Young Information Scientist 4 (2019), 1–15
DOI:10.25365/yis-2019-4-1

News analysis for the detection of cyber security issues in digital healthcare

A text mining approach to uncover actors, attack methods and technologies for cyber defense

Markus Bertl, MSc.*

Objectives — This research reviews the possibilities of text mining in the area of cybercrime in digital healthcare showing how advanced information retrieval and natural language processing can be used to get insights. The aim is 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.
Methods — 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 or rely on a comparably large datset as this study. 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 Defense Academy (DocCenter/NDA) of the Austrian Armed Forces. About 300,000 news articles were processed and analyzed. Additionally, the open source GDELT dataset was investigated.
Results & Conclusion — Text mining is an important tool for cybersecurity and trend research. The data points out that cyber threats are present in digital health technologies and cyberattacks are more and more threatening to organizations, governments, and individuals. Not only hacker groups, firms, and governments are involved in these attacks, also terroristic organizations use cyber warfare. That, together with the amount of technology in healthcare like pacemakers, IoT, wearables but also the importance of healthcare as critical infrastructure and the growing dependence on electronic health records makes our society vulnerable.

Keywords — digital healthcare, cybercrime, text mining, media mining, new technologies, Watson Explorer, GDELT, OSInfo, OSINT, association rule mining

Nachrichtenanalyse zur Erkennung von Cybersicherheitsproblemen im digitalen Gesundheitswesen: ein Text-Mining-Ansatz zur Aufdeckung von Akteuren, Angriffsmethoden und Technologien für die Cyber-Abwehr

Zielsetzung — Diese Publikation untersucht die Möglichkeiten, welche Text Mining im Bereich Cybercrime in Digital Healthcare bietet. Die Zielsetzung dieser Arbeit ist, herauszufinden was über Digital Healthcare berichtet wird, welche Akteure in dieser Domäne agieren und welche Angriffsmethoden und Technologien eine Rolle spielen.
Forschungsmethoden — Verschiedene Projekte verwenden Text Mining erfolgreich im Cyber-Bereich, allerdings nicht spezifisch adaptiert auf die Anforderungen des Gesundheitswesens. Dazu wurden verschiedene Text-Mining-Methoden, wie Fact Extraction oder semantische Felder, sowie statistische Methoden wie Korrelationen oder Trendanalysen angewendet. Die Datengrundlage kam aus der Zentraldokumentation der Landesverteidigungsakademie (ZentDok/LVAk) des österreichi-

* Markus Bertl, MSc. | dh171823@fhstp.ac.at | ORCID: 0000-0003-0644-8095

Young Information Scientist (YIS) wird vom Verein zur Förderung der Informationswissenschaft (VFI), Wien, herausgegeben. Alle Beiträge unterliegen einem Peer Review. ISSN: 2518-6892

schen Bundesheeres. Insgesamt wurden zirka 300.000 Artikel ausgewertet. Zusätzlich wurden die Metadaten des GDELT Datasets untersucht.
Ergebnisse & Schlussfolgerung — Text Mining ist ein zentrales Werkzeug für Cybersecurity und Trendanalysen. Die Daten zeigen, dass Technologien im Bereich Digital Healthcare stetig 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 Terrorangriffen. 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.

Schlagwörter — Digital Healthcare, Cybercrime, Text Mining, Media Mining, neue Technologien, Watson Explorer, GDELT, OSInfo, OSINT, Assoziationsanalyse

Diesem Beitrag liegt folgende Abschlussarbeit zugrunde / This article is based upon the following thesis:
Bertl, Markus: Cyberthreats in Digital Healthcare — An exploratory analysis using text mining on news data. Masterarbeit (MSc), St. Pölten University of Applied Sciences, 2019.

1 Introduction

IT-based healthcare technology is on the rise, buzzwords like e-Health, electronic health record (EHR) or telemedicine are increasingly used in the literature. Governments begin to build systems in order to save, connect, share, and analyze 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 et al. 2017; Zubiaga et al. 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 et al. 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 targets of cyberattacks in the past. These examples show how vital defense 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. In order to detect quickly what kind of cyber threats emerge in the healthcare industry, a text mining system based on news data is proposed in this research.

2 Research Questions

The overall research question is: Which additional value is achieved by text mining news data regarding the analysis of cyber threats in healthcare?

