Evolution in the Research Pattern of Healthcare IT Security & Privacy

NKU | subramanis1@nku.edu

MBI 692 – RESEARCH PROJECT RePORT

Selva damodaran

2021

Contents

[Introduction 2](#_Toc71309739)

[Background 2](#_Toc71309740)

[What is Healthcare Privacy & Security? 2](#_Toc71309741)

[HIPAA and PHI 3](#_Toc71309742)

[Current State of Data Security & Privacy in Healthcare IT 3](#_Toc71309743)

[Researches surrounding the Privacy of Healthcare Data 5](#_Toc71309744)

[Related Works 5](#_Toc71309745)

[Methodology and Results 6](#_Toc71309746)

[Loading Data 6](#_Toc71309747)

[Exploratory Data Analysis 7](#_Toc71309748)

[LDA Model Development and Training 10](#_Toc71309749)

[Validation of LDA Model Results 12](#_Toc71309750)

[Conclusion 17](#_Toc71309751)

[References 18](#_Toc71309752)

# Introduction

Healthcare information systems are largely viewed as the single most important factor in improving US healthcare quality and reducing related costs. According to the Office of the National Coordinator for Health Information Technology, 95% of hospitals across the US have adopted certified EHR technology. Anecdotal evidence from recent years suggest that a lack of adequate security measures has resulted in numerous data breaches, leaving patients exposed to economic threats, mental anguish and possible social stigma (Health Privacy Project, 2007). A recent survey in the USA suggests that 75% of patients are concerned about health websites sharing information without their permission (2). Possibly, this patient perception is fueled by the fact that medical data disclosures are the second highest reported breach (3). In response to these increasing threats to health information and privacy, new regulations at both the state and the federal level have been proposed in the USA, e.g., Health Insurance Portability and Accountability Act (HIPAA).

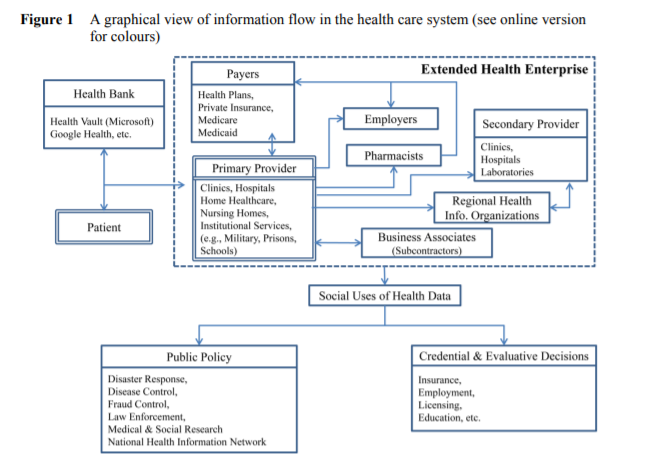
Over the past two decades, information security research has become a well-established area within the information systems discipline. Researchers have adopted several underlying theories from reference disciplines such as psychology and sociology to analyze information security risk management (4, 5, 6) and economic theories to characterize investment decisions and information governance (7, 8). Since then, there were several studies conducted in the preserving the security and privacy of patient data shared in Healthcare IT systems. In this paper, our aim is to study the trend of research in Healthcare IT Security and decipher a pattern that the EHR Privacy and Security has taken through these years. We first aim to explain the data / information flow in these EHR landscapes, then address the hole which creates the risk of security & privacy violations, and finally build a classification model to detect the pattern of researches conducted in EHR security.

# Background

## What is Healthcare Privacy & Security?

In patient physician relationship, the privacy is viewed as one of the key principles that governs it. In general, to facilitate appropriate treatment & diagnosis and to counter adverse drug reactions, patients are required to share information to their physicians. However, situations where patients experience health problems like psychiatric behavior, HIV etc. may reject to divulge important information’s as it might lead to social stigma and discrimination (10). Over a period of time, a patient’s medical file could accrue significant PII (Personally Identifiable Information), and many other information such as history related to medical diagnosis, medical images, past treatments, psychological summaries, sexual choices, dietary behaviors, income and physicians subjective valuations of personality and mental state (9).

Figure 1 illustrates a graphical view of information flow in healthcare sector. Apart from diagnosis and treatment provision a patient’s health record serves a variety of purpose in hospital. To name few, it will be used to improve healthcare system efficiency, development of public policy and administration, and in oversight of medical research (12). In order to justify the rendered payment services, patient’s medical records are also shared with payer company. (e.g., private insurance or Medicare/ Medicaid) and it is also used to improve their operations and service quality across different verticals of healthcare systems.



## HIPAA and PHI

One important law that is crucial in healthcare is HIPAA which emphasis the importance of protecting Valuable health information from the criminals and hackers and therefor it is always believed to be complicated and overwhelmed law that health care companies have wriggled to make sure they are fully complaint with.

Any information within a person’s medical record that are used to identify them and if it is held by a covered entity it is known as Protected Health Information or PHI. There are 18 specific identifiers are used under HIPAA’s privacy rule to handle with strict safeguards. Many PHI information breaches happen through hacking or any IT incidents. Root cause for these incidents are because of its high value in off market and these are used to facilitate illegal activities on dark web. (e.g., Create fake medical claims, purchase prescriptions etc.) So, it is extremely important to keep them protected with any breach of PHI.

