Cyberthreats in Digital Healthcare

An exploratory analysis using text mining on news data

Master Thesis

For attainment of the academic degree of
Master of Science in Engineering (MSc)

in the Master Program Digital Healthcare
at St. Pölten University of Applied Sciences

by
Markus Bertl, BSc
1710756823

First advisor: Andreas Jakl, MSc
Second advisor: Dr. Joachim Klerx

[St. Pölten, 14.05.2019]

Declaration

I declare that I have developed and written the enclosed Master Thesis completely by myself and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. This work was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

....

Place, Date Signature

Preface

“Whenever a theory appears to you as the only possible one, take this as a sign that you have neither understood the theory nor the problem which it was intended to solve.”

— Karl Popper, 1902 - 1994

“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”

— Sir Arthur Conan Doyle, 1859 - 1930

“Information is not knowledge.”

— Albert Einstein, 1879 - 1955

Abstract

This research reviews the possibilities of text mining in the area of cybercrime in digital healthcare. Different projects already apply text mining successfully in the cyber domain. However, none of these are specifically tailored to threats in the digital healthcare sector or uses an as big data foundation for analysis. The following research aims to mine news data to find out what is reported about digital healthcare, what security-related critical events happened, and what actors, attack methods, and technologies play a role there. To achieve that goal, different text mining methodologies like fact extraction, semantic fields as well as statistical methods like frequency, correlation and trend calculations were used. The news data for the analysis was provided by the DocCenter from the National Defence Academy (DocCenter/NDA) of the Austrian Armed Forces. About 300,000 news articles were analyzed. Additionally, the open source GDELT dataset was investigated.

The data points out that cyberthreats are present in digital health technologies and cyberattacks are more and more threatening to organizations, governments, and every person them self. Not only hacker groups, firms, and governments are involved in these attacks, also terroristic organizations use cyberwarfare. That, together with the amount of technology in digital healthcare like pacemakers, IoT, wearables but also the importance of healthcare as critical infrastructure and the dependence on electronic health records makes our society vulnerable.

Keywords: digital healthcare, cybercrime, text mining, new technologies, Watson Explorer, Google Trends, GEDLT, DocCenter/NDA

Kurzfassung

Diese Publikation untersucht die Möglichkeiten die Text Mining im Bereich Cybercrime in Digital Healthcare bietet. Verschiedene Projekte verwenden Text Mining erfolgreich im Cyber Bereich allerdings nicht spezifisch adaptiert auf die Anforderungen des Gesundheitswesens. Diese Arbeit analysiert deshalb News Daten um herauszufinden was speziell in der Domäne Digital Healthcare berichtet wird, welche IT Security Vorfälle passiert sind und welche Akteure, Angriffsmethoden und Technologien hier eine Rolle spielen. Dazu wurden verschiedene Text Mining Methoden wie Fact Extraction oder semantische Felder sowie statistische Methoden wie Korrelationen oder Trend Analysen angewendet. Die Datengrundlage kam aus der Zentraldokumentation der Landesverteidigungsakademie (ZentDok/LVAk) des österreichischen Bundesheeres. Insgesamt wurden zirka 300.000 Artikel ausgewertet. Zusätzlich wurden die Metadaten des GDELT Datasets untersucht.

Die Daten zeigen, dass Technologien im Bereich Digital Healthcare stätig zunehmen und Gefahren bergen. Diese werden auch gezielt von Organisationen, Staaten und Einzelpersonen ausgenutzt. Auch Terroristengruppen bedienen sich immer mehr Methoden der digitalen Kriegsführung als Ergänzung zu klassischen Terror Angriffen. Das zeigt gemeinsam mit der Durchdringung des Gesundheitswesens von digitalen Technologien wie Herzschrittmacher, IoT, Wearables aber auch Krankenhausinformationssysteme und elektronische Patientenakten die Gefahr die auf uns zukommt.

Keywords: Digital Healthcare, Cybercrime, Text Mining, neue Technologien, Watson Explorer, Google Trends, GEDLT, ZentDok/LVAk

Table of Content

1 Introduction

1.1 Motivation

IT-based healthcare technology is on the rise creating a huge potential for better patient outcome and for a transformation of care delivery. Buzz words like e-Health, electronic health record (EHR), telemedicine or Healthcare 4.0 are increasingly used in the literature. Governments begin to build systems in order to save, connect and share health data.

Another example where IT is used in healthcare is the implantation of medical devices into the human body. About 8,000 pacemakers are implanted in Austria per year (Raatikainen et al., 2015). We use insulin pumps, cochlear implants or robotic prostheses. All this is nowadays a routine procedure. Additionally, the Internet of Things (IoT) gains also popularity in the healthcare industry (Aktypi, Nurse, & Goldsmith, 2017; Zubiaga, Procter, & Maple, 2018).

With the advance of information technology, society becomes more dependent on these systems and more and more data is stored. However, the consequences if a pacemaker or insulin pump is hacked and not working correctly (Camara, Peris-Lopez, & Tapiador, 2015), if personal health data gets published (Ponemon Institute, 2017) or if a hospital has an IT system breakdown due to a hacker attack (Mertz, 2018) can have a significant impact on healthcare companies and especially on patients. For example, an average data breach in the healthcare sector costs around 380$ per capita for a company making an average cost of 2.8 million US Dollar per data breach (Ponemon Institute, 2017). In the previous example of a hacking attack on a pacemaker or hospital infrastructure, the consequences for the patient are potentially life-threatening.

According to Marsh & McLennan Companies (2017), 25% of companies in the healthcare sector have been victims of cyberattacks in the past. These examples show how vital defence against cyberattacks is and raise the question of the security of these crucial systems we are so dependent on.

In contrast, increasing high-quality monitoring of attacks, new trends, and criminal threats has the potential to positively impact cybersecurity in the healthcare sector. Subsequently, the monitoring's output can be used to adapt the security protocols and policies of companies. To quickly detect what kind of

cyberthreats emerge in the healthcare industry, a text mining system based on news data is proposed in this research.

The literature review in chapter 2 shows that there is a research gap when it comes to applying text mining and artificial intelligence on news data to prevent cybercrime in the healthcare industry because qualitative as well as quantitative studies to the research question “Which additional value is achieved by text mining news data regarding the analysis of cyberthreats in healthcare?” are very rare. One potential reason for this is the novelty of this research. This gap in the literature underlines the importance of this work and is one factor of why it was chosen to study the impact of text mining news data on cybercrime analysis in the healthcare industry. This research is the foundation for getting periodical market insights for trends, threats, and capabilities in digital healthcare. Consequentially, this study can have a huge business impact.

1.2 Definition of important terms

- Natural Language Processing (NLP)

Natural Language Processing describes a set of methods of how computers can process and analyze natural language data (text).

- Natural Language Understanding (NLU)

Natural Language Understanding is a subtopic of NLP. After the processing of the text data, NLU is used to make sense out of the data and analyzes its meaning and context.

- Natural Language Querying (NLQ)

NLQ means that searches can be formulated as questions in natural language without knowledge of any search syntax or IT system. As a result, the NLQ algorithm generates the search query out of the question.

- Critical Infrastructure

Critical Infrastructure are assets that are essential for the basic functioning of the state, society, and economy like water supply, public health, transportation, communication or security services.

- Open source intelligence (OSINT)

Open source intelligence means the analysis of publicly available, non-classified, sources in an intelligence context. For instance data could be print media, online publication, social media, academic work economic or financial data.

- Crowd open source information (Crowd OSInfo)

Crowd OSInfo describes the process of OSINT using a group of people called crowd.

1.3 Research Questions

The overall research question is “Which additional value is achieved by text mining news data regarding the analysis of cyberthreats in healthcare?”.

In detail, it shall be analyzed what kind of news is broadcasted about digital healthcare to identify new topics and trends, when security incidents related to healthcare happened, what methods are used and when and where they were published.

The above statement results in the following, more detailed, research questions:

- What topics are important in digital healthcare?
- What events regarding digital healthcare happened and when?
    - What actors play a role in this domain?
    - What attack methods were involved?
    - What new technologies arise in digital healthcare?
- Which sources report about incidents?

1.4 Methodology

A systematic literature review according to Kitchenham & Charters (2007) was conducted to get an overview of the state of the art in cybersecurity for healthcare, relevant existing text mining projects to uncover cyberthreats, and important text mining approaches in general.

Based on the above-mentioned research questions, facts from data sources described in chapter 3 are extracted using ontologies and rule-based approaches. In a second step, the extracted information is analyzed using explorative statistics like frequency, correlation and trend analysis as well as visualization techniques like tables and diagrams. Using the correlation between facts and different datasets, topic maps and network representations can be built to provide an overview of security in digital healthcare. This approach uncovers security threats, the moment when they were first discovered and first exploited by criminals together with the caused damage.

Additionally, new trends in healthcare can be uncovered and investigated according to their potentials and risks for enhancing security. New trends are

found using frequency and timeline analysis of the terms in the data sources. This approach is described by Michel et al. (2011). The researched terms can subsequently be further analyzed in Google Trends[1] to acquire additional insights on when and where they arose. A Google Trends analysis can add further value because news are published only once but Google Trends can show how long and how often searchers looked for the news indicating the time interest in these news persisted.

The ethics committee of the Federal State of Lower Austria stated that there is no formal obligation for this research to be submitted to an ethics board.

1.5 Structure of this research

Following this introductory chapter, the second chapter presents a literature review to show the current state of scientific investigation with regards to text mining for cybersecurity in healthcare, existing projects in this area and techniques for text mining. Chapter 3 aims for describing the data sources used for this research. The mining model that is built on top of the data and the used software is analyzed in chapter 4. Basically, this model describes what and how knowledge is extracted from the corresponding text sources. Apart from the model, also the tools and method that are used are introduced. The fifth chapter deals with the empirical results of the mining process, answering the research questions and presenting, therefore, the output of the work. These results are subsequently analyzed and critically discussed in chapter 0. The seventh chapter contains a summary of the findings, the conclusion, and provides an outlook for further work in the field of investigation. The last chapter states conflicts of interests and acknowledgements.

[1] trends.google.com

2 State of the art

This chapter presents the academic importance and the current state of research in the area of knowledge generation in cybersecurity using text mining for news data to uncover best practices and to investigate what facts have been extracted in previous projects.

2.1 Cybersecurity in healthcare

Kruse, Frederick, Jacobson, & Monticone (2017) and Coventry & Branley (2018) state that the healthcare industry is a vulnerable target group for cyberattacks because due to historical reasons, the ability to defend against attacks is very low; hospital information systems are old and people could not imagine for a long time that the health industry would be a possible target. Additionally, medical devices are increasingly connected to each other, resulting in the fact that getting into one system, more devices can be affected from there. It is also easier to gain access to systems because there are more possible entry points (Coventry & Branley, 2018). As stated in the Introduction, the harm that can be done is very high. The above-mentioned facts were also established by Camara et al. (2015), O'Doherty et al. (2016) and Gagneja (2017).

According to Sulleyman (2017) and Coventry & Branley (2018), the reasons why hackers attack this industry do not differ from other industry sectors. The strongest motivation is financial gain. The value of medical data can be up to 1,000$, which is worth more than ten times the amount of credit card numbers (Sulleyman, 2017). Using this number for calculating the damage of the cyberattack against Anthem Inc., one of the largest health insurances in the US, where 87 million data sets were stolen in 2015, it can be seen that there is a substantial financial gain in attacking healthcare industries (Cheng, Liu, & Yao, 2017). In the case mentioned above, about 87 billion US Dollar damage can be estimated.

In addition to the respective problems of data theft, interference with medical devices and critical infrastructure are also a major concern. Cyberattacks have a direct effect on patients' health and safety. According to military strategist Brigadier Mag. Dr. Walter Feichtinger, 53% of the Austrian citizens are afraid of transnational terrorism (Feichtinger, 2017). Despite the high concern about

bombings and shootings, hacking critical healthcare industry is nothing people are afraid of (Mertz, 2018). Even if the concerns are not high, Anura S. Fernando, principal engineer with UL's Medical Software states, that this is a possible scenario. Especially if the hacking attack is combined with a classic terror attack. The impact would be maximized if hospital infrastructure is not working due to a cyberattack so the many injured victims cannot be treated appropriately. However, not only terrorists have an interest in hacking critical infrastructure. The cyber topic is also an important point in modern warfare and defence. Hacking critical infrastructure like healthcare, water supply or even military defence systems has been carried out by different nations. Examples for the so called cyberwarfare are the Iraq war (Saltzman, 2013), Syria (Eilstrup-Sangiovanni, 2018) or North Korea (Sanger & Broad, 2017). These examples show that cyberattacks against the healthcare industry has not just consequences for the company itself but is also dangerous from a national security perspective (Göllner et al., 2010; Mertz, 2018).

Another reason why cybersecurity is so critical in healthcare is that the industry needs a high amount of certainty and trust. In a digital lab report, only a few bytes distinguish a healthy person from one with HIV. Just one single number on a prescription can turn a lifesaving medicine into a deadly overdose. One doctor's note can change if you are allowed to decide on your own or if you get an organ donation. However, what happens if these data cannot be trusted? Should the hospital administer all tests again? Should all health record entries not be trusted? What happens if the hospital cannot access all data or is not sure if some information is missing? That is why cyberattacks like ransomware are such a big problem in healthcare. It has the potential to destroy the trust in the systems (Gagneja, 2017; Mertz, 2018). According to Mertz (2018), it is not even necessary to carry out such attacks. A simple statement of a group that they hacked a hospital could result in damaging this trust in the society, regardless if the statement was true or not. If a healthcare provider cannot prove that its security is rock solid, it makes itself vulnerable.

Not only major institutions are targets of cyberattacks, also the end-users get increasingly focused by attackers. Examples are hacking of IoT devices, wearables, personal health records or mobile health apps (Aktypi et al., 2017; Martínez-Pérez, de la Torre-Díez, & López-Coronado, 2015; Zubiaga et al., 2018).

2.2 Text mining approaches

Historically, text data was analyzed using qualitative data analysis approaches like manual coding (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010). Since the

amount of data is fast-growing, it begins to exceed the information processing capabilities of single researchers or even research teams. Additionally, manual text analysis is always biased due to the subjective interpretation of data (Indulska, Hovorka, & Recker, 2012). Text mining brings a scalable and reliable solution to these problems (Müller, Junglas, & vom Brocke, 2016). Text mining is an extension of data mining techniques focusing on gaining knowledge out of unstructured data. The fuzzy nature of language adds additional complexity compared to standard data mining approaches, for instance on homogenous numbers. Various algorithms try to deal with that problem. This section describes the most important ones.

Topic modelling is one way to discover effectively what a corpus of large text documents is about. In other words, finding out the topics that run through a text corpus. Blei (2012) describes a topic as a multinomial distribution over a fixed vocabulary. One popular topic modelling algorithm is the Latent Dirichlet Allocation (LDA) developed by Blei, Ng, & Jordan (2003). Martin & Johnson (2015) argue that topic modelling algorithms are most efficient if all non-nouns are filtered out prior to the topic modelling process. While sentiment analysis mostly focuses on adjectives or change & trend detections on verbs, topics are particularly indicated by nouns.

Classification, as one kind of supervised learning, is a mining method that classifies each text to a certain category (Irfan et al., 2015). It can be divided into machine learning text classification and ontology-based text classification. This work focuses on ontology-based text classification because the statistical approach of machine learning text classification has difficulties with the semantic relations between words, making it more challenging to identify the conceptual patterns in text (Luger, 2009). Ontologies can solve that problem by an explicit specification of concepts, descriptions and the semantic relations between them (Irfan et al., 2015).

Clustering describes a process that groups documents together based on a specific category (Jain, 2010). Irfan et al. (2015) divides clustering into hierarchical clustering, partitional clustering, and semantic-based clustering.

Frequency and trend analysis of terms, for instance using Google Trends, is often mentioned as a successful strategy in recent literature (Al-Imam & AbdulMajeed, 2017; Prakash, 2016). Google Trends gives insights into the time dimension of different topics that are uncovered using topic mining approaches. It can help to investigate how long, to what extent, and from which location terms are searched by Google users.

2.3 Text mining for cybersecurity

This subchapter argues that text mining can be a useful tool for cybersecurity.

Big Data is nowadays an essential concept in healthcare (Apurva, Ranakoti, Yadav, Tomer, & Roy, 2017; Haldorai & Ramu, 2018; Shafqat et al., 2018).

Alami & Elbeqqali (2015) proposed a project to apply text mining techniques in the domain of cybercrime. They used microblog posts (Twitter) as a data foundation and extracted cybersecurity relevant information from there. Using this technique, they were able to identify so-called “suspicious profiles” meaning profiles of individuals that are possible threat actors. They also linked different profiles based on metadata and content together. Analyzing the text, they were focusing mainly on hashtags like #Arabspring or #BostonAttack that are compared to a suspicious term database to detect relevant posts. Gupta, Sharma, & Chennamaneni (2016) used a similar approach to identify cybersecurity attitudes and behaviours on Twitter. Grover, Kar, & Davies (2018) and Zubiaga et al. (2018) used Twitter and text mining to identify technology trends related to cybersecurity and healthcare. Erkal, Sezgin, & Gunduz (2015) and Hernandez-Suarez et al. (2018) take this approach even further by using microblog data to create an alerting system for predicting cyberattacks. Thapen, Simmie, & Hankin (2016) proposed a project that combines twitter data with news articles for event detection and situational awareness, showing that linking data sources has a positive influence. Another example of a software product using this approach in a sophisticated way is IBM X-Force Exchange. It is a proprietary, cloud-based threat intelligence platform that aggregates knowledge obtained from many different sources about security threats (IBM Security, 2017).

One growing problem in this domain is fake news. An increasing number of extremists share their ideas and beliefs on social media to influence others or just for targeted disinformation (Ferrara, Wang, Varol, Flammini, & Galstyan, 2016). This untruth news can influence and distort the analysis. Abid et al. (2017) and Ferrara et al. (2016) developed approaches to find this malicious posts using machine learning and rule-based approaches.

Apart from media on the internet, also the darknet can hold valuable information for the proposed research (Al-Rowaily, Abulaish, Al-Hasan Haldar, & Al-Rubaian, 2015).

2.4 Conclusion

Even if literature states a fast-growing problem in cybersecurity in healthcare, the proposed text mining approaches mostly tackle only general cyberthreats. For

this broad domain, that approach seems to be successful. Projects that analyze specifically the cyberthreats emerging by digital healthcare could not be found. Since literature supports that cybercrime in the healthcare industry is different from other industries, a more industry-focused analysis can be valuable.

When looking at the data sources, scientific literature has a strong focus on mining Twitter data. One reason for that is easy availability. Nevertheless, the value of the data is questionable because the text is very short and contains not so much information. Additionally, not all information on Twitter is carefully researched and accurate. More data from different sources can lead to an increase of valuable results. Research focusing on a broader data foundation from different sources, as suggested in this project, could not be found. The IBM X-Force Exchange offers some functionality as proposed in this work. However, it is not open useable and not tailored to the healthcare industry. Additionally, it focuses not so much on the threat that technology itself can present but more on specific vulnerabilities in software and hardware.