In detail, it shall be analyzed what kind of news is broadcasted about digital healthcare to identify

new topics and trends, security incidents related to healthcare and what attack methods are used.

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

- What are the primary topics 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 and topics arise in digital healthcare?

3 Methodology

A systematic literature review according to Kitchenham and 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 cyber threats, and important text mining approaches in general.

Based on the above-mentioned research questions, facts from data sources described in the next chapter 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 (Fuchslueger 2016). This approach uncovers security threats, the moment when they were first discovered and exploited by criminals together with the caused damage. The IBM Watson Explorer Analytical Components v12 was used for the data analysis in this research. A similar approach based on information retrieval and natural language processing has been applied by Mak, Pilles et al. (2018).

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).

4 Underlying data

Documents from 01/01/2015 until 31/03/2019 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, which is publicly available, the Austrian Armed Forces as part of their Cyber Documentation & Research Center project provided all data.

4.1 Cyber Documentation & Research Center

The Cyber Documentation & Research Center, short CDRC, was a project at the DocCenter from the National Defense 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’ (KriMiSi), and ‘innovation and technology’ (InnoTech). 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 database has been designed especially for the purpose of storing data about news, incidents or disasters (Meier 2012). Because of this support of the Crowd OSInfo approach, it is an ideal foundation for storing data in the CDRC. More information on the data collection in the Cyber Documentation & Research Center can be found at Mak, Klerx et al. (2015).

4.2 Global Database of Events, Language and Tone

The Global Database of Events, Language and Tone (GDELT) project collects data about the world's broadcast, print, and web news since 1979 until now. It covers nearly every country in the world in more than 100 languages. All articles are translated into English automatically. 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.' (*The GDELT Project* 2019)

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 Mediation Event Observations and is a coding standard for political news and violence (Gerner et al. 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* (2019) and at Leetaru and Schrodt (2013).

5 Findings

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

Because different keyword searches were not precise enough, semantic fields were used in the second attempt to get increase precision. 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'. For example, the semantic field for 'Healthcare' contains words like 'medical', 'hospital', 'ill', 'cure' or 'treatment'. Inflections, synonyms and abbreviations were added to them to increase accuracy. 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').

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.

Since the German language has many compound words, 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.

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 feature of the Watson Explorer. The part of speech analysis automatically extracts not only grammatical information as nouns, verbs or numerals but also phrases like noun sequences or verb-noun combinations from text. A periodical check what nouns, verbs or phrases have a high correlation to healthcare, cybercrime or ICT reveals new terms in the area. The semantic field can then be updated with these terms.

After the relevant documents could be identified, the different datasets were explored using the metadata, annotations of the semantic fields and the different views that are offered by the IBM Watson Explorer Content Analytics Miner.

