsetzen. Folglich wird aus den Anforderungen der TISAX ausführbarer Code erzeugt.

7 Fazit und Ausblick

Abschließend wird in diesem Kapitel ein Fazit gezogen und ein Ausblick gegeben werden. In Kapitel 2 besitzen die ISO/IEC 27000-Reihe als Quelle und der IT-Grundschutz als deutsches Pendant außerhalb der Kapitel keine Struktur in sich. Der TISAX-Fragenkatalog, der als Tabellenkalkulations-Arbeitsmappe herausgegen wird ist in seiner Veröffentlichungsform menschenlesbar und die Struktur ist von Menschen besonders mit Expertenwissen gut erkennbar. Eine Maschinenlesbarkeit des Fragenkatalogs ist begrenzt gegeben, sodass in Kapitel 3 eine Überführung in das maschinenlesbare Format YAML erfolgte, dass aufgrund seiner Syntax auch von Menschen besser lesbar ist als eine JSON- oder XML-Datei. Eine Transformation in diese Datenformate ist weiterhin möglich.

Weiter konnte die YAML dann mittels Machine Learning automatisiert verarbeitet werden, sodass in Kapitel 4 die unterschiedlichen Teildisziplinen des NLP auf die Datenquellen angewandt werden konnten. Insbesondere funktionieren Part-of-Speech-Tagging sehr gut und Sentence Similarity mäßig gut mit schon untersuchten Algorithmen und vortrainierten Modellen. Eine Untersuchung der Daten mit fine trained Modellen bot sich nicht an, da nach einem Experteninterview die Erkenntnis erlangt wurde, dass nach aktuellem Stand sowohl Named Entity Recognition als auch Summarization nicht umsetzbar sind.

Eine Verwertung der Arbeitsergebnisse ist innerhalb der Arbeitsgruppe der ENX Association, die die TISAX weiterentwickelt möglich. So könnte eine zusätzliche Weiterentwicklung mit stärkerem Fokus auf maschinelle Verarbeitung erfolgen und an Stelle einer Arbeitsmappe für Tabellenkalkulationssoftware auf YAML oder XML als Datenformat setzen, welches dann von beliebigen Anwendungssystemen dargestellt werden und zwischen diesen ausgetauscht werden kann. Sollte sich die ENX Association langfristig für ausführbare Spezifikationen entscheiden könnte statt automatisierter Überführung einmalig der gesamte Fragenkatalog manuell überführt und eine zukünftige Weiterentwicklung in Gherkin durchgeführt werden. Dabei würden die gewonnenen Freiheiten der TISAX-YAML jedoch aufgegeben werden, solange Anwendungssysteme ausführbare Spezifikationen als Datenformat nicht unterstützen.

Außerdem können die Ergebnisse der Arbeit schließlich im sAInd Projekt als Grundlage für nachfolgende Meilensteine genutzt werden. Mit Hilfe der manuellen Durchführung kann eine automatisierte Überführung der TISAX in eine ausführbare Spezifikation entwickelt werden. Außerdem ist die TISAX-YAML als Datenformat für eine Datengrundlage in Anwendungssystemen, die die Lage der Informationssicherheit in Unternehmen beurteilen, geeignet. Neben den strukturierten Daten dienen auch die Ähnlichkeiten der Kontrollfragen zu den Normen einem kontinuierlichen Audit. Beispielsweise könnte ein Dashboard die Konformität mit dem Audit zeigen und neben den einzelnen Kontrollfragen zusätzlich Informationen aus der ISO/IEC 27000-Reihe und dem IT-Grundschutz anbieten, um die prozentuale Konformität weiter zu erhöhen oder Maßnahmen vorzuschlagen, die weiterhin eine dauerhafte Konformität gewährleisten.

Literaturverzeichnis

Bundesamt für Sicherheit in der Informationstechnik: BSI-Standard 200-1: Managementsysteme für Informationssicherheit (ISMS), Nov. 15, 2017, https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/BSI_Standards/standard_200_1.pdf,

Bundesamt für Sicherheit in der Informationstechnik: BSI-Standards, Mar. 11, 2022, https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/BSI-Standards/bsi-standards_node.html, zuletzt aufgerufen 03/14/2022.

Deerwester, Scott et al.: "Indexing by latent semantic analysis", en, in: *J. Am. Soc. Inf. Sci.* 41.6 (1990), pp. 391–407.

döt Net, Ingy et al.: Yaml Ain't markup language (YAML™) version 1.2, tech. rep., Oct. 2021, https://yaml.org/spec/1.2.2/,

ENX Association: Über ENX Association, June 25, 2022, https://portal.enx.com/de-DE/enxassociation/, zuletzt aufgerufen 06/25/2022.

ENX Association: Über TISAX, June 25, 2022, https://portal.enx.com/de-DE/TISAX/, zuletzt aufgerufen 06/25/2022.

Explosion: English transformer pipeline, Nov. 4, 2021, https://spacy.io/models/en#en%5C_core%5C_web%5C_trf,

Fellbaum, Christiane: WordNet: An electronic lexical database, MIT Press, 2000.