These facts clearly show that there is a research gap in this area, making the project proposed in this work even more important.

From a methodological point of view, the literature supports topic modelling, ontology-based text mining, and dictionary-based classification. Useful ontologies could be MESH[2] or SNOMED CT[3]. More recent literature also indicates that Google Trends can provide additional value for the proposed research questions, by giving insights into frequency, geolocation, and trends of Google searches. This feature could be beneficial on-top of previously mined topics as information enrichment.

[2] bioportal.bioontology.org/ontologies/MESH/

[3] www.snomed.org/

3 Data foundation

This chapter deals with the data sources that are investigated in the next parts of this research. Only documents from 01.01.2015 until 31.03.2018 in German and English language are investigated. Additionally to the GDELT data, about 300,000 articles were gathered for analysis. Except for the GDELT dataset (described at 3.4) which is publicly available, all data was provided by the DocCenter from the National Defence Academy (DocCenter/NDA) of the Austrian Armed Forces. This chapter only describes the data sources, the process of how the data was extracted from the sources of the DocCenter and imported into the text mining systems can be found at subchapter 4.1 and 4.2.

3.1 Cyber Documentation & Research Center

The Cyber Documentation & Research Center, short CDRC, was a project at the DocCenter from the National Defence Academy (DocCenter/NDA) of the Austrian Armed Forces. It uses a Crowd Open Source Information (Crowd OSInfo) approach to gather information on cybersecurity matters from previously selected and evaluated high-quality news resources. The Cyber Documentation & Research Center collects information about three different topics, “cyber”, “crises, military, and security policy”, and “innovation and technology”. From 01.01.2015 until 31.03.2019 about 160,000 documents in German and English language are indexed in different Ushahidi databases.

The Ushahidi platform consists of a MySQL based database system that has been designed especially for the purpose of storing data about news, incidents or disasters (Meier, 2012). It offers a transparent, near real-time insight into what is really happening. Ushahidi offers interfaces allowing people who are at the location of an indecent to upload reports and updates about the event using different channels like web forms, social media, email or text messages. Because of this support of the Crowd OSInfo approach, it is an ideal foundation for storing data in the CDRC.

The reported data is accessible using different web UIs which are pictured in the following subsections. Each platform allows keyword search and filtering based on metadata. More information on the Cyber Documentation & Research Center can be found at Mak, Klerx, Pilles, & Göllner (2015).

3.1.1 Cyber[4]

This repository of the CDRC holds news articles about the cyber domain. From 01.01.2015 until 31.03.2019 about 130,000 documents in German and English language are indexed. Additionally to the news text, every report is manually tagged with information about the category and news type. In addition, different metadata like language, date, author or location information are saved. Cyber soldiers update the cyber Ushahidi every working day.

The web UI of the Cyber Ushahidi platform is pictured in Figure 1.

Figure 1 Cyber UI

3.1.2 KriMiSi[5]

Crises, military, and security policy („Krisen-, Militär- & Sicherheitspolitik“, short KriMiSi) focuses on reports about crises and other military relevant situations around the world. About 16,000 German and English geolocated entries are included in this repository. Cyber soldiers update the KriMiSi Ushahidi every second week.

Figure 2 shows the KriMiSi web interface.

[4] cyber.cdrc.at

[5] Krimisi.cdrc.at

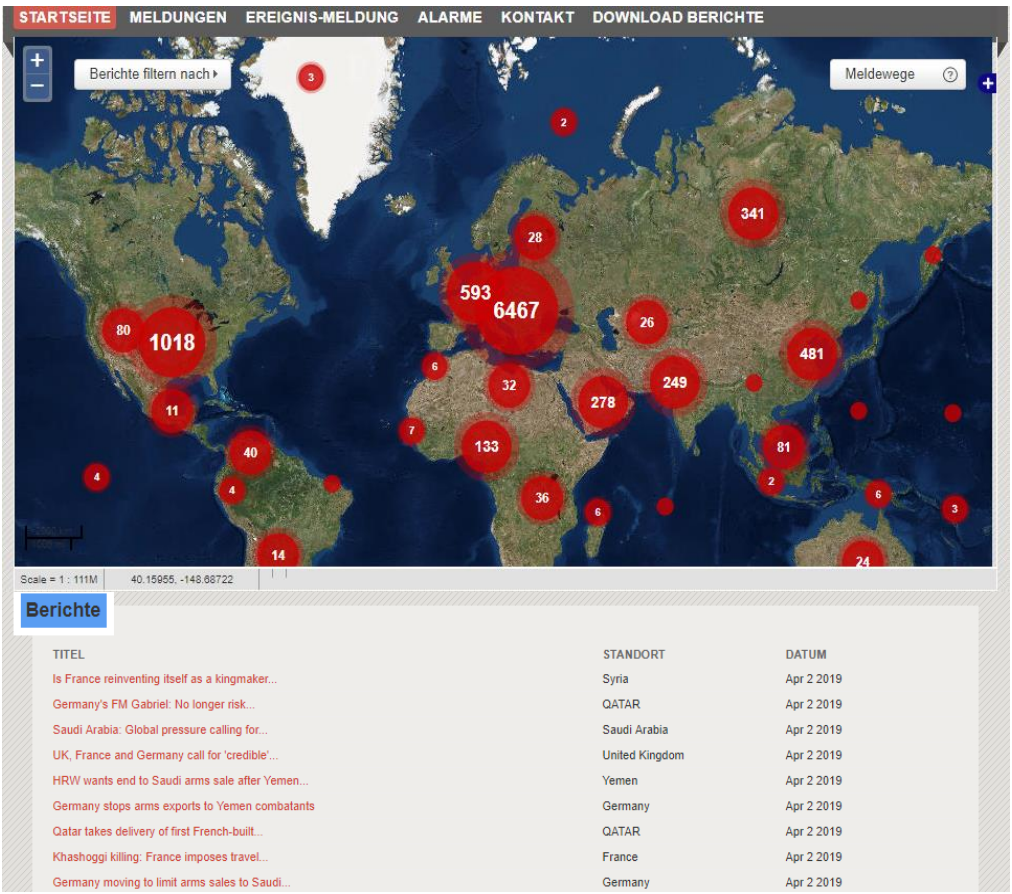


Figure 2 KriMiSi UI

3.1.3 InnoTech[6]

InnoTech focuses on articles, patents and other publications on innovation and technology. The InnoTech repository holds about 6,000 records in German and English language. Each article is manually tagged with geo information about the origin of the entry and the kind of innovation or technology it deals with. Cyber soldiers update the InnoTech Ushahidi biweekly.

The InnoTech website is pictured in Figure 3.

[6] innotech.cdrc.at

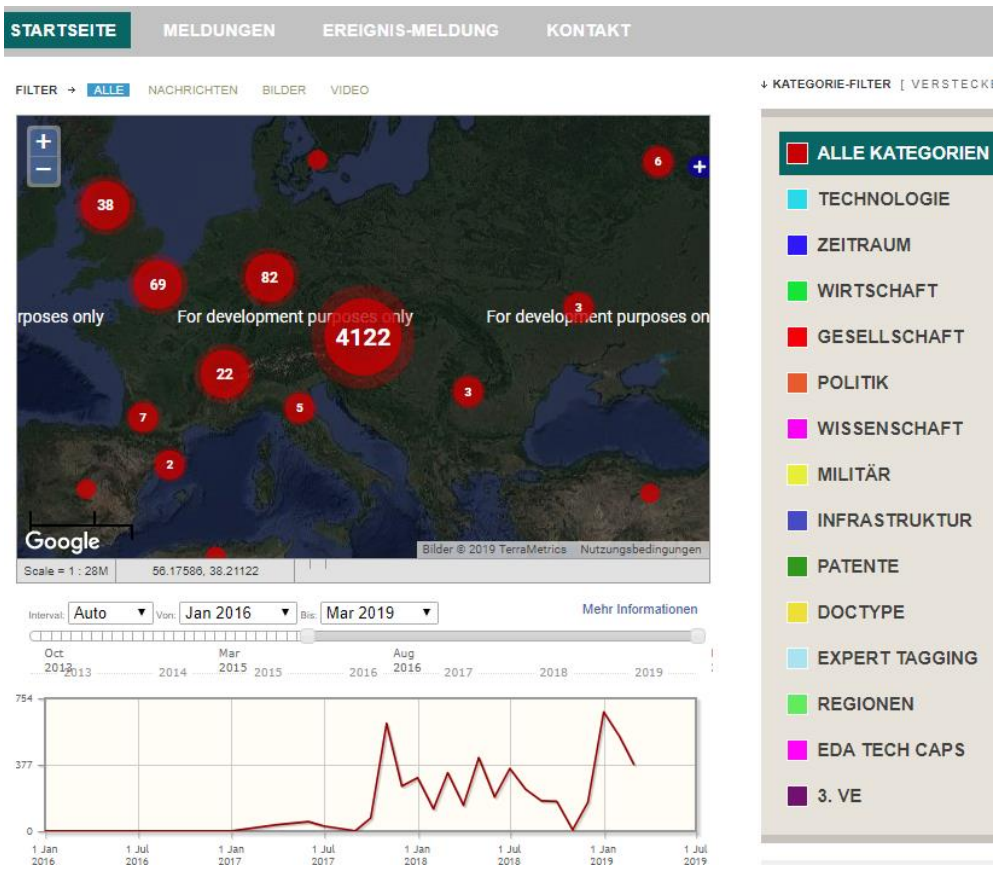


Figure 3 InnoTech UI

3.2 Austrian News Data

This dataset contains about 130,000 press releases from Austria in German language with a focus on political issues. Additionally to the news text, the dataset holds the date, heading, subheading, and an automated tag about the news kind.

3.3 Lexis Nexis Database

This dataset consists of about 12,000 articles divided into eleven categories from many different news sources in German and English language. For the analysis, the only categories used were science, technology, and medicine (about 1,600 articles). Every article is tagged automatically with different keywords that describe its content and a percentage score of how relevant this keyword is to the article.

3.4 Global Database of Events, Language and Tone

The Global Database of Events, Language and Tone (GDELT) project[7] collects data about the world's broadcast, print, and web news since 1979 until now about nearly every country in the world in more than 100 languages. All articles are translated into English automatically ('The GDELT Project', n.d.). Currently, it holds about 850 million geolocated and enriched records. It is described as

> "An initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day."
>
> - Kalev Leetaru & Philip Schrodt (Creators of GDELT)

The data is open source accessible and can be downloaded as raw files or directly analyzed using different web services or Googles BigQuery API. The data set is CAMEO coded. CAMEO stands for Conflict and Meditation Event Observations and is a coding standard for political news and violence (Gerner, Schrodt, Abu-Jabr, & Yilmaz, 2002). For this work, the GDELT 2.0 Knowledge Graph dataset was used. The complete GDELT Data has about 9.5 terabytes.

More information on the GDELT Project can be found at 'The GDELT Project' (n.d.) and at Leetaru & Schrodt (2013).

[7] www.gdeltproject.org

4 Fact extraction and analysis

This chapter explains the used methodology in detail. First, the general system architecture is described. Subsection 4.2 deals with the data gathering process, afterwards the text mining tool IBM Watson Explorer and Google Trends are explained, and the fact extraction process is shown. The last subsection contains an overview of the investigation methodology meaning which methodology the parsed data can be analyzed.

4.1 Data flow

To deepen the understanding of the workflow and to summarize the processes, the following subchapter explains the data flow of the Cyber Documentation and Research Center and the Global Database of Events, Language and Tone.

4.1.1 CDRC

According to the CDRC processes developed by Mak et al. (2015), previously selected high-quality news sources (high profile sites) are read by subject matter experts, also called high-performance crowd, and relevant content is saved in different content management and database systems. As part of this research, these systems were loaded into a text mining system using ETL processes. After the text mining process, the investigator can analyze the data to gain insights into the data. Figure 4 pictures this process.

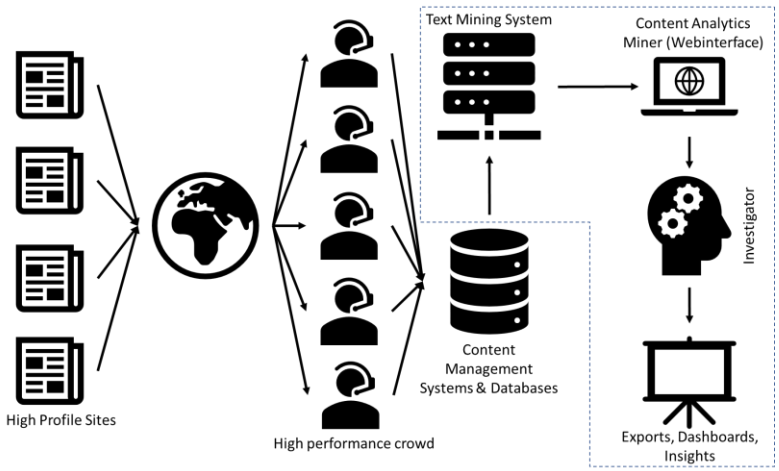


Figure 4 Data Flow CDRC

4.1.2 GDELT

The GDELT data was analyzed using the Google BigQuery API. To make the full text effectively searchable, the raw files have been downloaded and imported into a search index for text queries. The GDELT Analysis Service[8] also offers possibilities to investigate the dataset. Figure 5 shows the analyzation process of the GDELT data. The blue dotted area marks systems that have been used or developed directly for analyzation as part of this research.

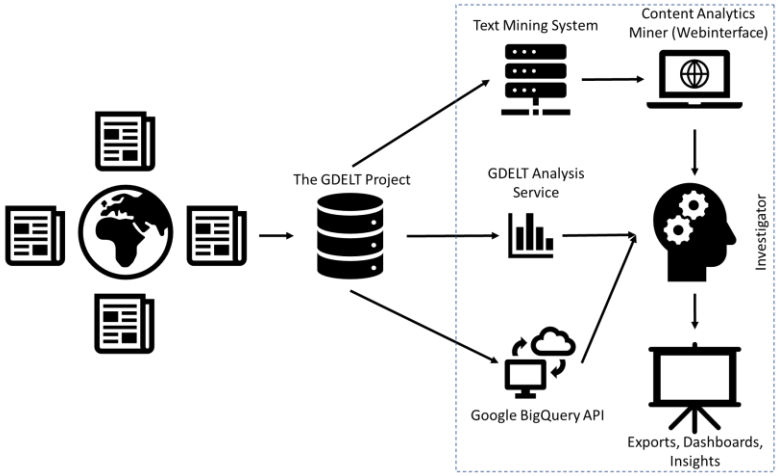


Figure 5 Data Flow GDELT

4.2 Data pre-processing

In content analytics, pre-processing is a time consuming task which has a high impact on the analysis results (Crone, Lessmann, & Stahlbock, 2006; Uysal & Gunal, 2014). To make analysis possible, all data was pre-processed and converted to a format that was usable for being loaded into the text mining system IBM Watson Explorer.

For the Cyber CDRC data, a database crawler has been developed to craw the data directly from the Ushahidi databases of the Cyber Documentation & Research Center. Since the out of the box database connection of the Watson Explorer does not support the structure of the Ushahidi databases used in the CDRC, this functionality had to be programmed additionally. To make crawling easier, a flat representation of the data was needed. For that, the relations of the database structure shown in Figure 6 have been joined together in a flat structure using a single database view. In this view, each row represents an article. Especially the join of the 1:n and m:n relationships into a flat representation in

[8] http://analysis.gdeltproject.org/

one row required a complex SQL logic. The then created crawler connects over a JDBC type four driver to the Ushahidi databases view to read each record and then load it into the Watson Explorer index.

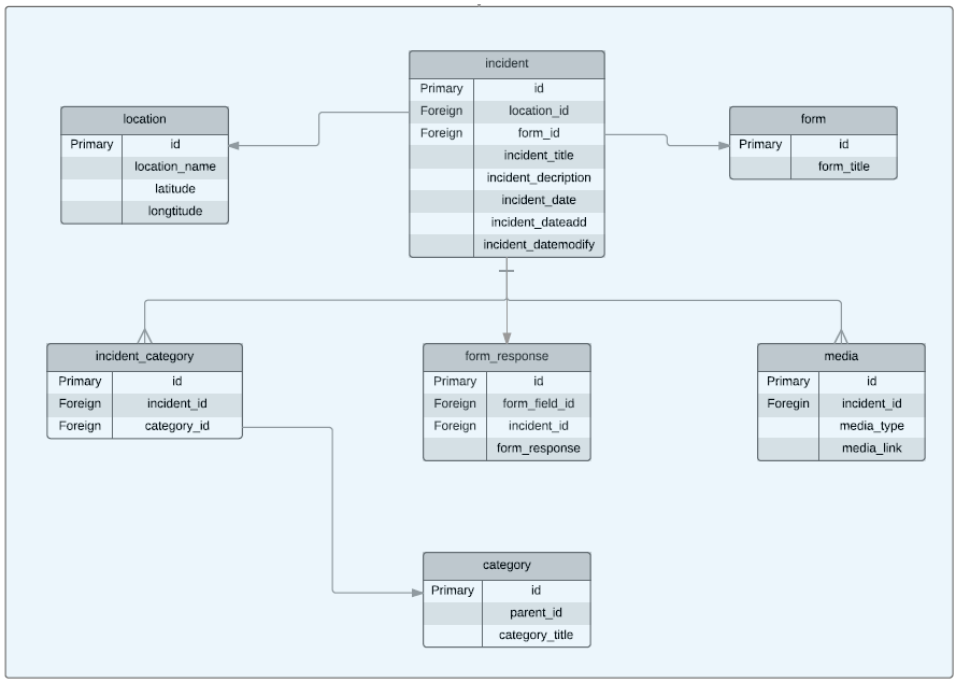


Figure 6 Ushahidi database structure

To crawl the Lexis Nexis datasets, the provided HTML files were first converted to PDF for better analysis results using a programmed script. The PDF files could then be crawled using a Windows File System Crawler.

The Austrian News Data was converted from txt files to XML to make the import of the metadata (like creation date, topic or tags) of each article possible. InnoTech, and KriMiSi data was exported from servers of the DocCenter from the National Defence Academy (DocCenter/NDA) of the Austrian Armed Forces using a specially programmed XML exporter and then re-imported in Watson Explorer using XML mapping strategies.

Since GDELT holds such a massive amount of data (about 9.5 TB of text), analysis, especially the download and storage, is difficult. The GDELT data is available through three different sources. The raw data can be downloaded as csv files. The GDELT database is also available through Google's BigQuery, a system to query large SQL databases on Google infrastructure over the Google cloud. Additionally, the GDELT project offers simple pre-configured analyze services on its website. The Googles BigQuery system offers SQL functionality but not advanced data and text analytics like the Watson Explorer does. Through the GDELT analysis services, not all filter options needed to get the relevant data could be performed. That is why the GDELT analysis service was only used

additionally. The primary analysis was done using BigQuery. There the data was filtered so that smaller sub-datasets could be loaded into the Watson Explorer for text mining. Since GDELT was mostly used to crosscheck specific analysis results that have been gained from the other datasets this previous filtering process even helped to narrow down the analysis scope.