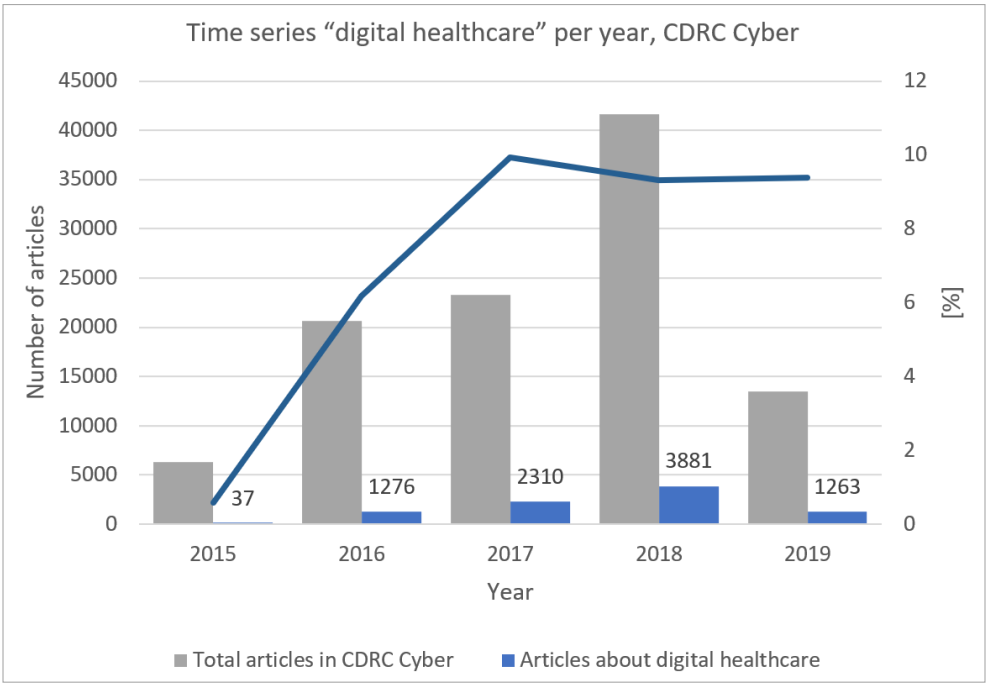


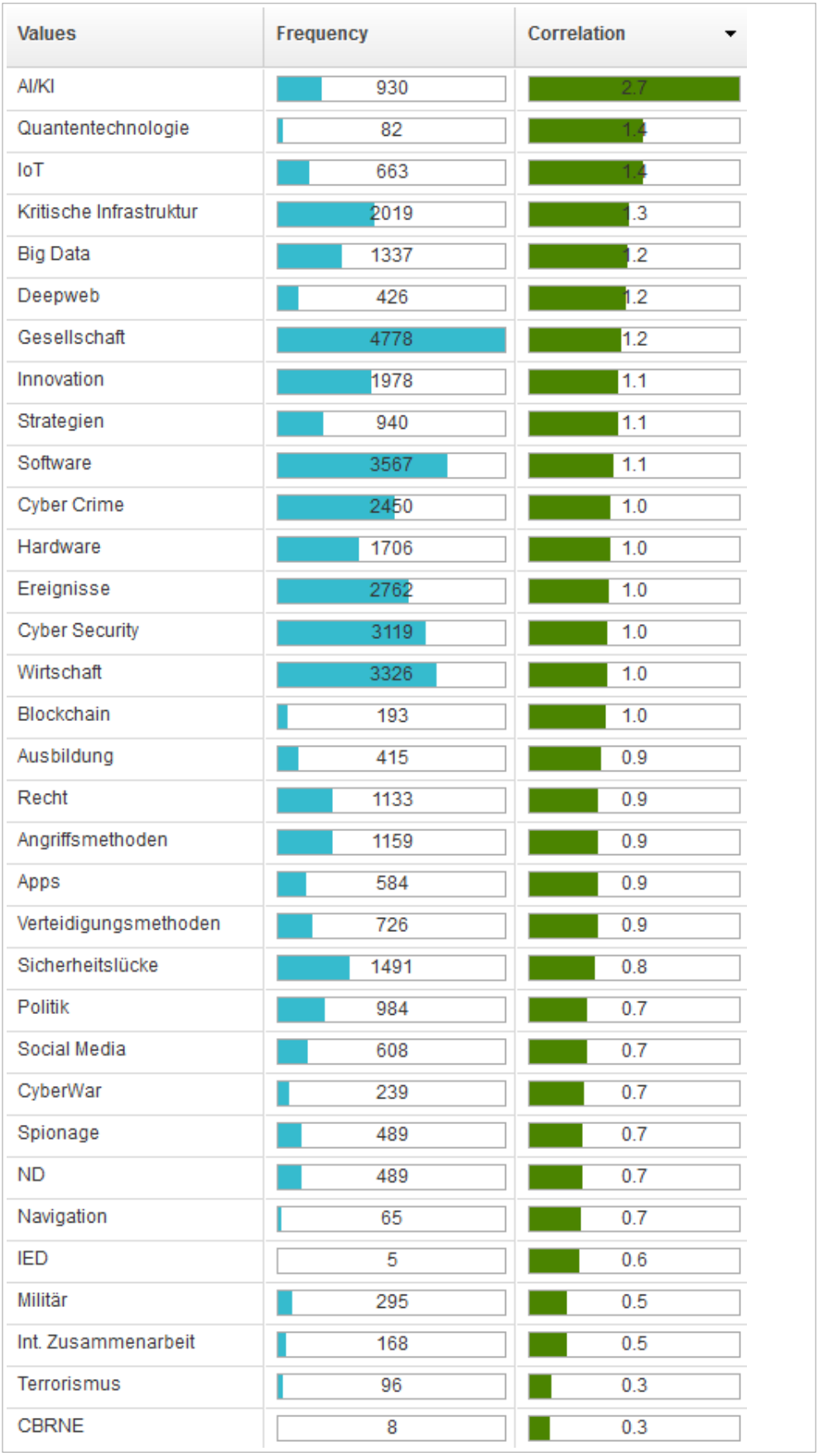
Figure 1: Time series ‘digital healthcare’ per year, CDRC Cyber.