Friedrichs, Cedric and Dennis Aufderheide: Experteninterview zur TISAX, Online-Interview, Mar. 30, 2022.

Fruchterman, Thomas M. J. and Edward M. Reingold: "Graph drawing by force-directed placement", in: *Software: Practice and Experience* 21.11 (Nov. 1991), pp. 1129–1164, DOI: 10.1002/spe.4380211102.

Gansner, Emden R. and Stephen C. North: "An open graph visualization system and its applications to software engineering", in: *SOFTWARE - PRACTICE AND EXPERIENCE* 30.11 (2000), pp. 1203–1233.

Gleich, Florian: TISAX Participant Handbook, Apr. 7, 2022, https://www.enx.com/handbook/tisax-participant-handbook.html, zuletzt aufgerufen 05/16/2022.

Higuchi, Koichi: KH Coder 3 Reference Manual, Mar. 16, 2016, http://khcoder.net/en/manual_en_v3.pdf,

Information technology — Security techniques — Information security management systems — Overview and vocabulary, vol. 2018, Feb. 2018.

Informationstechnik - Sicherheitsverfahren - Informationssicherheitsmanagementsysteme - Überblick und Terminologie, vol. 2020, June 2020, p. 6.

Informationstechnik - Sicherheitsverfahren - Informationssicherheitsmanagementsysteme - Überblick und Terminologie, vol. 2020, June 2020.

Jinho D. Choi, Martha Palmer: Guidelines for the Clear Style Constituent to Dependency Conversion, Jan. 2012.

Le, Quoc V. and Tomas Mikolov: Distributed Representations of Sentences and Documents, 2014, DOI: 10.48550/ARXIV.1405.4053, https://arxiv.org/abs/1405.4053,

Liu, Yinhan et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019, DOI: 10.48550/ARXIV.1907.11692, https://arxiv.org/abs/1907.11692,

Martín Abadi et al.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, Software available from tensorflow.org, 2015, https://www.tensorflow.org/,

Plattform Industrie 4.0: Leitbild für Industrie 4.0, Bundesministerium für Wirtschaft und Energie (BMWi), May 22, 2019, https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/Leitbild-2030-f%C3%BCr-Industrie-4.0.pdf?__blob=publicationFile&v=11, zuletzt aufgerufen 02/07/2022.

Řehůřek, Radim and Petr Sojka: "Software Framework for Topic Modelling with Large Corpora", English, in: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, http://is.muni.cz/publication/884893/en, Valletta, Malta: ELRA, May 2010, pp. 45–50.

Romesburg, Charles: Cluster Analysis for Researchers, Lulu.com, Apr. 1, 2004, 344 pp., ISBN 1411606175, https://www.ebook.de/de/product/4208765/charles_romesburg_cluster_analysis_for_researchers.html,

T. Bray, Ed.: The JavaScript Object Notation (JSON) Data Interchange Format, RFC 8259, Internet Engineering Task Force, Dec. 2017, DOI: 10.17487/RFC8259, https://www.rfc-editor.org/rfc/pdfrfc/rfc8259.txt.pdf,

Thompson, H. and C. Lilley: XML Media Types, RFC 7303, Internet Engineering Task Force, July 2014, DOI: 10.17487/RFC7303, https://www.rfc-editor.org/rfc/pdfrfc/rfc7303.txt.pdf,

Universal Dependencies: Universal POS tags, Nov. 15, 2021, https://universaldependencies.org/u/pos/, zuletzt aufgerufen 05/16/2022.

Upton, Graham and Ian Cook: A Dictionary of Statistics, Oxford University Press, Jan. 2008, DOI: 10.1093/acref/9780199541454.001.0001.

dictionary, Python Software Foundation, May 22, 2022, https://docs.python.org/3.10/glossary.html#term-dictionary, zuletzt aufgerufen 05/22/2022.

Verband der Automobilindustrie e.V.: Information Security Assessment, Apr. 16, 2021, https://portal.enx.com/isa5-de.xlsx,

Webb, Geoffrey I.: "Data Preparation", in: *Encyclopedia of Machine Learning*, ed. by Sammut, Claude and Geoffrey I. Webb, Boston, MA: Springer US, 2010, pp. 259–260, ISBN 978-0-387-30164-8, DOI: 10.1007/978-0-387-30164-8_194, https://doi.org/10.1007/978-0-387-30164-8_194,

Weischedel, Ralph et al.: OntoNotes Release 5.0, 2013, DOI: 10.35111/XMHB-2B84.

Wynne, Matt: Behaviour-Driven Development, ed. by SmartBear, Aug. 28, 2019, https://cucumber.io/docs/bdd/,

Zhang, Jingqing et al.: PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization, 2019, arXiv: 1912.08777 [cs.CL].

Eidesstattliche Erklärung zur Bachelorarbeit

Ich erkläre hiermit, dass ich die Arbeit selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benützt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

Ingolstadt, den 07.07.2022

Cedric Friedrichs