4.3 IBM Watson Explorer

IBM Watson Explorer Deep Analytics Edition 12.0.2.2 (WEX DAE) was used to mine the data described in the chapter before. This software bundle holds tools for enterprise search, natural language processing, and machine learning.

The program part used for the text mining in this research is called IBM Watson Explorer Analytical Components. It consists of three main interfaces.

- Administration Console

In the Administration Console, all server configuration can be managed. Watson Explorer Analytical Components stores different data sets in so-called collections. Each collection consists of three parts as shown in Figure 7.

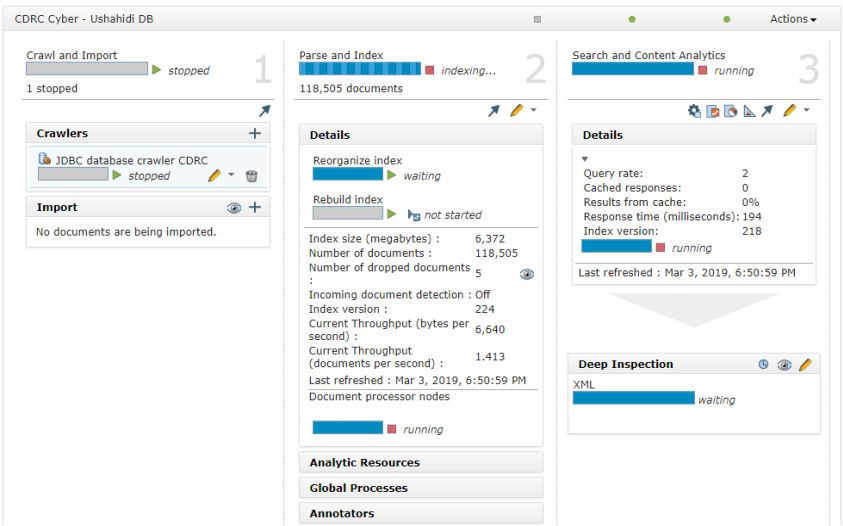


Figure 7 Watson Explorer Admin Console

The first part, “Crawl and Import”, defines how data is accessed for importing into collections.

The second part, “Parse and Index” performs the natural language processing and saves the text, metadata and extracted facts into the index. This part is described in more detail in subsection 4.4.

The third part, “Search and Content Analytics” holds functionality to configure the analysis in the Content Analytics Miner.

- Content Analytics Studio

The Content Analytics Studio is an application for Microsoft Windows based on the Eclipse IDE. It is used for developing Apache UIMA Annotators for fact extraction based on RegEx, dictionaries, rules or ontologies. UIMA stands for Unstructured Information Management Architecture and is an open source project for text analytics developed by the Apache Software Foundation. Figure 8 shows the user interface of the Content Analytics Studio.

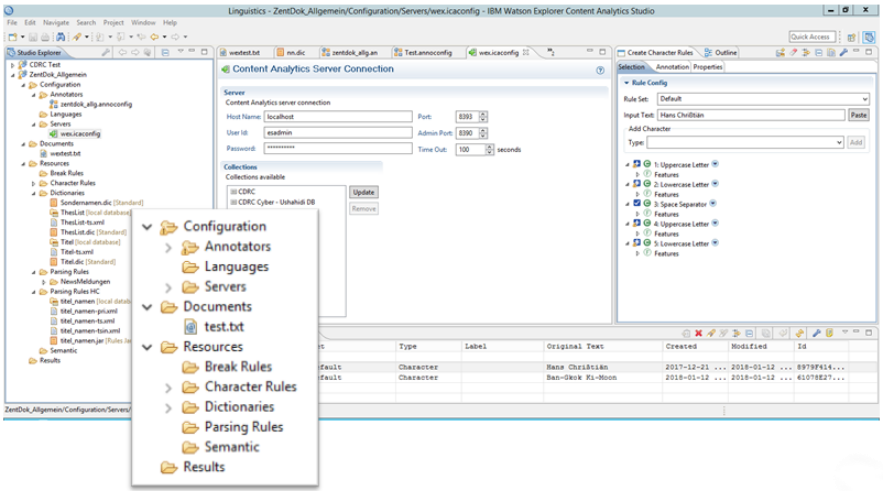


Figure 8 Content Analytics Studio UI

- Content Analytics Miner

The Content Analytics Miner, as shown in Figure 9, is a web user interface for exploring collections for the end-user. It offers different methods of search, performing drill downs, investigating the extracted facts and calculating correlations. In this part of the software, the whole investigation process takes place. Further descriptions can be found in chapter 4.6.

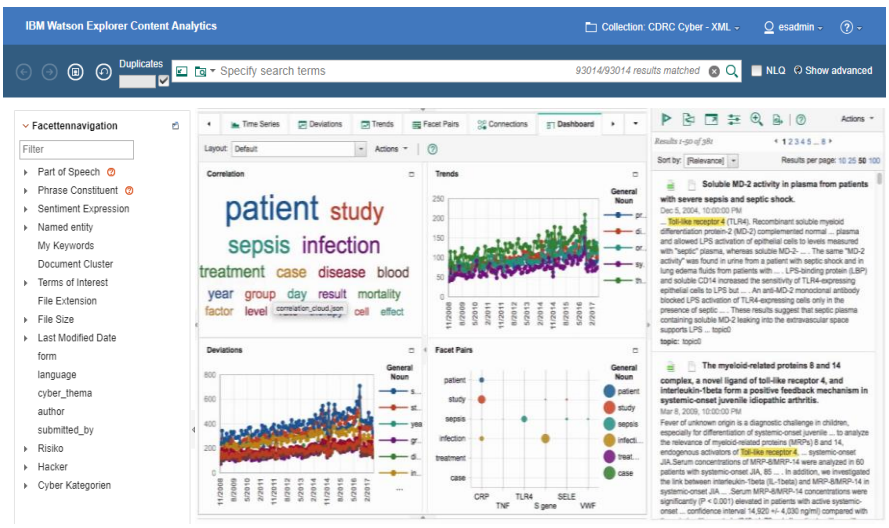


Figure 9 Content Analytics UI

- REST API and near real-time NLP API

The whole system is configurable and query able using a REST API. Additionally to the functionality of the content analytics miner, Watson Explorer Analytical Components offers also a real-time NLP API to provide natural language processing as a service.

A detailed explanation of the functionality and the possibilities of the Watson Explorer can be found at Mak, Pilles, Bertl, & Klerx (2018) and at Zhu et al. (2014).

4.4 Entities, rules and extraction process

According to the research questions stated in chapter 1.3, annotations for the natural language processing have been designed. For that, first the topics of interests were compiled manually. Examples are:

- Health terminology
- Cyber terminology
- Actors (hacker, hacker groups, hackerspaces, healthcare institutions, politicians, doctors, professors, researchers, etc.)
- Targets (medical devices, IT systems, wearables, applications, apps, websites, data breach)

After that, terminology for building the annotations for the topics of interest was extracted using manual research (reading through the datasets, internet searches, etc.), available terminology lists (like the semantic fields of the CDRC categories) and available ontologies (e.g. MASH or SNOMED CT). With the extracted terminology, rules for parsing out the information from the news articles were developed in the Content Analytics Studio. Rules are language dependent, meaning that for each language-specific rules had to be created. The developed rules can then be deployed to the UIMA pipeline in the Watson Explorer for the extraction process. The UIMA pipeline contains different steps for the text analytics process. The different stages every text goes through in the analysis process are pictured in Figure 10.

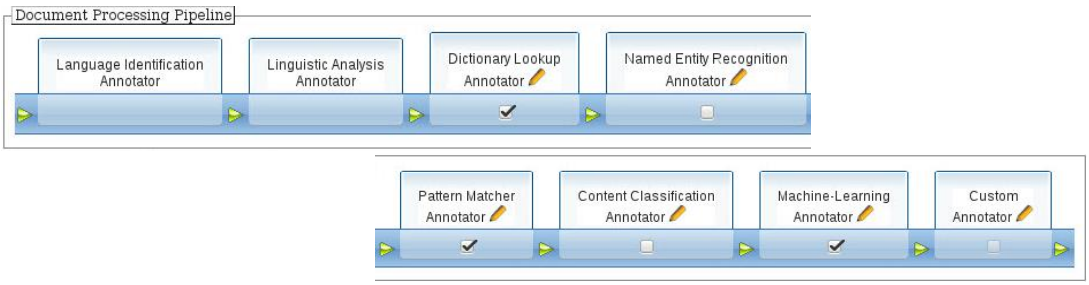


Figure 10 UIMA Pipeline

After the crawling of the news articles, the first step in the UIMA pipeline, language identification is done. Dependent on the detected language, the linguistic analysis annotator performs tokenization, meaning text is split up into paragraphs, sentences, and words. These steps are the necessary minimum; everything else is optional and depending on the analysis. The part of speech analysis in the Pattern Matcher phase categorizes words using grammar to nouns, verbs, adjectives, stop words and so on. Additionally, different other steps for natural language processing can be configured in the UIMA pipeline as well. Entities can be parsed out using dictionaries, automated named entity detection, patterns, word combinations, rules or machine learning. The content can be classified using artificial intelligence or statistical algorithms. Custom text analysis can be implemented using Content Analytics Studio or self-programmed Java annotators. In the indexing stage, the parsed-out entities and the text together with the metadata available is saved in index fields of an Apache Lucene like search index. These fields can then be mapped to facets for analyzation in the Watson Explorer Miner. Facets augment traditional search with more dimensions to narrow down search results with different filters on the extracted facts (Persons, Organizations, etc.) and metadata (date, source, etc.).

4.5 Google Trends

Google Trends gives insights into the search behaviour of Google user. It shows how often a search term was googled over a specified time period. Google Trends also gives insights into geo locations and related topics or search terms. The number of searches indicates user trends. The trends of different keywords can also be compared.

As one example, Figure 11 shows a Google Trend analysis of the term “Digital Healthcare”. It can be seen how “Digital Healthcare” increased its popularity steadily over the last years, especially in North America and India. The section “related queries” states the search queries that other users have additionally searched related to “Digital Healthcare”. The “related topics” section show on a higher-level the summarized topics that were searched together with “Digital Healthcare”. The found terms “digital transformation”, “healthcare administration” or “investment” can lead to the logical assumption that “Digital Healthcare” is part of the digital transformation, one major part is healthcare administration and it requires new investments.

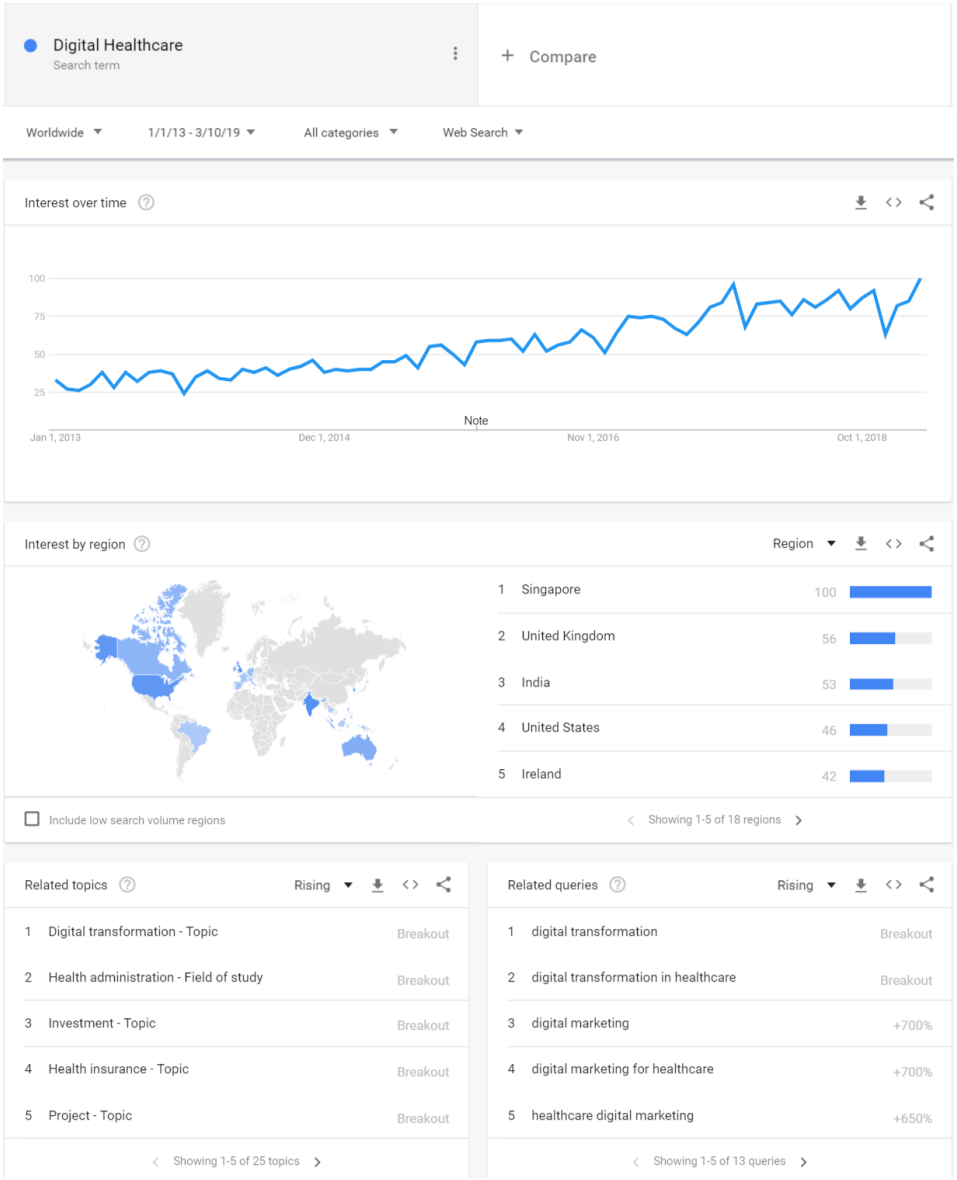


Figure 11 Google Trends, Data source: Google Trends (https://www.google.com/trends)

4.6 Investigation Methodology

After the steps done in section 4.4 (entities, rules and extraction process), the annotations can then be used together with the metadata in the facets of the Content Analytics Miner to analyze the data. The following section gives an overview of the possibilities of the Watson Explorer Analytical Components which has been used for text mining in this research. Chapter 5 shows the application of the here described methodology on the topic of this work. The in this subchapter proposed methodology on investigation big amounts of unknown text data has been proven successful in previous projects (Fuchslueger, 2016; Mak et al., 2018) and has been therefore used in this research.

Analyzing the different facets can bring quick insights into what the data is about and what the main topics are. The two values that the Watson Explorer calculates in the facet view pictured in Figure 13 are the frequency (how many documents contain the facet value) and the correlation (how strongly a facet value is related the current search query or another facet value). The correlation value indicates how relevant the facet value is to the documents matching the currently active search condition. In this context, the correlation is not the common mathematically known value but a measurement that is used to gauge the relevance of a particular keyword as it compares to other data in a document corpus (Zhu et al., 2014, p. 16 f.). In other words, correlation measures the level of uniqueness of the facet value as compared to other documents that match a query. A correlation bigger than 1.0 means an anomaly in the data that should be investigated. High correlation does not necessarily mean high frequency or the other way around. In the example demonstrated in Figure 12, "IoT" has a higher frequency than "AI" in the documents about digital healthcare but the correlation value of "AI" with "Digital Healthcare" would still be bigger because "AI" is mentioned in the digital healthcare context more often than in the rest of the document set. From this perspective, the correlation can also be interpreted as the level of uniqueness of a facet value in the context of the current search compared to the rest of the dataset.

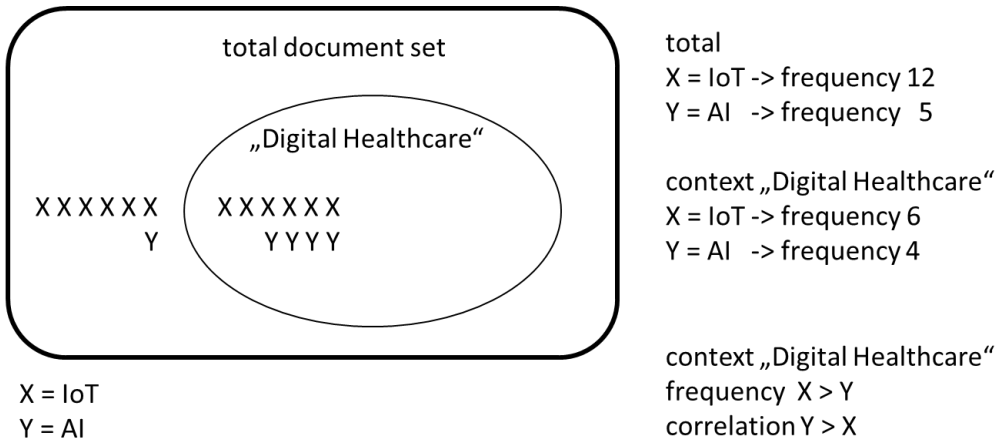


Figure 12 Frequency vs. Correlation example

To show that this approach produces logical and reliable results, the GDELT dataset in Figure 13 is filtered to show only data where the US democratic party is mentioned. Using the v1person facet of the GDELT dataset and correlation calculation, persons with a connection to the democratic party can be filtered out. The names with the highest correlation value ("Kristen Gillibrand", "Jay Inslee" and "Eric Giddens") are indeed members of the democratic party in the US. This example indicates the reliability of this method and shows how correlations can bring insights into data without any prior added ontology or knowledge model behind the analysis process. This approach has been used in chapter 5.