A time series analysis across all datasets with only the relevant documents (documents about digital healthcare) showed that news about cybercrime in healthcare has increased steadily over the last years (figure 1) indicating the rising importance of digital healthcare. In the last three years, articles about digital healthcare accounted for about 10 % of the whole news data investigated.

To extract important topics in the area of digital healthcare, the metadata tags have been analyzed. According to figure 2, the facet values artificial intelligence (‘AI/KI’), quantum technology (‘Quantentechnolgie’), Internet of Things (‘IoT’) or critical infrastructure (‘Kritische Infrastruktur’) have 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 cybercrime, cybersecurity, defense methods (‘Verteidigungsmethoden’), intelligence (‘Spionage’), vulnerabilities (‘Sicherheitslücken’), cyberwar and terrorism (‘Terrorismus’) were mentioned. The high frequency of these terms also supports this relevance.

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 2 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, pp. 16 sq.). 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 3, ‘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.

An extraction of the cyber threats was conducted using the intersection of documents dealing with digital healthcare, actors in the cyber domain and hacking methods. The actors (hacker, hacker



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 2: Topics of digital healthcare, Cyber.

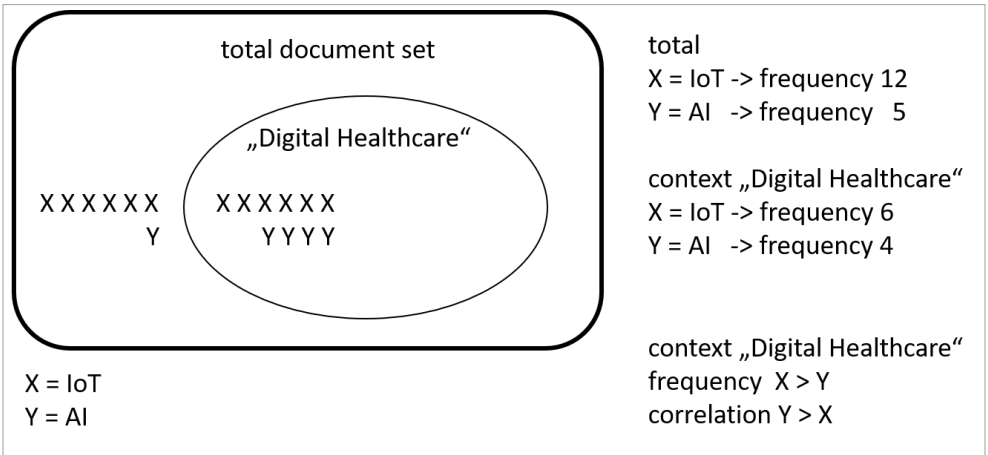


Figure 3: Frequency vs. Correlation example.

groups, and terror groups) and hacking methods were modelled as dictionary using already available lists compiled by the Cyber Documentation & Research Center as well as manual research to improve accuracy.

A time series view of documents containing something about digital healthcare and actors or attack methods gave insights into when cybersecurity relevant events were reported. This is demonstrated in figure 4.

The near-duplicate detection of the Watson Explorer makes it also possible to hide more mentions of the same event.

A comparison of figure 1 with figure 4 shows that the number of news articles about digital healthcare as well as the number of news articles 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 5. 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 terror groups is because they are mentioned outside the scope of digital healthcare more frequently in the KriMiSi dataset. However, the high frequency still indicates that terrorists are involved in cybercrime in digital healthcare.

The number of terrorists involved in cybercrime in digital healthcare was put into relation to the number of total cyberattacks mentioned in the KriMiSi dataset in table 1 and the Cyber dataset in table 2.

The relative numbers of terrorists involved in hacking events in digital healthcare even decreased in KriMiSi and stayed approximately constant at the Cyber dataset. The high percentage of terrorist involvement in table 1 is because KriMiSi focuses in general more on terrorist activities than the other datasets. However, the data clearly points out that terrorists are involved in cybercrime in digital healthcare.

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

Figure 6 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.

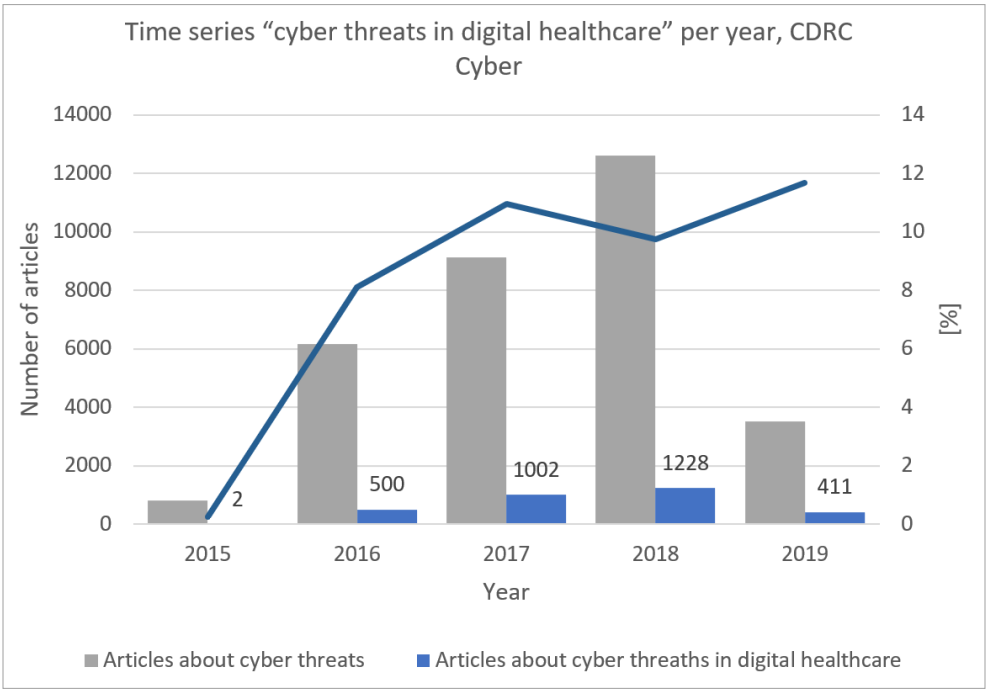


Figure 4: Time series ‘cyber threats in digital healthcare’ per year, CDRC Cyber.

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

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 highly 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

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 5: Actors in digital healthcare, KriMiSi.

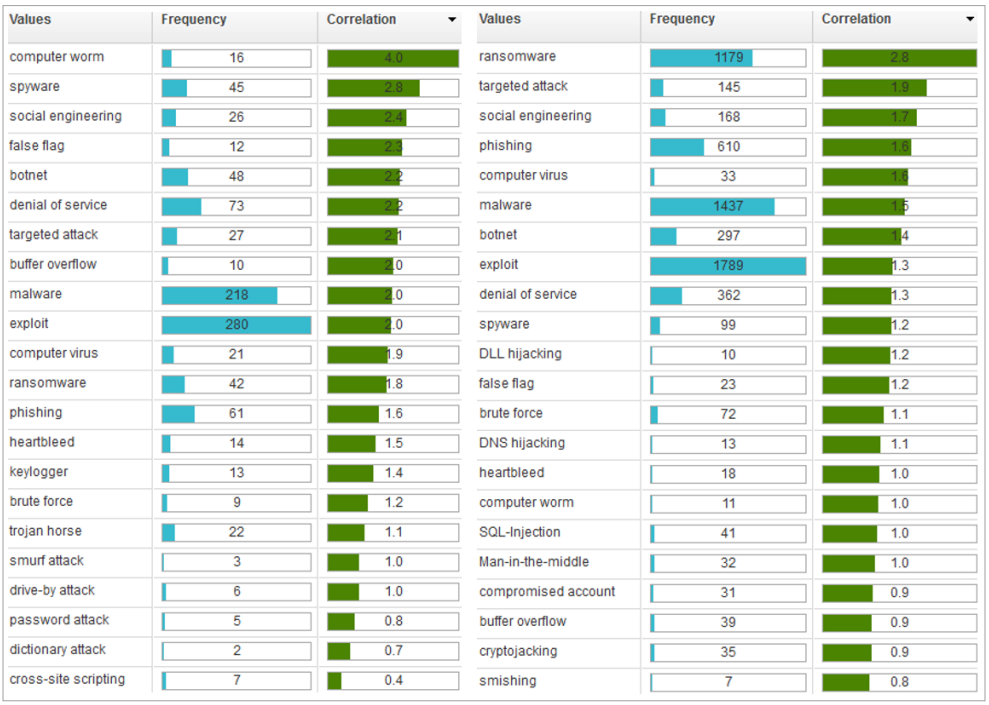
are the most dominant attack methods in the unfiltered datasets making them less noticeable in the context of digital healthcare.