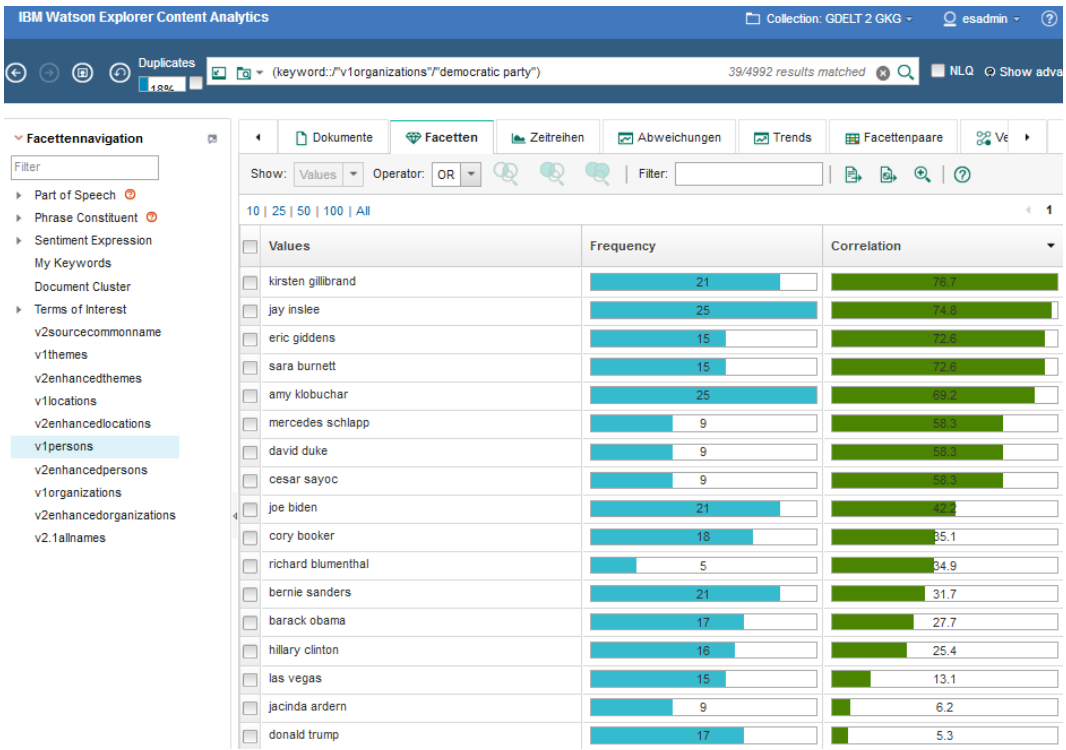


Figure 13 Facet view

To understand the time component of the data, the time series view is used. A bar chart represents the frequency of documents distributed over time. Using search or faceted search, the absolute frequency of documents containing different terms can be investigated. As one example, Figure 14 shows the rise of articles about digital healthcare per month in the Cyber dataset.

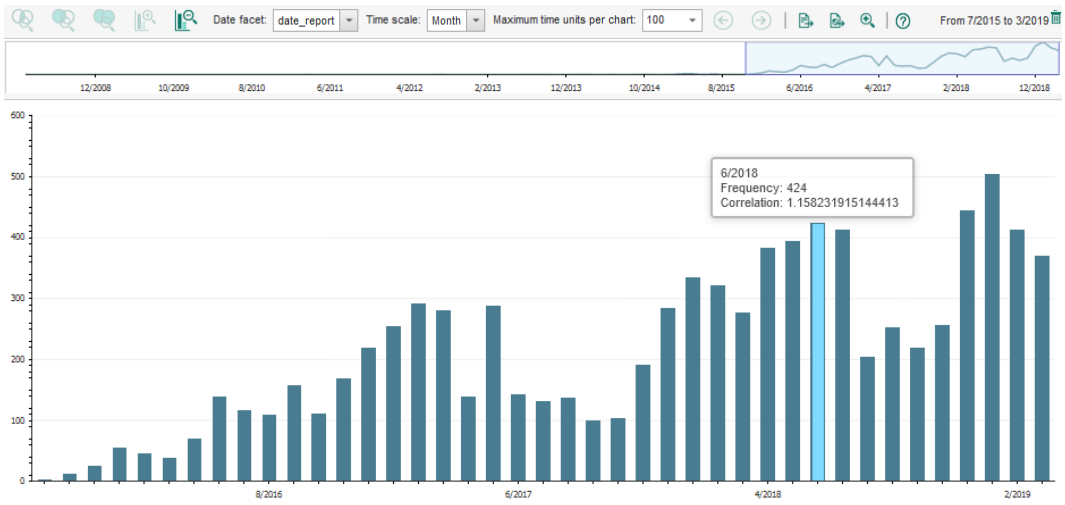


Figure 14 Time Series view

The trends view identifies anomalies (sharp or unexpected increases in the frequency) of a facet over the last time periods. The x-axis for each facet value behaves similarly to the time series chart, it shows the time according to the

currently set time scale. The left side of the y-axis is the frequency count of the facet value. The right side of the y-axis represents the trend index. This value measures the increase ratio of the frequency for a given time interval as compared to the expected average frequency that is calculated based on changes in the past of the frequency over a given time interval. A modified Poisson distribution is used to estimate the expected change in frequency (Zhu et al., 2014, p. 180). The trend index is shown as a line graph in the chart. The colour of the bar indicates additionally if the trend index is higher than expected.

The charts can be sorted by the highest frequency of the facet values, highest trend index (to see which values are operating out of the norm the most), latest index (similar to the highest index sort but looks at the highest index values for the most recent time to see which values are most recently operating out of the norm) and name. An explanation of how the index values are calculated can be found at Zhu et al. (2014, p. 181).

Figure 15 shows the trend graphs for different hacker related to digital healthcare that were found in the Cyber dataset. It gives insights into when the hackers were mentioned more often than usual.

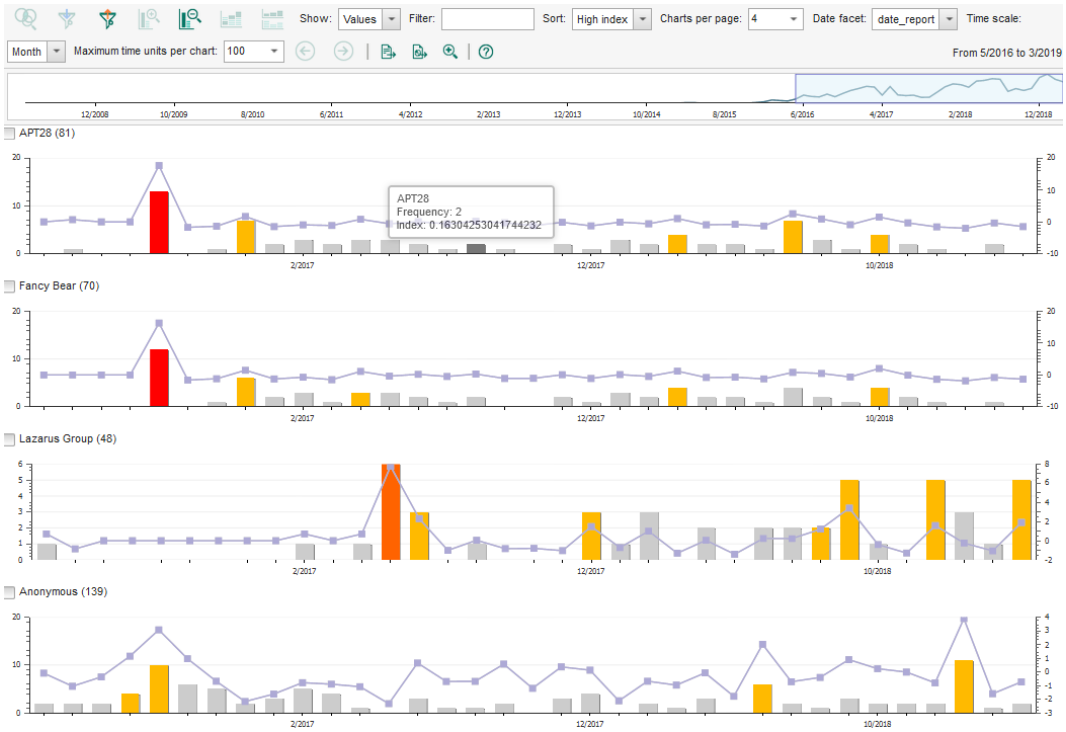


Figure 15 Trend view

Trends from different facet values can be compared by combining several chards into one. This way, it can be investigated how different trends relate to each other. Figure 16 shows the combined trend view of Anonymous and the Lazarus Group in the context of digital healthcare.

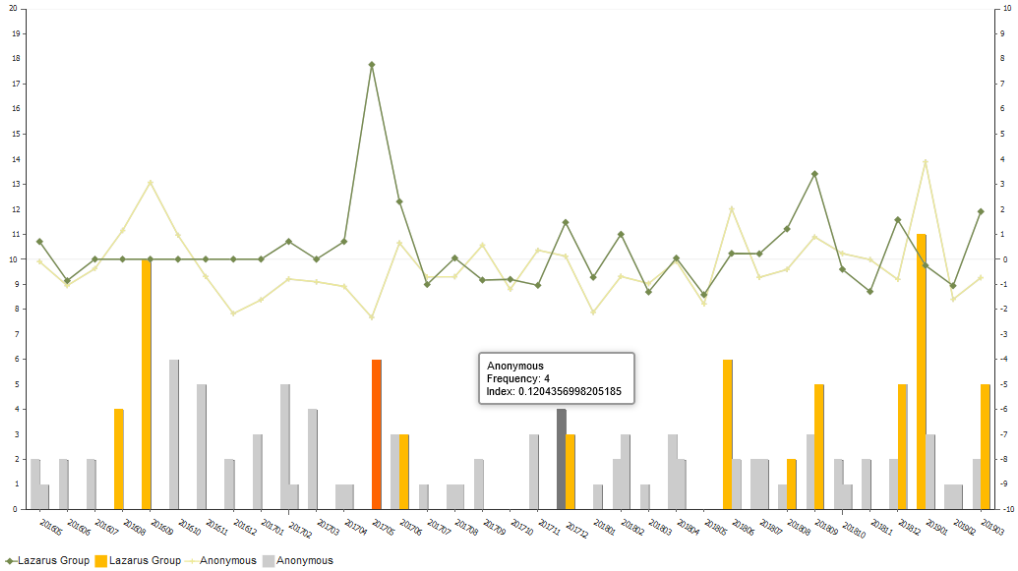


Figure 16 Combined Trend view

The list of bar charts in the deviation view shows how much the frequency of a facet value deviates from a calculated expected average over the entire time period. Figure 17 shows the deviation graphs for different hacker related to digital healthcare that were found in the Cyber dataset. The deviation index score measures the average occurrence of a facet value on a selected date across the currently searched documents. Deviations can be increases or decreases from the average value. It is useful to identify patterns that occur cyclically (seasonally, monthly, weekly, etc.). The deviation view can alert if cyclic patterns have unexpected changes. In contrast to the trend view which focuses only on past periods, the deviation view takes into account the whole time period.

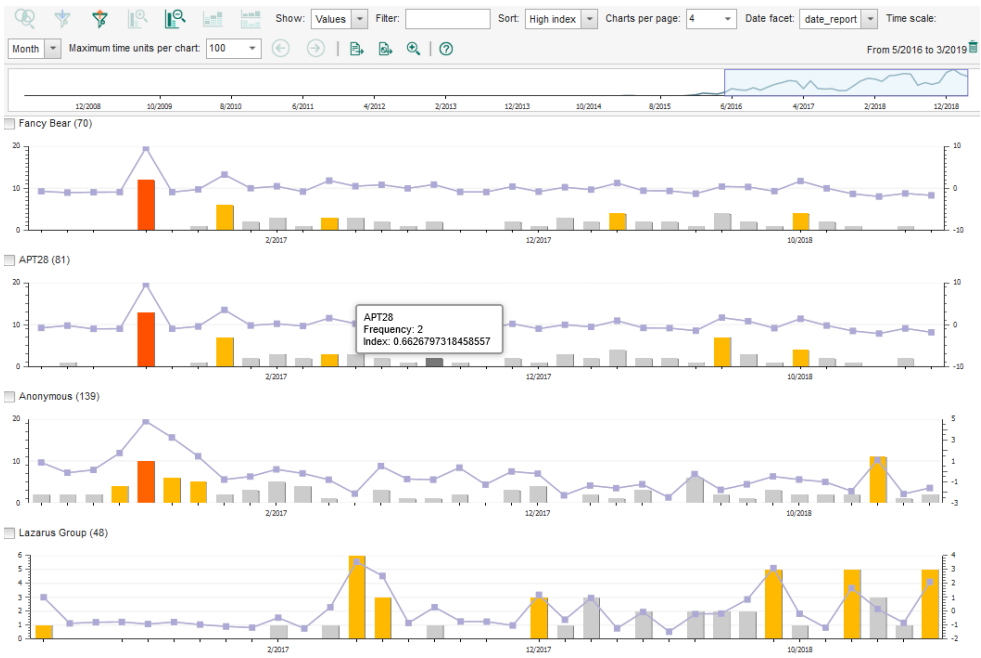


Figure 17 Deviation view

To investigate possible connections between facets, the facet pair analysis is used (see Figure 18 and Figure 19). It shows, how often two facet entries are mentioned in a document together and calculates a correlation between these two values in context to the current search. The calculation of the correlation is shown using the following example modified from Zhu et al. (2014, p. 192).

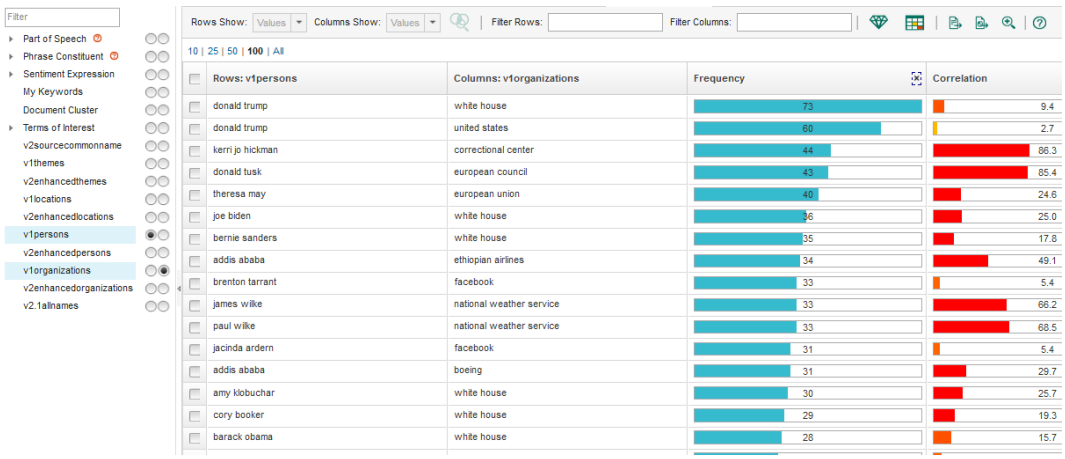
- Total number of documents 852
- Total number of occurrences of the facet value “cybercrime” 123
- Total number of occurrences of the facet value “DDoS” 86
- Total number of documents containing “cybercrime” and “DDoS” 67

According to the example, about 14% of the documents (123/852) contain the keyword “cybercrime”. Zhu et al. (2014, p. 192) refers to this as the density of “cybercrime” in the total document corpus. The density of “cybercrime” in the document set containing the term “DDoS” is 78% (67/86). The correlation value is now calculated as the ratio of these two density values making it about 5 (78/14). In other words, the correlation value is the ratio of the density of the facet value “cybercrime” in the document set for keyword “DDoS” and the density of the facet value “cybercrime in the whole text corpus. To put it in general terms the following formulas can be derived:

$$density = \frac{frequency}{frequency\ of\ total\ corpus}$$

$$correlation = \frac{density\ intersection}{product\ of\ densities\ of\ intersected\ sets} * reliability\ correction$$

Because that correlation value is not so reliable when the number of documents which include both keywords is relatively small, a reliability correction using statistical interval estimation is used additionally. That makes the correlation more accurate. For easier understanding, again an off-topic example is used. Figure 18 shows the correlation table between the facets v1persons and v1organistions of the GDELT dataset. Figure 19 pictures this in the grid view.



Rows: v1persons	Columns: v1organizations	Frequency	Correlation
donald trump	white house	73	9.4
donald trump	united states	60	2.7
kerri jo hickman	correctional center	44	86.3
donald tusk	european council	43	85.4
theresa may	european union	40	24.6
joe biden	white house	36	25.0
bernie sanders	white house	35	17.8
addis ababa	ethiopian airlines	34	49.1
brenton tarrant	facebook	33	5.4
james wilke	national weather service	33	66.2
paul wilke	national weather service	33	68.5
jacinda ardern	facebook	31	5.4
addis ababa	boeing	31	29.7
amy klobuchar	white house	30	25.7
cory booker	white house	29	19.3
barack obama	white house	28	15.7

Figure 18 Facet Pair table view

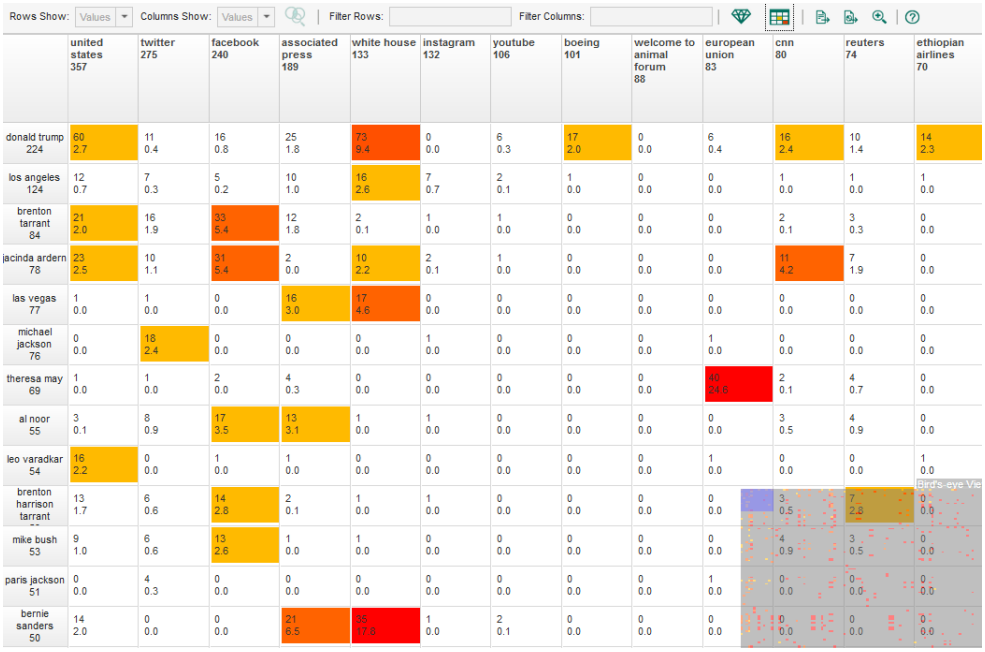


Figure 19 Facet Pair grid view

Both Figure 18 and Figure 19, show a strong correlation between Donald Trump and the White House or between Theresa May and the European Union. Given the fact that Trump is currently the president of the United States and Theresa May, as president of the United Kingdom, is negotiating the Brexit Deal with the European Union these values seem quite logical and is an indication that the proposed methodology brings correct and explainable results.