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.

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



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	1.6
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 6: Attack methods that correlate with digital healthcare, left KriMiSi, right Cyber.

it in general terms the following formulas can be derived:

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

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

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. Further information on this can be found at Zhu et al. (2014, pp. 192–193).

As pictured in figure 7, 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.

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 trends 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. The trend view can give insights into unexpected changes in frequency or correlation values of facet entries. Figure 8 shows the noun sequence facet of all documents mentioning terms from digital healthcare in the trend analysis.

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 7: Facet pair view with attack methods and actors, Cyber.

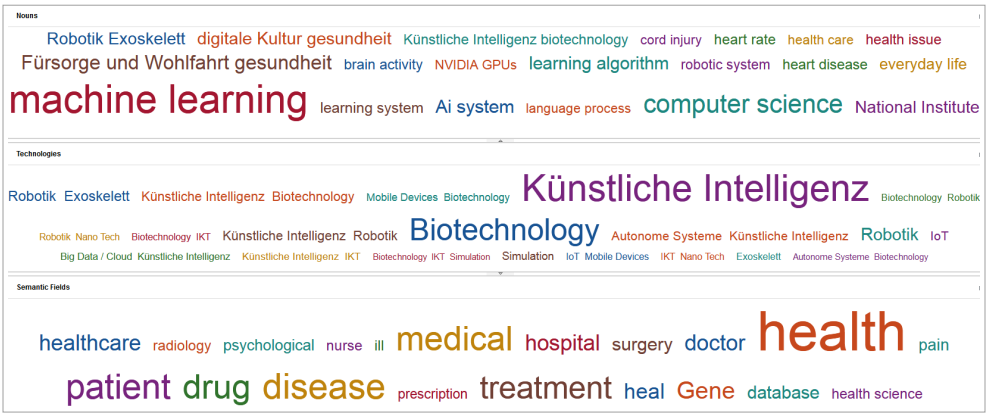


Figure 8: Noun Trends digital healthcare, InnoTech.

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 Discussion

The results presented underline again that text mining is an important tool for cybersecurity and trend research. 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. If the model, in this case the semantic fields, is not accurate, the whole analysis can produce misleading results. Also the cross-check of the outcomes of the analysis is vital. This is a task which is difficult to automate at the moment. For now, 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 et al. 2018). Because of that, humans need to perform quality control of the system's output themselves.

The already argued quality control was done by manually evaluating the articles behind the statistics to see if the correlations and results are plausible. 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. 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.

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.

This research only analyzes the publication dates of the articles in the data sets. An extraction of the event dates was not conducted. It might thereby be plausible that the results could be lagged as long as the publication date is used as a proxy for the event date. However, as both dates usually do not differ too significantly, the respective effect is expected to be small. However, it should be kept in mind that not all security critical events are published.

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. 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 but what training data is responsible for which result cannot be traced. In addition, 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 spurious 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 Defense 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 has found their way into the CDRC datasets, the analysis could be biased because of that.

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.

The results of the research seem in general plausible. That digital healthcare is mentioned more and more was expected. Also 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 target 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 specially for digital healthcare. They also revealed that the targets are very diverse reaching 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 during the whole lifecycle.

7 Conclusion

According to the findings, all stated research questions could be answered.

An overview of the topics of digital healthcare can be seen in figure 2 but also in the word clouds pictured in figure 8. 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 from a medical point of view. The topic analysis also showed first indications of cybercrime in digital healthcare.

The events were minded and can be derived from the time series view pictured in figure 4. An increase in cyber events in the domain of digital healthcare was observed over the last years.

Using facet and correlation analysis, the actors described in figure 5 were extracted. Additionally to the traditional hacking groups, also terrorist cells and networks involved in cybercrime in digital healthcare were found. The percentage of involvement of terrorists in cybercrime in digital healthcare seems to be declining.

Attack methods in combination with digital healthcare were listed as shown in figure 6. The facet pair analysis in figure 7 revealed what attackers used which attack method.

New technologies were found with the trend view pictured in figure 8.