Rows: Attack Methods	Columns: Cyber Thema	Frequency 2	Correlation 1
carding	Deepweb	13	5.7
side-channel attack	Quantentechnologie	2	4.4
buffer overflow	Sicherheitslücke	27	3.5
false flag	CyberWar	6	3.0
buffer overflow	Angriffsmethoden	19	2.7
botnet	IoT	71	2.7
buffer overflow	Verteidigungsmethoden	13	2.7
botnet	Angriffsmethoden	113	2.6
malware	Angriffsmethoden	554	2.5
SQL-Injection	Angriffsmethoden	24	2.4
cross-site scripting	Sicherheitslücke	22	2.4
targeted attack	CyberWar	9	2.4
spyware	ND	34	2.3
spyware	Spionage	34	2.3
cross-site request forgery	Sicherheitslücke	12	2.3
exploit	Sicherheitslücke	532	2.3
denial of service	Angriffsmethoden	153	2.2
cross-site request forgery	Angriffsmethoden	10	2.2
computer worm	Angriffsmethoden	9	2.2
computer worm	CyberWar	4	2.2
social engineering	Angriffsmethoden	65	2.1
malware	Cyber Crime	1033	2.1
phishing	Angriffsmethoden	225	2.1

Figure 20 Facet Pair table view, example two

Figure 20 illustrates in another example of how the correlation of facet pairs can be used to get insights. The facet attack method and cyber topic of the CDRC Cyber dataset were correlated, showing that different attack methods all correlate to topics specific to cyberattacks.

Additionally to the shown diagrams, all documents can be queried using keyword search, Boolean logic, and linguistics. Search hits are highlighted in an automatic generated intelligent document summary to make effective cross reading possible.

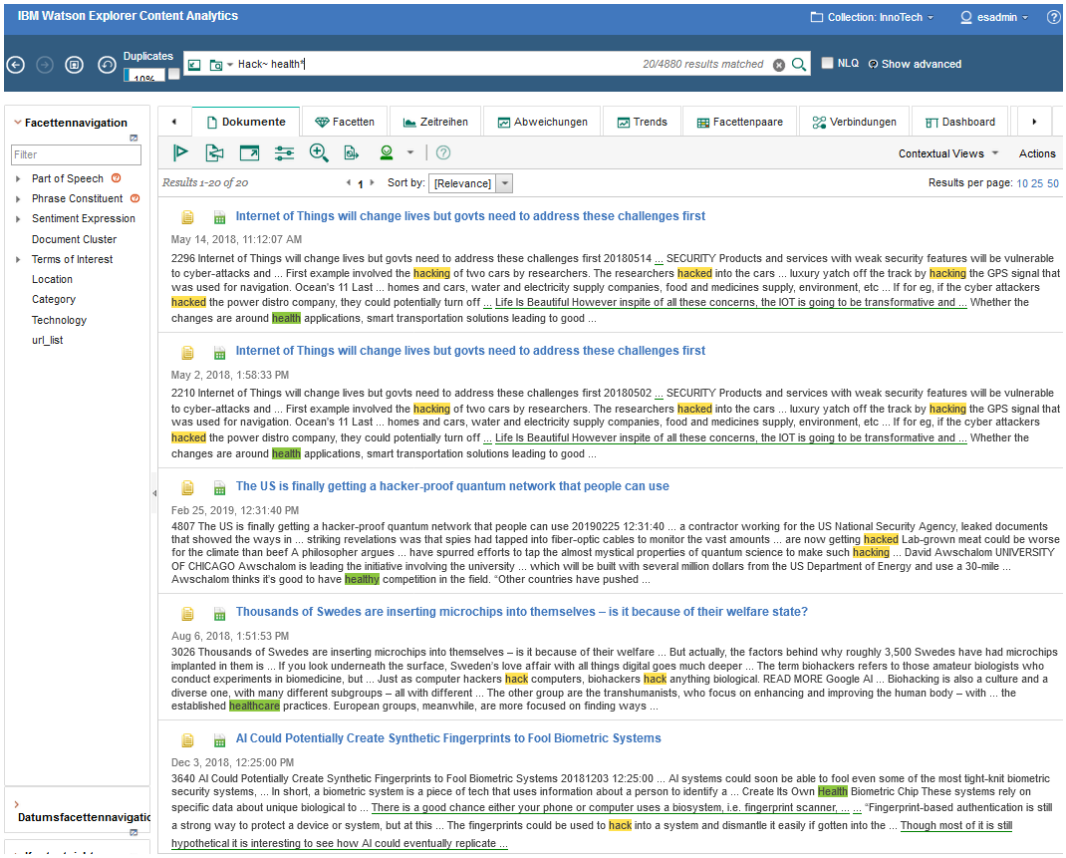


Figure 21 Document view with a search query

The search query shown in Figure 21 (`Hack~ health*`) returns documents containing both the words with the lemma Hack (like hacking, hacked, hack, etc.) and the words that begin with health (`health*`) like health, healthy or healthcare. The search terms are highlighted using different colours. Searches can be saved to be available for later.

The green and red underlined, words, phrases and sentences indicate positive and negative terms that can be investigated using sentiment analysis. Figure 22 shows a more detailed sentiment analysis using the sentiment view. It was investigated how companies were mentioned in the context of digital healthcare. The following view of the Cyber dataset shows the positive (green) and negative

(red) expressions for the selected firm, in this case IBM. Additionally, the phrase that indicates the sentiment is underlined in red or green in the text summary of the document preview. It is possible to analyze positive and negative phrases, expressions as well as the targets of them. The sentiment analysis can be combined with the trend analysis to get insights into how the sentiment of different facets changed over time.

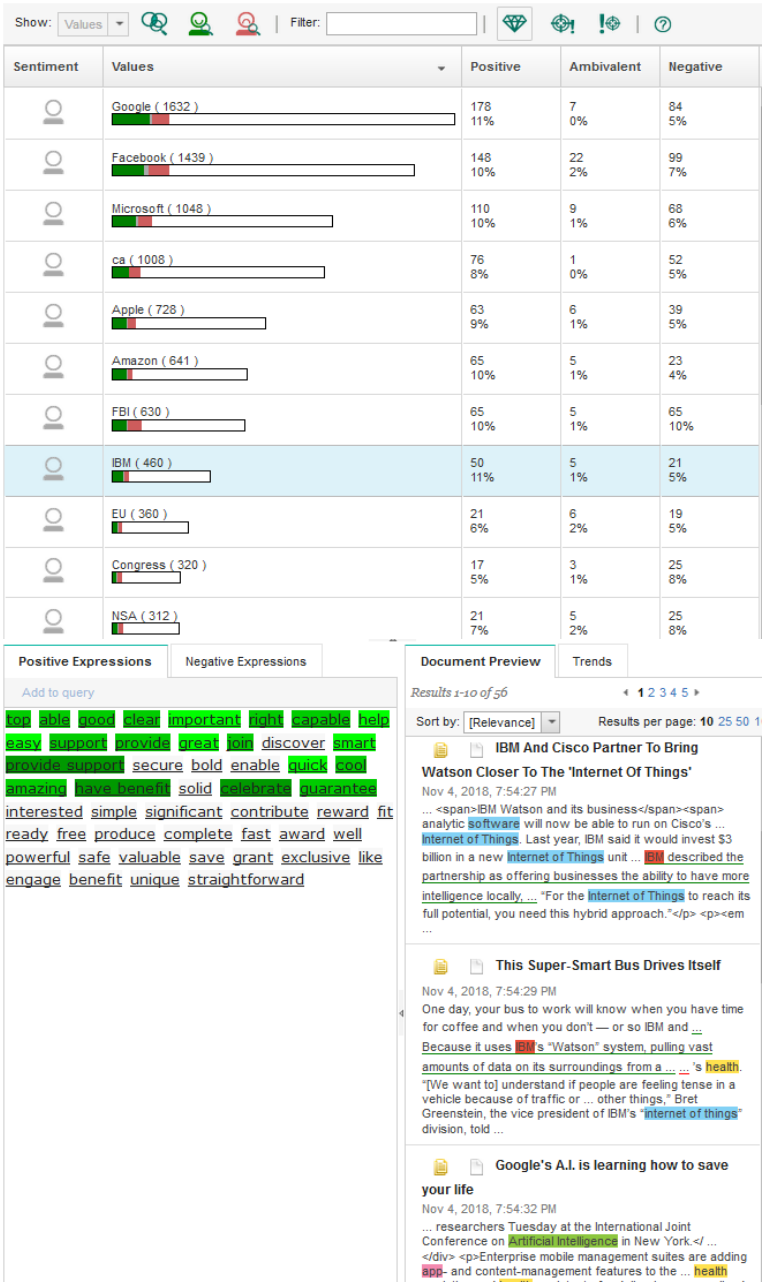


Figure 22 Sentiment view

More detailed information on the functionality of the Content Analytics Miner and the Watson Explorer in general can be found at Mak et al. (2018) and at Zhu et al. (2014).

5 Findings

The following chapter presents the results of applying the in chapter 4 described methodology on the data foundation shown in chapter 3.

The first step was to develop a filter to get the relevant data for this research. A news article was considered as relevant if it contains information about digital healthcare. To find these articles in the datasets different approaches were tested.

A short manual analysis of the document content was performed for each dataset to get an overview of what the data is about. After that, a linguistic keyword search was developed. To take into account the different languages, keywords in English as well as Germen were used. To get more precise results, not only keywords but also all grammatically related words were searched using the word stem search with the ~ operator. The following query has been developed "`(Health~ OR Gesundheit~) AND ((Security~ OR Crime~ OR Attack~ OR Angriff~ OR Hack*) OR (Cyber* OR Digital~ OR IT OR Information Technology))`". This method was quite successful to get an overview of the data however also false positive articles were found. Especially articles about heart or panic attacks were frequent. Comparing the results to the manual analysis of the text, also documents were missing. This led to the assumption that a much bigger list with keyword would be needed for this approach to succeed.

For that reason, semantic fields were used in the second attempt to get more precise results. A semantic field is a set of words that are grouped semantically together, meaning that all words in the field share a common semantic property (Brinton, 2000, p. 112). For this research, digital healthcare was assumed as the intersection of healthcare and information and communication technology. Because of that, two semantic fields were used, one for "Healthcare" and one for "Information and Communication Technology". Semantic fields are also created by cyber soldiers as part of their work in the CDRC. They contain words that are somehow related to a given topic. For example, the semantic filed for "Healthcare" contains words like "medical", "hospital", "ill", "cure" or "treatment". Since the semantic fields developed by the cyber soldiers had a strong context to defence, the word lists had to be updated and adapted to cover the whole scope of healthcare and ICT. Additionally, the semantic fields of the cyber soldiers did not contain inflections, synonyms or abbreviations. These have been added to

get a more precise result. Subsequently, the semantic fields were manually translated to English. An automated approach using Google Translate did not succeed because some words that had been automatically translated did not fit to the context that was needed. One example is the word “behandeln” which was translated to “dealing” (“dealing with a problem” – “ein Problem behandeln”). However, in the context of healthcare this should be “treating” (“treating a patient” – “einen Patienten behandeln”).

To get the relevant documents for this research, annotators with the terms of the two semantic fields have been created. In the Content Analytics Miner, the two semantic fields were intersected using a search query for words from both fields with the Boolean AND operator. This search now only returns documents that contain at least one term from the semantic field healthcare and one term from the semantic field ICT.

The process of adapting the semantic fields was performed iteratively to add useful words and remove words that may have a double meaning until the results were satisfying. One example of a term that produced many false positive results was the German term for genome “Gen”. This was mentioned as an abbreviation for the military rank “General” as well as in the context of “next gen” (next generation) technology. Another example of a term that had to be removed because it produced many false positive results was the word “Operation” (German for “surgery”) which can also be a military operation in English.

Since the German language has many compound words like “Krankenhausbett” (hospital bed) or “Gesundheitsministerium” (ministry of health) it was defined additionally for every entry in the semantic fields what place a term is allowed to have in a word. Meaning if the term should stand alone as a whole word, at the beginning of a word, in the middle, or at the end of a word. In the above-mentioned examples, “Krankenhaus” (hospital) as well as “Gesundheit” (health) were defined as terms that are allowed to stand alone and in the begin of words.

Another problem was American and British English, since the dataset contained both languages, all different spellings had to be implemented into the semantic field. The iterative implementation of all these findings in adaption processes was time consuming because each step required a full index rebuild so that the new semantic fields could be tested on the data.

Synonyms, inflections, words with double meaning or compound words demonstrate, that a simple keyword list is not enough to extract the relevant documents. As described above, complex linguistic rules have to be explicitly created to every use case and for every language to get a high-quality result. Just using available terminology lists and translation services is not enough because it misses the specific context of the data and the research questions.

Since the language use of societies is always changing (Aitchison, 2005), updating the semantic fields is crucial if they are used over a longer time period. This adaption can be done using the part of speech analysis in the Watson Explorer. A periodical check what nouns, verbs or phrases have a high correlation to healthcare, cybercrime or ICT reveals new terms in the area. Using that information, the keywords can be added to the semantic fields to make them more precise.

After the relevant documents could be identified, the central part of the content investigation was conducted using the in chapter 4 described methodology. The different datasets were explored using the metadata, annotations of the semantic fields and the different views that are offered by the Content Analytics Miner.

A time series analysis over all datasets with only the relevant documents (meaning documents that contain words from the semantic field healthcare and ICT) showed that news about cybercrime in healthcare has increased steadily over the last years. This increase indicates the importance of digital healthcare. Additionally, it demonstrates also that each dataset holds information for analysis. None of the used datasets was without relevant documents for this research. As one example, Figure 23 shows the increase of relevant data in the Cyber dataset.

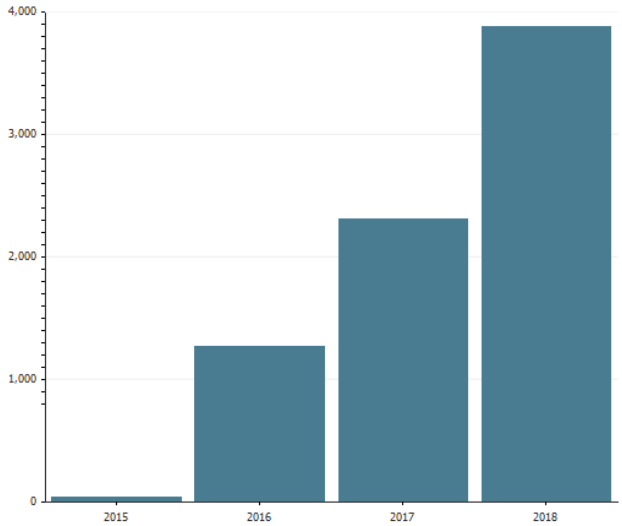
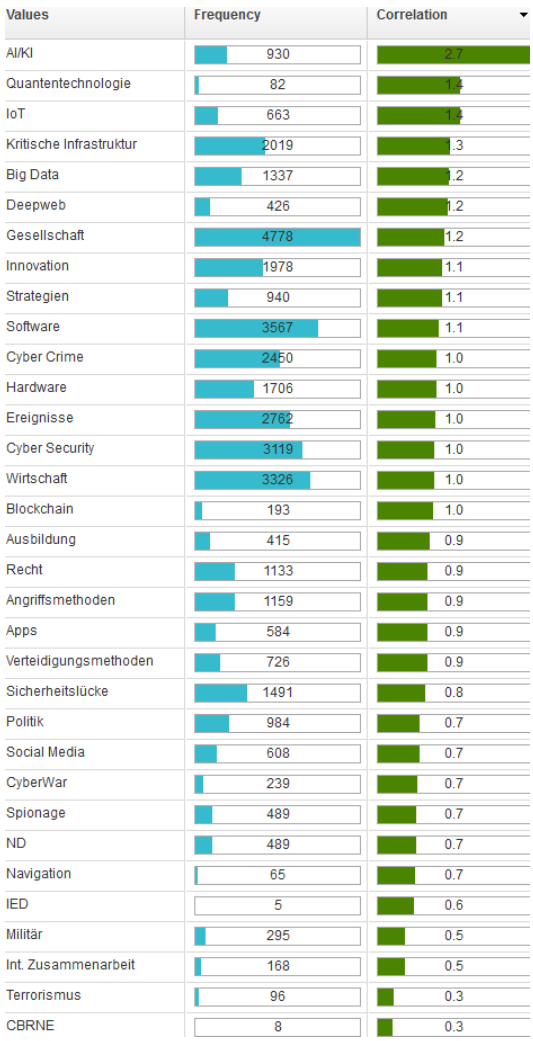


Figure 23 Time series “digital healthcare” per year, CDRC Cyber

To answer the research question what topics are important in the area of digital healthcare, the metadata tags of all articles that contain information about digital healthcare have been analyzed. Figure 24 shows the cyber topic facet values of all articles about digital healthcare in the Cyber dataset. The most relevant topic seems to be artificial intelligence (“AI/KI”) but also Quantum Technology (“Quantentechnolgie”), Internet of Things (“IoT”) or critical infrastructure (“Kritische Infrastruktur”) have high a high correlation score to digital healthcare.

This analysis was the first indication that security is also vital in the context of digital healthcare since Cyber Crime, Cyber Security, defence methods (“Verteidigungsmethoden”), intelligence (“Spionage”), vulnerabilities (“Sicherheitslücken”), CyberWar and terrorism (“Terrorismus”) were mentioned. The high frequency of these terms also supports this relevance.

As described in Figure 12 of the chapter before, the correlation value indicates the level of uniqueness of the topic. A topic with only a high frequency could also mean that the topic is overrepresented in general in the dataset meaning that it is maybe a global trend, not related to digital healthcare. Since the correlation calculation compares the frequency in the context of digital healthcare with the frequency of the tags in the whole dataset, topics with a high correlation are can be assumed to be more connected to digital healthcare.



Values	Frequency	Correlation
AI/KI	930	2.7
Quantentechnologie	82	1.4
IoT	663	1.4
Kritische Infrastruktur	2019	1.3
Big Data	1337	1.2
Deepweb	426	1.2
Gesellschaft	4778	1.2
Innovation	1978	1.1
Strategien	940	1.1
Software	3567	1.1
Cyber Crime	2450	1.0
Hardware	1706	1.0
Ereignisse	2762	1.0
Cyber Security	3119	1.0
Wirtschaft	3326	1.0
Blockchain	193	1.0
Ausbildung	415	0.9
Recht	1133	0.9
Angriffsmethoden	1159	0.9
Apps	584	0.9
Verteidigungsmethoden	726	0.9
Sicherheitslücke	1491	0.8
Politik	984	0.7
Social Media	608	0.7
CyberWar	239	0.7
Spionage	489	0.7
ND	489	0.7
Navigation	65	0.7
IED	5	0.6
Militär	295	0.5
Int. Zusammenarbeit	168	0.5
Terrorismus	96	0.3
CBRNE	8	0.3

Figure 24 Topics of Digital Healthcare, Cyber

An extraction of the cyberthreats was conducted using the intersection of the previously modelled document collections about digital healthcare, actors in the cyber domain and hacking methods. The actors and hacking methods were

modelled as dictionary using already available lists on the internet as well as manual research to get better precision. In the context of this work, actors are hacker, hacker groups, and terror groups. More than 500 actors have been compiled to dictionaries as part of this research.

A time series view of documents containing something about digital healthcare and actors or attack methods gives insights into when cybersecurity relevant events were reported. This is demonstrated in Figure 25. Every bar in the diagram can be selected to get the articles with the relevant event behind it. The articles contain a description of the events. This is demonstrated in Figure 26.

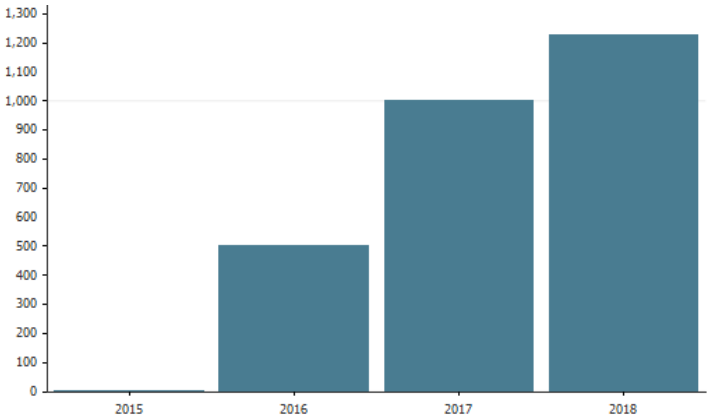


Figure 25 Time series “cyberthreats in digital healthcare” per year, CDRC Cyber

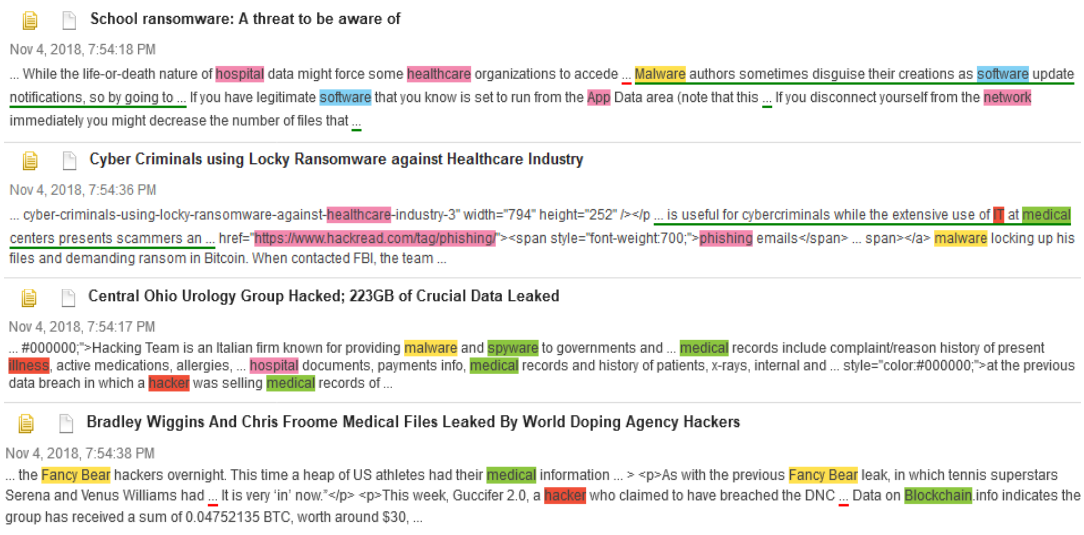
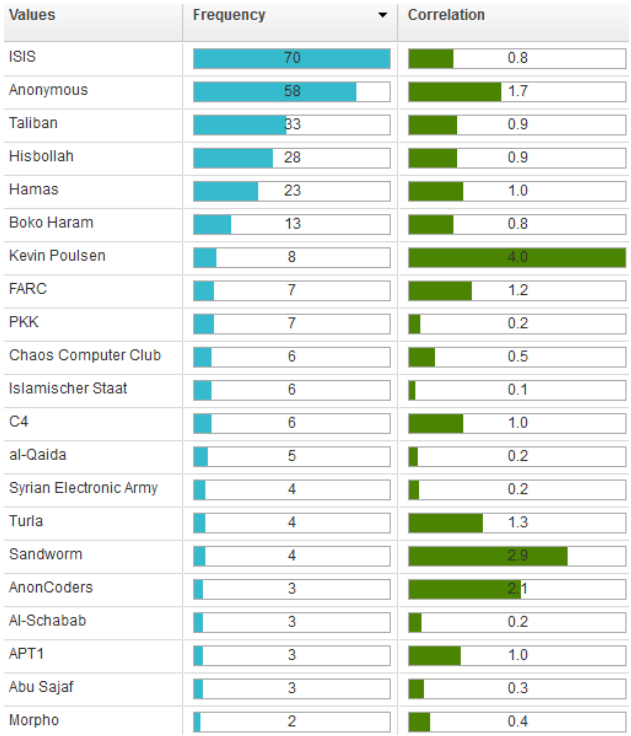


Figure 26 Document view of articles about events, Cyber

The attack methods, targets, and actors are automatically highlighted in the summary of the document view. This gives a quick overview of the context of the event. The near duplicate detection of the Watson Explorer makes it also possible to hide more mentions of the same event.

Comparing the chart of Figure 23 to the one from Figure 25, it can be seen that both, the numbers of news about digital healthcare as well as the number of news about cybercrime in digital healthcare rose steadily over the last years.

To better visualize the actors that play a role in the domain cybersecurity for digital healthcare, the extracted actors have been pictured in Figure 27. There terror groups like Hamas, Hezbollah or Taliban as well as hacker groups are mentioned indicating that terror networks are also involved in cybercrime. The low correlation of the terror groups is due to the fact that they are outside of the context digital healthcare mentioned more frequent. However the high frequency still indicates that terrorists are involved in cybercrime in digital healthcare.



Values	Frequency	Correlation
ISIS	70	0.8
Anonymous	58	1.7
Taliban	33	0.9
Hisbollah	28	0.9
Hamas	23	1.0
Boko Haram	13	0.8
Kevin Poulsen	8	4.0
FARC	7	1.2
PKK	7	0.2
Chaos Computer Club	6	0.5
Islamischer Staat	6	0.1
C4	6	1.0
al-Qaida	5	0.2
Syrian Electronic Army	4	0.2
Turla	4	1.3
Sandworm	4	2.9
AnonCoders	3	2.1
Al-Schabab	3	0.2
APT1	3	1.0
Abu Sajaf	3	0.3
Morpho	2	0.4

Figure 27 Actors in Digital Healthcare, KriMiSi

The dataset of the Austrian press releases supports that finding too. Figure 28 shows a facet analysis using the documents of the Austrian press releases that have something to do with digital healthcare (documents that contain words from the semantic field healthcare and ICT) and the corresponding tags facet. The tags were part of the metadata that came with the dataset. The correlation and frequency values underline the previously stated assumption that digital technologies in healthcare are indeed related to terrorism (“Terrorismus”), bombing (“Bombenanschlag”), crime (“Kriminalität”) and cybercrime (“Computerkriminalität”). Even though some topics have higher correlation, which is quite logical since the dataset deals with a diverse topic of political news and not only terror or crime news, every terror related tag (the blue highlighted) has a correlation value higher than one which means it is noticeable.

Values	Frequency	Correlation 1
Nikotin	6	6.5
Medizin	6	6.2
Gesundheit	12	4.9
Innovationen	2	4.8
Europäisches Forum Alpbach	5	4.5
Computerkriminalität	11	4.1
Videospiele	2	4.0
Biowissenschaften	1	4.0
Normen	1	4.0
Pharma	2	3.3
Krankenhäuser	3	3.2
Gesundheitspolitik	6	2.9
Forschung	5	2.9
IT	6	2.9
Gastronomie	4	2.7
Kriminalität	8	2.5
Ermittlung	18	2.0
Tirol	6	2.0
Wien	34	1.9
Strafverfolgung	5	1.8
ÖVP	14	1.6
Anschlag	33	1.6
Deutschland	40	1.4
Armut	5	1.4
Europol	2	1.4
Ärzte	2	1.4
Terrorismus	50	1.3
China	8	1.1
Bombenanschlag	7	1.1

Figure 28 Tags correlating with digital healthcare

Querying only documents about digital healthcare mentioning terror groups and creating a time series analysis (pictured in Figure 29), a drastic increase in mentions in the year 2019 can be observed. More than twice as much articles with cybercrime and terrorists were found in the first three months 2019 compared to 2018.

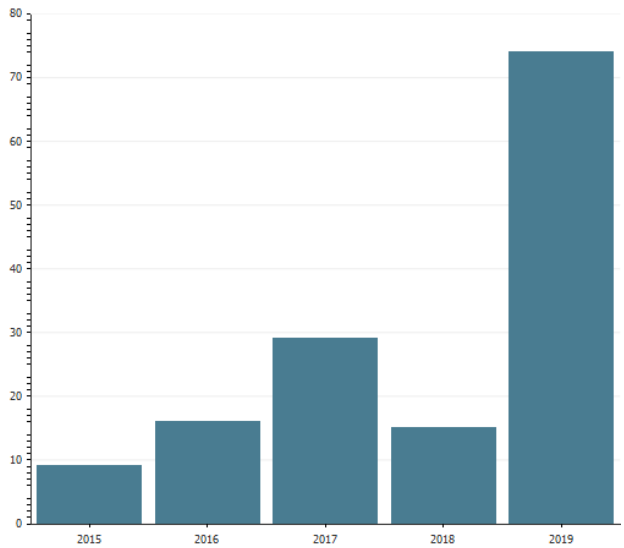


Figure 29 Raise of terrorists involved in cybercrime in Digital Healthcare, KriMiSi

This unusual increase was investigated further and put in relation to the number of total cyberattacks mentioned in the KriMiSi dataset in Table 1 and in the Cyber dataset in Table 2.

Year	involvement of hackers & terrorists	involvement of terrorists	% of involvement of terrorists
2015	18	9	50%
2016	17	16	94%
2017	30	29	97%
2018	16	15	94%
2019	129	47	36%

Table 1 percentage of mentioned hacking attacks by terrorists, KriMiSi

Year	involvement of hackers & terrorists	involvement of terrorists	% of involvement of terrorists
2015	0	0	0%
2016	75	21	28%
2017	99	28	28%
2018	98	21	21%
2019	35	8	23%

Table 2 percentage of mentioned hacking attacks by terrorists, Cyber

Table 1 and Table 2 both show that the relative numbers of terrorists involved in hacking events in digital healthcare even decreased in KriMiSi and stayed approximately constant at the cyber dataset.

Using the text mining approach from above the attack methods were extracted.

Figure 30 shows attack methods that have a high correlation with digital healthcare, indicating what attack methods are used in this area. On the left side are the extracted attack methods from the KriMiSi dataset, on the right side the extracted attack methods from the Cyber dataset. Both have been filtered to show only articles in the context of digital healthcare. Even though the same attack methods have been extracted from both datasets, they derive in frequency and correlation because each dataset contains data from a different domain. The attack methods that have a high correlation in KriMiSi show how digital health technology has been attacked from a military point of view while the high correlating attack methods from the Cyber dataset are more related to the general area of digital healthcare. These differences underline the importance to investigate domains out of different perspectives to get a holistic view.

Important to notice is that this analysis only reveals the attack methods that are mentioned in news articles. Especially in the military context, some attacks are not even discovered or broadcasting is suppressed. If so, they will not show up using this methodology and are therefore not detected in this research.

Values	Frequency	Correlation
computer worm	16	4.0
spyware	45	2.8
social engineering	26	2.4
false flag	12	2.3
botnet	48	2.2
denial of service	73	2.2
targeted attack	27	2.1
buffer overflow	10	2.0
malware	218	2.0
exploit	280	2.0
computer virus	21	1.9
ransomware	42	1.8
phishing	61	1.6
heartbleed	14	1.5
keylogger	13	1.4
brute force	9	1.2
trojan horse	22	1.1
smurf attack	3	1.0
drive-by attack	6	1.0
password attack	5	0.8
dictionary attack	2	0.7
cross-site scripting	7	0.4

Values	Frequency	Correlation
ransomware	1179	2.8
targeted attack	145	1.9
social engineering	168	1.7
phishing	610	1.6
computer virus	33	1.5
malware	1437	1.5
botnet	297	1.4
exploit	1789	1.3
denial of service	362	1.3
spyware	99	1.2
DLL hijacking	10	1.2
false flag	23	1.2
brute force	72	1.1
DNS hijacking	13	1.1
heartbleed	18	1.0
computer worm	11	1.0
SQL-Injection	41	1.0
Man-in-the-middle	32	1.0
compromised account	31	0.9
buffer overflow	39	0.9
cryptojacking	35	0.9
smishing	7	0.8

Figure 30 Attack methods that correlate with Digital Healthcare, left KriMiSi, right Cyber

In a military context, computer worms, spyware, and social engineering seem to be relevant as the investigation of the KriMiSi dataset showed. Out of the scope of the Cyber dataset, ransomware, targeted attacks, and social engineering are high correlating attack methods in the digital healthcare domain. In both datasets, malware and exploits have the highest frequency but are only in the upper middle when sorting by correlation value. This lower correlation score is because both malware and exploits are the most dominant attack methods in the unfiltered datasets making them less noticeable in the context of digital healthcare.

Using the facet pair analysis, attack methods were correlated to actors. This way, it can be found out which attack method is used by what actor. The semantic field approach was applied again to get only the actors that are mentioned in the context of digital healthcare.

As pictured in Figure 31, this combination of the analysis above shows some strong correlations between actors and attack methods. Doing a manual crosschecking of the articles containing the facet pairs with high correlations supports that these hackers are in connection to the found attack methods. Additionally, the correlations were investigated using internet searches to make sure that the articles in the analysis have not been biased. As one example, a web search for Albert Gonzalez revealed that he is a known hacker who used SQL injections to steal computer data from internal cooperate networks. Because of that, the high correlation between SQL Injection and Albert Gonzalez seems logical.

Rows: Attack Methods	Columns: Hacker Groups	Frequency	Correlation
SQL-Injection	Albert Gonzalez	3	30.2
false flag	Lazarus Group	7	19.4
targeted attack	APT29	4	16.7
targeted attack	APT28	11	9.2
targeted attack	Fancy Bear	9	8.1
computer worm	Tarh Andishan	1	7.2
targeted attack	Anonymous	13	6.8
denial of service	Anonymous	56	6.2
targeted attack	Ghost Squad Hackers	2	5.3
smishing	Shortcut	1	5.1
false flag	Sandworm	2	5.1
spyware	APT32	2	5.1
targeted attack	Turla	3	5.0
targeted attack	Lazarus Group	6	5.0
false flag	Scarcruft	1	4.6
botnet	APT28	19	4.5
botnet	Fancy Bear	16	4.1
DNS hijacking	Anonymous	4	4.0
false flag	APT28	4	4.0
malware	Lazarus Group	45	3.9
denial of service	Lizard Squad	4	3.5
phishing	Fancy Bear	29	3.5
phishing	APT28	32	3.4
denial of service	APT28	20	3.0
Man-in-the-middle	Scarcruft	1	2.9
DNS hijacking	APT17	1	2.6

Figure 31 Facet pair view with attack methods and actors, Cyber

When investigating technologies that correlate to digital healthcare a correlation analysis can be valuable as well. Figure 32 indicates that nouns like “robotics”, “exoskeleton”, “AI”, “biotechnology”, “simulation” or “mobile devices” have a high correlation with the topic digital healthcare.

Values	Frequency	Correlation
Robotik Exoskelett	28	3.1
Künstliche Intelligenz Biotechnology	24	2.8
Mobile Devices Biotechnology	12	2.1
Künstliche Intelligenz	123	1.6
Biotechnology Robotik	7	1.6
Robotik Nano Tech	5	1.4
Biotechnology IKT	6	1.4
Biotechnology	106	1.1
Künstliche Intelligenz Robotik	20	1.1
Exoskelett	9	1.1
Robotik	39	0.9
Autonome Systeme Künstliche Intelligenz	19	0.9
Simulation	15	0.9
Big Data / Cloud Künstliche Intelligenz	8	0.8
Biotechnology Simulation	5	0.8
Biotechnology Nano Tech	13	0.8
IoT	19	0.8

Figure 32 Technologies in Digital Healthcare, InnoTech

An analysis of these terms using the time series view shows that mentions have increased steadily over the past years. A cross-check with Google Trends underlines this result.

Since trends cannot be known so easily before they appear and are mentioned, a dictionary approach would not be helpful to uncover them. The dictionary would most probably not contain the values needed to uncover the trend. Because tends are well described by nouns, the nouns were extracted and investigated to see if there are sharp increases that could indicate a trend. The trend and deviation view, together with the noun facet was used to identify new technologies in digital healthcare. As described in 4.6, the trend view can give insights into unexpected changes in frequency or correlation values of facet entries. Figure 33 shows the noun sequence facet of all documents mentioning terms from digital healthcare in the trend analysis. The peaks of the terms found with the trend analysis in Figure 33 were also found using Google Trends. As one example, Elon Musk was picked out in Figure 34 to demonstrate this.

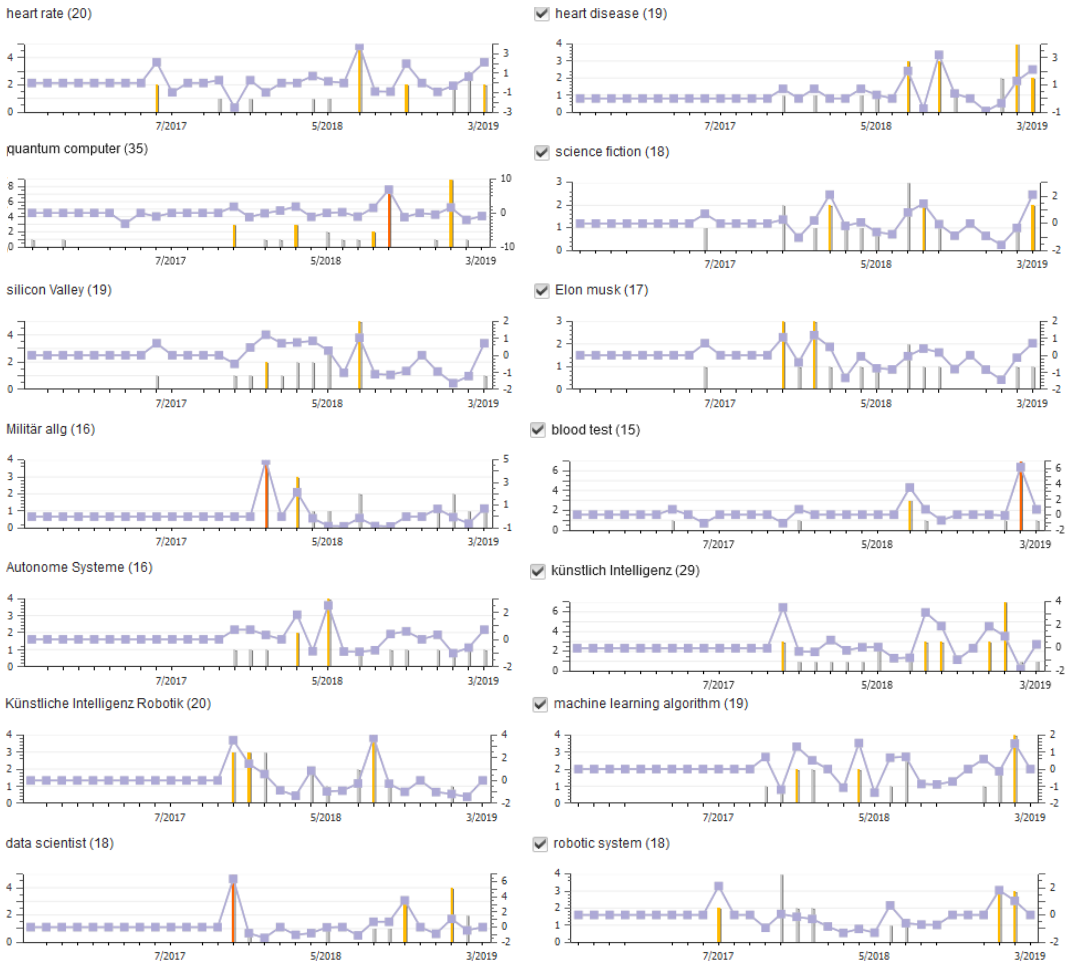


Figure 33 Noun Trends Digital Healthcare, InnoTech

Figure 34 Google Trends Elon Musk, Data source: Google Trends (https://www.google.com/trends)

To get a more human-friendly representation of trending words, the charts from Figure 33 were reorganized to word clouds using the Dashboard function. These generated word clouds are pictured in Figure 35.