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 have increased steadily over the last years. That underlines the importance of the topic. The statements found in the literature also back up the found results. Cybercrime in digital healthcare is an increasing threat for society, companies, and the individual. The proposed methodology and tools were capable of answering 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.

The text mining and investigation approach developed in this research can be also beneficial for investigating completely different domains.

In conclusion, this work demonstrated successfully 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 proposed methodology.

References

Aitchison, J. (2005). Language change. In *The routledge companion to semiotics and linguistics*. Ed. by Cobley, P. London: Routledge, 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. In *Proceedings of the 2017 on multimedia privacy and security – MPS '17* (Dallas, TX, 30th October 2017). New York, NY, USA: ACM, pp. 1–11. DOI: 10.1145/3137616.3137617.

Brinton, L. J. (2000). *The structure of modern english: A linguistic introduction*. John Benjamins.

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

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

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*. Paper prepared for delivery at the Annual Meeting of the International Studies Association.

Indulska, M.; Hovorka, D. S.; Recker, J. (2012). Quantitative approaches to content analysis: Identifying conceptual drift across publication outlets. In *European Journal of Information Systems* 21 (1), pp. 49–69. DOI: 10.1057/ejis.2011.37.

Kitchenham, B.; Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.471 visited on 30th September 2019.

Leetaru, K.; Schrodt, P. A. (2013). *GDELT: Global data on events, location, and tone*. Paper presented at the ISA annual convention.

Lu, H.; Li, Y.; Chen, M. et al. (2018). Brain intelligence: Go beyond artificial intelligence. In *Mobile Networks and Applications* 23 (2), pp. 368–375. DOI: 10.1007/s11036-017-0932-8.

Mak, K.; Klerx, J.; Pilles, H. C.; Göllner, J. (2015). *Wissensentwicklung mit „Crowd OSInfo"*. Schriftenreihe der Landesverteidigungsakademie 80. Wien: BM für Landesverteidigung und Sport.

Mak, K.; Pilles, H. C.; Bertl, M.; Klerx, J. (2018). *Wissensentwicklung mit IBM Watson in der Zentraldokumentation (ZentDok) der Landesverteidigungsakademie*. Schriftenreihe der Landesverteidigungsakademie. Wien: BM für Landesverteidigung und Sport.

Marsh & McLennan Companies (2017). *MMC cyber handbook 2018*.

Meier, P. (2012). Ushahidi as a liberation technology. In *Liberation technology: Social media and the struggle for democracy*. Ed. by Diamond, L.; Plattner, M. F. Baltimore: The Johns Hopkins University Press, pp. 95–109.

Mertz, L. (2018). Cyber-attacks to devices threaten data and patients: Cybersecurity risks come with the territory. Three experts explain what you need to know. In *IEEE Pulse*. DOI: 10.1109/MPUL.2018.2814258.

Michel, J.-B.; Shen, Y. K.; Aiden, A. P. et al. (2011). Quantitative analysis of culture using millions of digitized books. In *Science* 331 (6014), pp. 176–182. DOI: 10.1126/science.1199644.

Müller, O.; Junglas, I.; Brocke, J. vom (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. In *Communications of the Association for Information Systems* 39, pp. 110–135. DOI: 10.17705/1CAIS.03907.

Ponemon Institute (2017). *2017 cost of data breach study: Global overview*. https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=SEL03130WWEN visited on 5th January 2019.

Raatikainen, M. J. P.; Arnar, D. O.; Zeppenfeld, K. et al. (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. In *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, pp. i1–75. DOI: 10.1093/europace/euu300.

The GDELT Project (2019). https://www.gdeltproject.org/ visited on 12th March 2019.

Zhu, W.-D.; Foyle, B.; Gagné, D. et al. (2014). *IBM Watson content analytics: Discovering actionable insight from your content*. 3rd ed. IBM Redbook. IBM International technical support organization. http://www.redbooks.ibm.com/redbooks/pdfs/sg247877.pdf visited on 30th September 2019.

Zubiaga, A.; Procter, R.; Maple, C. (2018). A longitudinal analysis of the public perception of the opportunities and challenges of the internet of things. In *PLoS ONE* 13 (12). DOI: 10.1371/journal.pone.0209472.

Eingereicht: 27.06.2019; revidiert: 09.08.2019; akzeptiert: 12.08.2019; publiziert: 04.10.2019