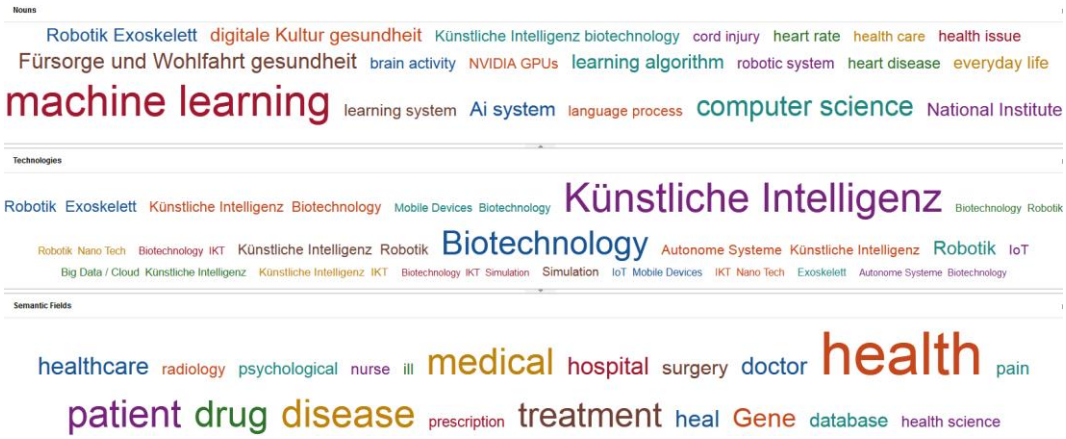


Figure 35 Trends in Digital Healthcare Word Clouds, InnoTech

Trends shown in Figure 35 underline the previously stated assumption that AI, biotechnology, robotic and IoT play an increasing role in the domain. The health-related trending words as health or medical are noise for this analysis. However, they can be interpreted as indicators for the correctness of the filtering of digital healthcare related documents because mostly ICT or health-related words were found.

To conduct an extraction of the sources that report about digital healthcare, the domains of the source URLs of the CDRC articles were annotated using dictionary and parsing rules. The Lexis Nexis data had this information in the filename. To extract the URL from the filename, a character rule was developed. In the GDELT dataset, this information is already in the metadata so no additional annotations were needed.

Figure 36 shows the extracted sources that report about digital healthcare in the InnoTech dataset.

Some datasets like the CDRC Cyber contains also author information that can be correlated to digital healthcare. Using this information it can not only be seen

which source but also which author reports about digital healthcare. This is pictured in Figure 37.

Values	Frequency	Correlation
medicalfuturist.com	16	2.6
medgadget.com	22	2.1
nvidia.com	34	2.0
intelligentrobots.org	8	1.5
earthlink.net	8	1.5
singularityhub.com	14	1.4
sciencedirect.com	8	1.4
sciencedaily.com	87	1.3
technologyreview.com	46	1.3
nature.com	25	1.2
wired.co.uk	20	1.2
ox.ac.uk	13	1.1
futurism.com	13	1.0
technewsworld.com	10	1.0
arxiv.org	14	1.0
phys.org	31	1.0

Figure 36 Sources writing about digital healthcare, InnoTech

Values	Frequency	Correlation
Marianne Kolbasuk McGee	169	9.6
Jessica Davis	116	8.9
Ryan Fahey	13	5.4
Paolo Passeri	38	5.3
Frank Landman	11	3.9
Paramita Ghosh	11	3.7
Ryan Ayers	6	3.6
Ryan F. Mandelbaum	7	3.2
Oleg Svet	4	2.9
Dimitar Kostadinov	4	2.9
Adi Gaskell	16	2.9
Olivia Solon	13	2.7
Elie Dolgin	6	2.7
Julia Carrie Wong	13	2.4
Con	6	2.4
Louis Columbus	6	2.4
Rob Enderle	11	2.4
TIM STARKS	16	2.3
Brian Fung	12	2.3
KRISTIN HOUSER	4	2.3
Emily Waltz	4	2.3
Bruce Y. Lee	4	2.3

Figure 37 Authors reporting about digital healthcare, Cyber

Figure 38 shows the facet pair analysis of sources and the semantic field healthcare of the document collections dealing with digital healthcare. This facet pair analysis gives an indication which news source publishes information on which healthcare topic. Interesting is the high correlation despite the low frequency of the Huffington Post and paramedic. A manual investigation showed, that paramedic was found only once in the whole InnoTech dataset and there in a context of digital healthcare and in a Huffington Post article. This makes the level of uniqueness (correlation) of the term “paramedic” and its connection to Huffington Post so high. Eye-catching is also, that NVIDIA as Tech company reposts in strong correlation to radiology. A manual investigation of the nine documents that back this correlation value showed that they deal with technology NVIDIA provided for analyzing radiology images using artificial intelligence. That interesting fact indicates again how the correlation analysis can be used to find new, unexpected facts in the data.

Rows: Source	Columns: Healthcare	Frequency	Correlation
medicalfuturist.com	pharmacy	5	6.4
medicalfuturist.com	healthcare	14	4.5
nvidia.com	radiology	9	4.1
medicalfuturist.com	doctor	13	4.1
medicalfuturist.com	nurse	6	4.1
medicalfuturist.com	hospital	12	3.5
huffingtonpost.co.uk	paramedic	1	3.3
handelsblatt.com	ill	3	3.2
twitter.com	psychological	4	3.0
technologyreview.com	Gene	27	2.9
medicalfuturist.com	medical	16	2.3
nvidia.com	healthcare	16	2.3
medicalfuturist.com	patient	15	2.0
discovermagazine.com	HIV	3	2.0
medicalfuturist.com	radiology	4	2.0
technewsworld.com	prescription	3	1.9
medicalfuturist.com	drug	14	1.8
news.mit.edu	protein	11	1.8
innovationmanagement.se	prescription	2	1.7
futureoflife.org	prescription	2	1.7

Figure 38 Sources and the medical subject they report about, InnoTech

Using the same approach, sources dealing with specific technologies in digital healthcare were found out as well.

After applying the developed rules, and the different views of the Content Analytics Miner, the document sets could be reduced to an amount that made a manual investigation of the articles behind the different analyzes possible. This step was essential to evaluate the quality of each dataset and the annotations.

Additionally, the results of each analysis in this chapter could be checked for plausibility to reduce the risk of false correlations.

6 Evaluation

This chapter deals with a short evaluation of the findings presented in chapter 5. To not overload this research, a detailed mathematical evaluation of all results is out of the scope. However, the key parts of this work shall be analyzed using a small sample for manual cross checking to get an indication of the validity of the results.

The most important part for the further analysis of the research topic was the extraction of relevant documents (documents that deal with digital healthcare). The semantic field approach increased the number of found documents that are relevant to the analysis with a factor 10 compared to the linguistic keyword search. To review the performance of the semantic field approach, documents that have been manually tagged with healthcare from the CDRC Cyber dataset were compared to the documents found using the semantic fields. This revealed a recall score of 0.686 (306/446). A manual analysis of the difference showed that the manual tagging was not precise in about 30% of the cases. This underlined the success of the semantic field approach. However, one explanation for this lack of precision in the manual tagging is that this has been performed by the cyber soldiers of the DocCenter out of a defence context which differs from a general point of view.

To overcome this, a sample of 100 documents out of the 985 digital healthcare related ones that have been found in the InnoTech dataset using the semantic fields healthcare and ICT were manually analyzed to see how many articles really deal with digital healthcare. Out of the 100, six articles were found that are not related to digital healthcare. Two examples (document id 914 and 727) that were false classified mentioned “pain points”, where “pain” was classified as a health-related term. Another false classified document (id 845) was explaining the Doppler Effect with the example of an ambulance driving by. Two other wrongly classified articles had phrases like “because of her patient approach” (id 653) or “regulations like in medicine” (id 747) in them.

One difficulty in the evaluation of the semantic field approach was the definition of digital healthcare. Since this is an interdisciplinary cross-domain issue, it is very subjective which topics are about digital healthcare and which are not. One especially difficult example were technologies that are used for improving the health of animals. The semantic field approach defined this as digital healthcare

related. Another article dealt with new technology for increasing the health of plants and the last on was classified digital healthcare related and dealt with technology to improve the “health” of cars. In this example, the articles about animal and plants were categorized as true positives, the one about the car health (id 596) as false positive since a car is not a living object. However, this manual classification is open to discussion.

To evaluate the found technology topics relevant to digital healthcare (Figure 24) the articles behind the five topics with the highest correlation value to digital healthcare (“AI/KI”, “Deepweb”, “IoT”, “kritische Infrastrukur”, “BigData”) were manually analyzed. 100 articles of the five topics have been read to determine if the technology topic really has a context to digital healthcare. Every article showed a context to the tagged topic, supporting the validity of the correlation factor. Five articles were wrongly classified and had nothing to do with digital healthcare (ids 2966, 3015, 7855, 14932, 17536).

The same was performed for the actors in cybercrime in digital healthcare (Figure 27). A sample with 100 articles about digital healthcare where the algorithm used found at least one of the top five actors (“ISIS”, “Anonymous”, “Taliban”, “Hisbollah”, “Hamas”) was manually investigated if the found actor really has something to do with a cybercrime in digital healthcare. 34 articles were found where the mentioned actors had nothing to do with the digital healthcare. 26 data entries had nothing to do with digital healthcare and were classified wrongly.

Common problems were advertising in the news article (id 4132), user comments that had been crawled together with the article (id 4028), different use of the word “anonymous” (id 4032), the phrase “satellite picture” where satellite was not related to the article but only the tool for a picture (id 3635), misspellings like radiological as radiology (id 3591) or different meanings of abbreviations like AI for a newspaper not for artificial intelligence (id 3286).

The higher rate of false positives in the KriMiSi dataset was expected since it contains much more articles that are neither cyber nor health-related than for instance the Cyber or InnoTech datasets. Additionally, many long articles (20 pages and more) were found. One paragraph dealt with something about digital healthcare; all other 20 pages were unrelated to the topic. This example was still coded as true positive in the manual review since a cyberthreat in digital healthcare was mentioned even if it is not the primary content or even the aim of the article. The intelligent document summary helped with the manual review because it showed on one glance where the annotations were found. Especially by articles with many pages this was a useful feature.

To further investigate the correctness of detecting the mentions of terrorists in the data, the articles of the Cyber dataset which have been tagged manually by cyber soldiers with terrorism (“Terrorismus”) has been compared to the one found

with the for this research compiled dictionary of terrorists and terror cells. 127 articles were tagged with terrorism, 97 were found using the dictionary approach. A manual investigation (query of the document containing information about digital healthcare and the tag “Terrorismus” and not any entry from the dictionary) showed that only 24 of the 127 tagged documents dealt with a concrete example of terrorism in digital healthcare. The other 103 tagged articles described technology or vulnerabilities that could hold a potential for terrorists. They did not contain any mention of terrorism or terrorist groups. Because of these differences, a comparison of the articles did not make sense. Since the article have been read anyway, the count of false positive digital healthcare documents was statistically investigated. Out of 100 retrieved digital healthcare documents tagged with “Terrorismus” were 15 false positive.

The mean of the false positive rate of the detection of articles about digital healthcare in the above examples is about 13% with a standard deviation of 9.8%. A boxplot representing the statistic of the evaluation of the semantic fields is visualized in Figure 39.

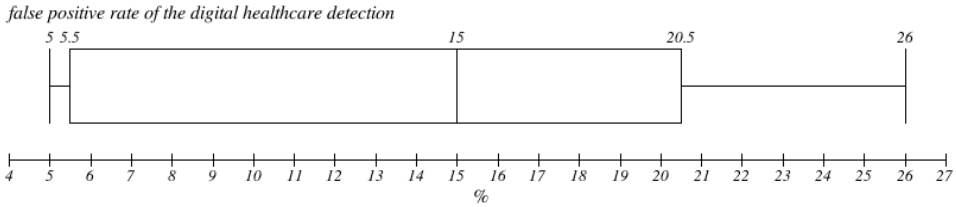


Figure 39 Boxplot - False positive rate of the digital healthcare detection

The evaluation underlined that the semantic field approach, as well as the investigation approach using correlation, delivered good results. Since the wrong classified articles were so less, no further improvements have been done on the semantic fields after the evaluation. Especially, because the removal of terms from the semantic fields could lead to many false negatives.

7 Discussion

The results presented in chapter 5 underline again that text mining is an important tool for cybersecurity and trend research. The findings of this work prove that this technology brings additional value. It seems unlikely that the same results could have been achieved using manual coding techniques. One reason for that is the high and vast growing amount of data. Especially large datasets like GDELT make manual coding time consuming (Müller et al., 2016) and imprecise (Indulska et al., 2012). The approach used in this research offers the same precision for small data sets as well as for large ones.

An important fact is that the used approach still requires manual work for high-quality results. The knowledge base, dictionaries, annotations, and modelling the fact extraction involves manual work. These steps have a substantial impact on the analysis results. If the model, in this case the semantic fields, is not accurate, the whole analysis can produce misleading results. Because of that, also the cross-check of the outcomes of the analysis is so important. This check is a task which is challenging to automate at the moment. Artificial intelligence lacks the semantic general knowledge that humans have to understand and analyze an interdisciplinary cross-domain subject like digital healthcare for validating the results of the text mining analyzes in this work. In literature, this lack of semantic knowledge is also referred to as association function problem and symbol grounding problem (Lu, Li, Chen, Kim, & Serikawa, 2018). Because of that, humans need to perform quality control of the system's output themselves.

Still, the text mining algorithms used in this research have the advantage that once the model has been created, it can be massively parallelize using multithreading and server cluster to improve performance for big data analysis. In contrast to manual coding it scales easier not only from a time and cost perspective but also from a quality point of view. It offers the same precision and recall scores regardless of the size of the datasets.

Furthermore, the Watson Explorer supports the already argued checked for plausibility well by allowing to explore the documents behind a result that back the calculated statistics or diagrams anytime. Every analysis outcome like a facet value, facet pair, or chart can be cross-checked that way. This was done to manually evaluate the plausibility of the answers to the research question in this

work. Using this manual cross-reading, the rules for the analysis can be enhanced as well for a continuous improvement of the text mining model.

Since the GDELT dataset only contained structured metadata, text analytics was more difficult. There the Watson Explorer, as a text mining tool, could not bring the full value. That is why most of the investigations have been done using SQL on Googles BigQuery interface. However, the different structure makes it questionable if it should be investigated together with highly unstructured data like from the CDRC. Nevertheless, the GDELT dataset supported the findings of the other data sources also showing that large-scale international datasets underline the results of this research. This positive crosscheck is also an indicator that shows that the choice of high-profile sites and other data sources from the DocCenter of the Austrian Armed Forces was well selected to provide the information needed for this research.

The more global datasets become, the more language problems begin to matter. Because both German and English articles were together in on dataset, the analysis could be less precise. On the other hand, the more the datasets are split up the more difficult it gets to have a holistic view over everything. On the other hand, putting all data together in one big collection makes the text mining more difficult because of the heterogeneous data out of different context. For this research, the decision was made to investigate the datasets without manipulation meaning with mixed languages and not transformed into one collection. The reasons for keeping the different datasets separated were the data size and the different metadata and structure of the datasets. Additionally, each dataset had one special focus that helped to investigate the research questions out of different perspectives. While the cyber dataset had more general articles about ICT, InnoTech focused more about technology. KriMiSi had a strong focus on the military side of the research questions. Nearly the same entities could be extracted from each dataset but the frequency and correlation derive strongly. The deviation can be traced back to the different context of the data collection. These differences underline the importance to investigate domains out of different perspectives, meaning using different datasets, to get a holistic view and to demonstrate the robustness of the results.

Because the CDRC project began in 2015, that date is the starting point of the documents in the study. The end of data investigated here was 30.05.2019. Since the research pointed out that most changes are happening now, a further investigation in how the trends proceed should be conducted.

Even though Google Trends has an entirely different approach than the text mining carried out in this research (counting user searches not mentions in a text), the terms found using trend analysis in Watson Explorer also have similar

peaks in the Google Trend charts. This is one indicator of the validity of the performed trend analysis.

The text mining models developed for this work were mostly dependent on rules and statistics. No machine learning or artificial intelligence approach was used. Even though the Watson Explorer would support that methodology through its oneWEX and Watson Knowledge Studio component, the research would lose its adaptability, explainability, and reproducibility. The same rules and statistics applied to the dataset always will result in the same outcome. The other way around, every result of the text mining process in this work can be traced back to specific rules. Facets, facet pairs, charts or bars can be selected to investigate the documents and rules that extracted the annotations in this document that back the statistic. This fact makes the research highly explainable and gives additional insights to uncover the “why behind the what” meaning not only results but also the context of the results can be investigated. In contrast, the supported machine learning methods for the Watson Explorer (Watson Knowledge Studio) are not explainable. Annotators can be trained and then they produce a result, what training data is responsible for which result cannot be traced. Also the adaption of machine learning annotators for the Watson Explorer is more complicated. The annotator can be retrained but then implications on the results cannot be estimated beforehand. A small adaption in the training data could lead to a completely different output. Because of the mentioned limitations in the used tool machine learning was out of the scope for this research.

The downside of the statistical approach used is the strong focus on correlation. The correlation-based investigation has the risk of finding fake correlations or misinterpreting correlations as causation. That risk was defused in this research by doing a manual plausibility check on the articles behind the statistics shown in this document. As stated above, the Watson Explorer supported that approach well with an intuitive user interface and an intelligent document summary and highlighting in the document view that made an effective cross-reading of many articles possible. This cross-reading also helped to adapt the rules and to understand the context of the statistical result in the text.

Special attention should also be given to the rising problem of fake news. The research assumes the truth of all articles and relies on the fact that the DocCenter from the National Defence Academy (DocCenter/NDA) of the Austrian Armed Forces checks every article in the CDRC for accuracy and plausibility. The GDELT dataset contains only non-validated statements. Since GDELT was only used additionally to the other datasets, the potential bias is low. However if fake news nevertheless have found their way into the CDRC datasets, the analysis could be biased because of that.

Additionally, the frequency count has to be interpreted carefully. A term with high frequency is not necessarily more important than a low frequency term. The fact that something was mentioned more often does not indicate importance or relevance, especially because importance always depends on the context of the observer. The same goes for the correlation. A high correlation not necessarily means high relevance. It is essential to know how the correlation is calculated (see Figure 12) to understand how results can be interpreted.

In general, the used methodology worked as good as predicted beforehand. Only the data gathering and pre-processing, so that all articles could be indexed, was more work than assumed; mainly because of the complexity of the data and the need to manually adapt the import interfaces of the Watson Explorer. Also, the iterative adaption process of improving the semantic fields was more time consuming than expected. The complexity of the German and English language was underestimated. Examples for the problems that have been dealt with are compound words, translation, American and British English or context-dependent meaning of words. This complexity plus the fact that every iteration comes with a full re-index of all datasets (about 28h processing time) so that the new semantic field annotations could be applied to all articles, made the adaption of the semantic fields one of the most time-consuming tasks.

Subsequently, the final evaluation showed that the long work on the semantic fields was successful. A low false positive rate has been scored by the identification of digital healthcare related articles. Since a detailed evaluation was out of the scope, the low sample rate of 400 articles (four times 100 articles on different datasets) was acceptable for this research.

The used semantic filed approach seems robust and not dependent on the data source. The precision varies, as it can be seen by the differences in false positive mentions between Cyber and KriMiSi however the false positive rate was all the time in the lower area. This high precision indicates that the semantic field approach is generalizable and can be used on different data sources.

The results of the research seem in general plausible. That digital healthcare is mentioned more and more was expected. It was also expected, that this new area of technology brings possibilities for cyberattacks. The found technology trends seem to be logical as well. The two main unexpected results of this research were first that the attack methods and actors could be found out using correlation. Manual investigation of the documents and events that led to these correlations showed good precision and proved the high quality as well as the efficiency of this approach. The second unexpected result was that terrorists were involved in such a high percentage of cyberattacks in the healthcare area. The articles showed that they were not only perpetrators but especially the Islamic State was also a victim of cyberattacks. Interesting as well is that the

percentage of terrorists involved in cybercrime seems to be declining in 2019. Because only data until 31.03.2019 was accessible at the time of this research, it is unsure how representative this time period is.

All findings indicate the importance of cybersecurity in general as well as special for digital healthcare. They also revealed that the victims are very diverse. They reach from large-scale companies, single entrepreneurs, consumers, and states even to terror groups. Because of that trend and our dependence on technology, security considerations have to be part of every technology, not only in the development process but also over the whole lifecycle.

8 Conclusion & further work

According to the findings in chapter 5, all stated research questions from section 1.3 could be answered.

An overview of the topics of digital healthcare is pictured in Figure 24 and in the word clouds in Figure 35. On one side technical terms like AI, robotics, machine learning, IoT or biotechnology play a role, on the other side also patient, drugs, diseases, treatment or genes are important concepts in the domain.

The events can be derived from the time series view pictured in Figure 25. To find out what events are behind the chart, the bars of the diagram can be selected and the different articles can be viewed. That is shown in Figure 26.

Using facet and correlation analysis, the in Figure 27 described actors were extracted. Traditional hacking groups together with terrorist cells and networks could be found. The percentage of terrorists involved in cybercrime in digital healthcare seems to be declining.

Attack methods in combination with digital healthcare could be listed using the same methods. This is pictured in Figure 30. The facet pair analysis in Figure 31 revealed what attackers use which attack method.

New technologies were found with the trend view pictured in Figure 33 and the facet analysis pictured in Figure 32.

Sources that report about digital healthcare could be parsed out using custom annotators (Figure 36, Figure 38). In some cases, this information was even available in the metadata (Figure 37).

The results of this research prove that text mining news data can achieve added value for cybersecurity in the domain of digital healthcare. The reports about digital healthcare in general, as well as the reports about cybercrime and vulnerabilities in this area increased steadily over the last years. That underlines the importance of the topic. The found results are also backed up by the statements found in the literature described in chapter 1 and subchapter 2.1. Cybercrime in digital healthcare is an increasing threat for society, companies, and the individual. The proposed methodology and tools were capable of answer all research questions. The findings prove the benefit of text mining for cybercrime and trend research in digital healthcare, especially in the security

sector. At the same time, this research also underlines how vital ongoing data collection is. Without big historic high-quality datasets like in the CDRC or GDELT the shown approaches would most likely not have been as successful.

As further work, other data foundations than news data could be evaluated to see if they can bring more or different insights. Examples for maybe relevant data sources could be research papers, the darknet, social media, or video news. The specific extraction of vulnerabilities, together with vulnerability databases, could also be beneficial for research as well as for cyber defence. This would allow fast insights and help experts to react timely to security threats. Using the techniques shown in this work, an alerting system for cyber defence could be developed that can warn experts when vulnerabilities arise, or patches should be installed. When it comes to a continuous adaption of the analysis model, correlation and trend analysis can be used to improve the developed annotators. Important terms that are mentioned in later added articles can be detected and added to raise precision and recall. As discussed in the previous chapter, this work only focused on statistical methods; further research can evaluate if machine learning methods can bring additional value to this area. Future research can also question the cause of the found facts. Especially the dramatical raise of the absolute numbers of articles mentioning cybercrime and terror organizations in 2019 could be further analyzed as well as the decrease of the percentage of terrorists involved in cybercrime in digital healthcare. There the question of why this happened seems from high significance, especially in a military context. In addition, a more detailed evaluation of the methodology used for the analysis can be carried out in future studies.

The text mining and investigation approach developed in this research can be also beneficial for investigating completely different domains. Every text data source can be examined using the methodology described in chapter 4.

In conclusion, this work successfully demonstrated the possibilities and benefits of text mining technologies in the area of cybersecurity in digital healthcare and proved that the stated research questions could be answered using the in subchapter 4.6 proposed methodology.

9 Acknowledgement and conflict of interests

The author of this publication has had research support from the DocCenter from the National Defence Academy (DocCenter/NDA) of the Austrian Armed Forces, the Austrian Institute of Technology (AIT) and the federal state of Lower Austria. He also consults for the DocCenter on the development of analysis tools in the area of text mining and artificial intelligence. The terms of these arrangements have been reviewed and approved by the University of Applied Sciences St. Pölten in accordance with its policy on objectivity in research.

The author would like to thank ObstdhmfD Ing. Mag. Klaus Mak and Hans Christian Pilles, ADir, RgR from the Austrian Armed Forces as well as Dr. Joachim Klerx from the Austrian Institute of Technology for the data access and research support.

References

Abid, A., Ameur, H., Mbarek, A., Rekik, A., Jamoussi, S., & Hamadou, A. B. (2017). An extraction and unification methodology for social networks data: an application to public security. *Proceedings of the 19th International Conference on Information Integration and Web-Based Applications & Services - IiWAS '17*, 176–180. https://doi.org/10.1145/3151759.3151836

Aitchison, J. (2005). Language Change. In *The Routledge Companion to Semiotics and Linguistics* (pp. 111–120).

Aktypi, A., Nurse, J. R. C., & Goldsmith, M. (2017). Unwinding Ariadne's Identity Thread: Privacy Risks with Fitness Trackers and Online Social Networks. *Proceedings of the 2017 on Multimedia Privacy and Security - MPS '17*, 1–11. https://doi.org/10.1145/3137616.3137617

Alami, S., & Elbeqqali, O. (2015). Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts. *2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, 1–5. https://doi.org/10.1109/SITA.2015.7358435

Al-Imam, A., & AbdulMajeed, B. A. (2017). The NPS Phenomenon and the Deep Web: Trends Analyses and Internet Snapshots. *Global Journal of Health Science*, *9*(11), 71. https://doi.org/10.5539/gjhs.v9n11p71

Al-Rowaily, K., Abulaish, M., Al-Hasan Haldar, N., & Al-Rubaian, M. (2015). BiSAL – A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security. *Digital Investigation*, *14*, 53–62. https://doi.org/10.1016/j.diin.2015.07.006

Apurva, A., Ranakoti, P., Yadav, S., Tomer, S., & Roy, N. R. (2017). Redefining cyber security with big data analytics. *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, 199–203. https://doi.org/10.1109/IC3TSN.2017.8284476

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, (3), 30.

Brinton, L. J. (2000). *The Structure of Modern English: A Linguistic Introduction*. John Benjamins Publishing.

Camara, C., Peris-Lopez, P., & Tapiador, J. E. (2015). Security and privacy issues in implantable medical devices: A comprehensive survey. *Journal of Biomedical Informatics*, *55*, 272–289. https://doi.org/10.1016/j.jbi.2015.04.007

Cheng, L., Liu, F., & Yao, D. (Daphne). (2017). Enterprise data breach: causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(5), e1211. https://doi.org/10.1002/widm.1211

Coventry, L., & Branley, D. (2018). Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. *Maturitas*, *113*, 48–52. https://doi.org/10.1016/j.maturitas.2018.04.008

Crone, S. F., Lessmann, S., & Stahlbock, R. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*(3), 781–800. https://doi.org/10.1016/j.ejor.2005.07.023

Eilstrup-Sangiovanni, M. (2018). Why the World Needs an International Cyberwar Convention. *Philosophy & Technology*, *31*(3), 379–407. https://doi.org/10.1007/s13347-017-0271-5

Erkal, Y., Sezgin, M., & Gunduz, S. (2015). A New Cyber Security Alert System for Twitter. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 766–770. https://doi.org/10.1109/ICMLA.2015.133

Feichtinger, W. (2017). *Feichtinger kompakt: Terrorismus - Wie groß ist die Gefahr wirklich?* Retrieved from https://www.youtube.com/watch?v=YG4FD1zEIrc

Ferrara, E., Wang, W.-Q., Varol, O., Flammini, A., & Galstyan, A. (2016). Predicting online extremism, content adopters, and interaction reciprocity. *ArXiv:1605.00659 [Physics]*, *10047*, 22–39. https://doi.org/10.1007/978-3-319-47874-6_3

Fuchslueger, J. (2016). Semantische Analyse unstrukturierter Daten. *Austrian Law Journal*, *3*, 10.

Gagneja, K. K. (2017). Knowing the ransomware and building defense against it - specific to healthcare institutes. *2017 Third International Conference on Mobile and Secure Services (MobiSecServ)*, 1–5. https://doi.org/10.1109/MOBISECSERV.2017.7886569

Gerner, D. J., Schrodt, P., Abu-Jabr, R., & Yilmaz, Ö. (2002). *Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions*. Presented at the Annual Meeting of the International Studies Association.

Göllner, J., Kienesberger, G., Peer, A., Schönbacher, P., Weiler, M., & Wurzer, G. (2010). *Wissensmanagement im ÖBH - Analyse und Betrachtung von Kritischer Infrastruktur* (Landesverteidigungsakademie). Retrieved from http://www.bundesheer.at/wissen-forschung/publikationen/publikation.php?id=737

Grover, P., Kar, A. K., & Davies, G. (2018). “Technology enabled Health” – Insights from twitter analytics with a socio-technical perspective.

International Journal of Information Management, *43*, 85–97. https://doi.org/10.1016/j.ijinfomgt.2018.07.003

Gupta, B., Sharma, S., & Chennamaneni, A. (2016). *Twitter Sentiment Analysis: An Examination of Cybersecurity Attitudes and Behavior*. 11.

Haldorai, A., & Ramu, A. (2018). The Impact of Big Data Analytics and Challenges to Cyber Security. *Handbook of Research on Network Forensics and Analysis Techniques*, 300–314. https://doi.org/10.4018/978-1-5225-4100-4.ch016

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Perez-Meana, H., Olivares-Mercado, J., & Sanchez, V. (2018). Social Sentiment Sensor in Twitter for Predicting Cyber-Attacks Using ℓ1 Regularization. *Sensors (Basel, Switzerland)*, *18*(5). https://doi.org/10.3390/s18051380

IBM Security (Series Ed.). (2017). *IBM X-Force Exchange - Data Sheet*. Retrieved from https://www.ibm.com/downloads/cas/5LLKMXK3

Indulska, M., Hovorka, D. S., & Recker, J. (2012). Quantitative approaches to content analysis: identifying conceptual drift across publication outlets. *European Journal of Information Systems*, *21*(1), 49–69. https://doi.org/10.1057/ejis.2011.37

Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., … Li, H. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review*, *30*(2), 157–170. https://doi.org/10.1017/S0269888914000277

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*.

Kruse, C. S., Frederick, B., Jacobson, T., & Monticone, D. K. (2017). Cybersecurity in healthcare: A systematic review of modern threats and trends. *Technology and Health Care*, *25*(1), 1–10. https://doi.org/10.3233/THC-161263

Leetaru, K., & Schrodt, P. A. (2013). GDELT: Global data on events, location, and tone. *ISA Annual Convention*.

Lu, H., Li, Y., Chen, M., Kim, H., & Serikawa, S. (2018). Brain Intelligence: Go beyond Artificial Intelligence. *Mobile Networks and Applications*, *23*(2), 368–375. https://doi.org/10.1007/s11036-017-0932-8

Luger, G. F. (2009). *Artificial intelligence: structures and strategies for complex problem solving* (6th ed). Boston: Pearson Addison-Wesley.

Mak, K., Klerx, J., Pilles, H. C., & Göllner, J. (2015). *Wissensentwicklung mit „Crowd OSInfo“*. 80.

Mak, K., Pilles, H. C., Bertl, M., & Klerx, J. (2018). *Wissensentwicklung mit IBM Watson in der Zentraldokumentation (ZentDok) der Landesverteidigungsakademie*. Retrieved from http://www.bundesheer.at/wissen-forschung/publikationen/publikation.php?id=897

Marsh & McLennan Companies. (2017). *MMC Cyber Handbook 2018*.

Martin, F., & Johnson, M. (2015). More Efficient Topic Modelling Through a Noun Only Approach. *Proceedings of the Australasian Language Technology Association Workshop 2015*, 111–115. Retrieved from http://www.aclweb.org/anthology/U15-1013

Martínez-Pérez, B., de la Torre-Díez, I., & López-Coronado, M. (2015). Privacy and security in mobile health apps: a review and recommendations. *Journal of Medical Systems*, *39*(1), 181. https://doi.org/10.1007/s10916-014-0181-3

Meier, P. (2012). Ushahidi as a Liberation Technology. In *Liberation Technology - Social Media and the struggle for Democracy* (pp. 95–109).

Mertz, L. (2018). Cyber-Attacks to Devices Threaten Data, Patients. *IEEE Pulse*. Retrieved from https://pulse.embs.org/january-2018/cyber-attacks-devices-threaten-data-patients/

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, … Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Müller, O., Junglas, I., & vom Brocke, J. (2016). Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, *39*, 110–135. https://doi.org/10.17705/1CAIS.03907

O'Doherty, K. C., Christofides, E., Yen, J., Bentzen, H. B., Burke, W., Hallowell, N., … Willison, D. J. (2016). If you build it, they will come: unintended future uses of organised health data collections. *BMC Medical Ethics*, *17*(1). https://doi.org/10.1186/s12910-016-0137-x

Ponemon Institute. (2017). *Ponemon Institute's 2017 Cost of Data Breach Study: Global Overview*. Retrieved from https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=SEL03130WWEN

Prakash, B. A. (2016). Prediction Using Propagation: From Flu Trends to Cybersecurity. *IEEE Intelligent Systems*, *31*(1), 84–88. https://doi.org/10.1109/MIS.2016.1

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Raatikainen, M. J. P., Arnar, D. O., Zeppenfeld, K., Merino, J. L., Levya, F., Hindriks, G., & Kuck, K.-H. (2015). Statistics on the use of cardiac electronic devices and electrophysiological procedures in the European Society of Cardiology countries: 2014 report from the European Heart Rhythm Association. *Europace: European Pacing, Arrhythmias, and Cardiac Electrophysiology: Journal of the Working Groups on Cardiac Pacing, Arrhythmias, and Cardiac Cellular Electrophysiology of the European Society of Cardiology*, *17 Suppl 1*, i1-75. https://doi.org/10.1093/europace/euu300

Saltzman, I. (2013). Cyber Posturing and the Offense-Defense Balance. *Contemporary Security Policy*, *34*(1), 40–63. https://doi.org/10.1080/13523260.2013.771031

Sanger, D. E., & Broad, W. J. (2017, March 4). Trump Inherits a Secret Cyberwar Against North Korean Missiles. *The New York Times*. Retrieved from https://www.nytimes.com/2017/03/04/world/asia/north-korea-missile-program-sabotage.html

Shafqat, S., Kishwer, S., Rasool, R. U., Qadir, J., Amjad, T., & Ahmad, H. F. (2018). Big data analytics enhanced healthcare systems: a review. *The Journal of Supercomputing*. https://doi.org/10.1007/s11227-017-2222-4

Sulleyman, A. (2017, May 12). NHS Cyber Attack: Why stolen medical information is so much more valuable than financial data. *The Independent*. Retrieved from http://www.independent.co.uk/life-style/gadgets-and-tech/news/nhs-cyber-attack-medical-data-records-stolen-why-so-valuable-to-sell-financial-a7733171.html

Thapen, N., Simmie, D., & Hankin, C. (2016). The early bird catches the term: combining twitter and news data for event detection and situational awareness. *Journal of Biomedical Semantics*, *7*(1). https://doi.org/10.1186/s13326-016-0103-z

The GDELT Project. (n.d.). Retrieved 12 March 2019, from The GDELT Project website: https://www.gdeltproject.org/

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Zhu, W.-D., Foyle, B., Gagné, D., Gupta, V., Magdalen, J., Mundi, A. S., … Triska, M. (2014). *IBM Watson Content Analytics: Discovering Actionable Insight from Your Content* (3rd ed.).

Zubiaga, A., Procter, R., & Maple, C. (2018). A longitudinal analysis of the public perception of the opportunities and challenges of the Internet of Things. *PLoS ONE*, *13*(12). https://doi.org/10.1371/journal.pone.0209472

List of Figures

List of Tables