## Current State of Data Security & Privacy in Healthcare IT

Firstly, studies and research conducted on the healthcare consumers that include health record management and online EHR system have created number of security-related questions which include the primary concern on healthcare privacy and security for consumers, impact of medical identity theft on consumers well-being. Secondly, issues related to providers, like the handlers of IT adoption, effect of IT on medical errors, telemedicine, pervasive computing. (e.g., creation of access control systems, network security, privacy policy and risk management). Lastly, many information security and privacy directions have surfaced in the public privacy, particularly in the areas like, medical study, national information network, health service pricing

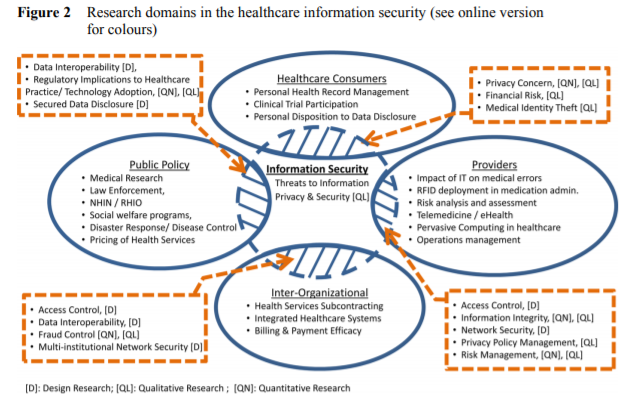
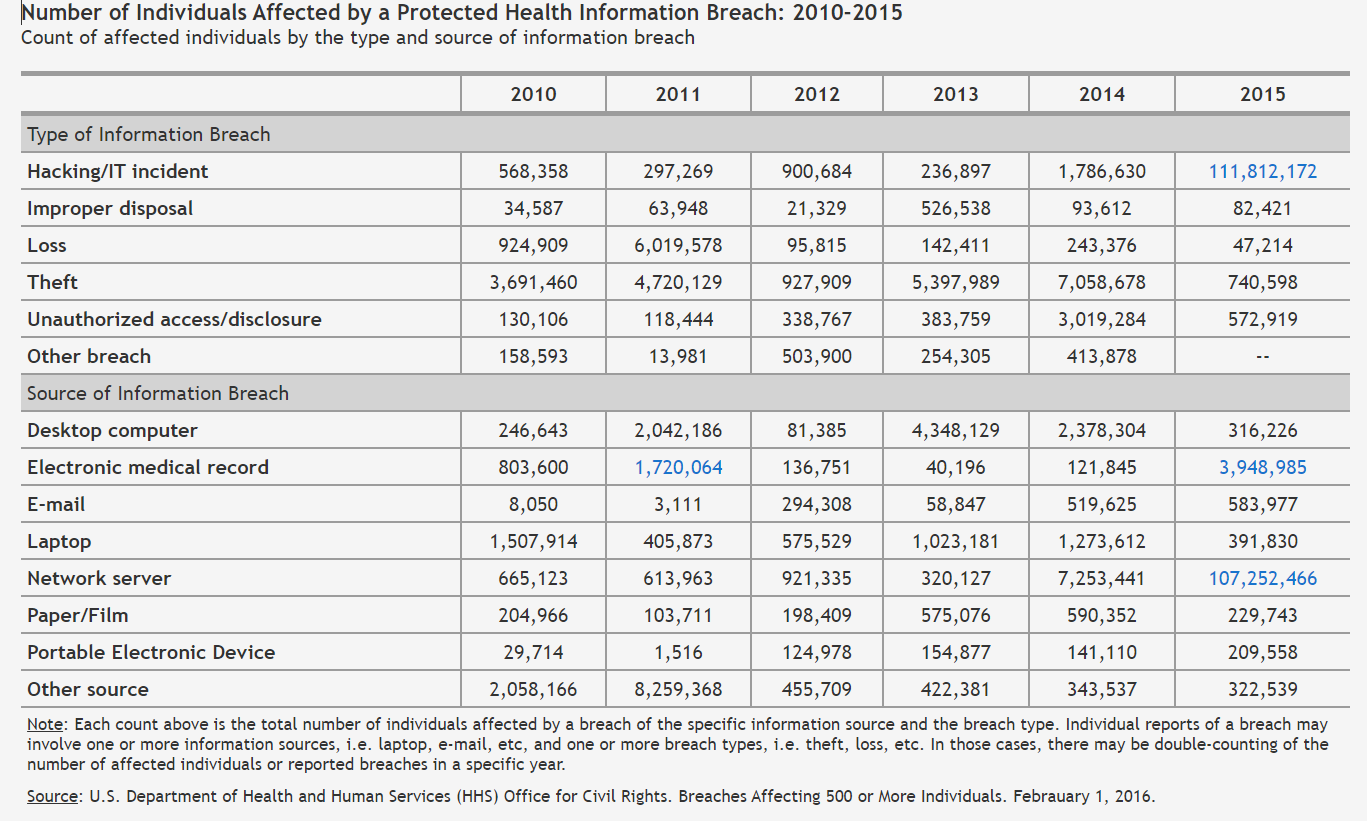


Figure below shows the number of PHI data attacks encountered between 2010 and 2015.



## Researches surrounding the Privacy of Healthcare Data

Several organizations have already put in place technological solutions for maintaining and managing their patient’s privacy over wired and wireless networks. Numerous research studies have been conducted in the area of information protection of patients’ PHI data. Majority of the design research focuses on developing artefacts such as models, algorithms, samples, and frameworks to resolve specific issues related to data protection and governance in EHR systems (11). Studies have also been performed aiming at a qualitative research, which involves examining a social experience using a span of qualitative instruments/data such as interviews, participants’ observation data, researcher’s impression (12). In healthcare research, much of the qualitative research revolves around the impact of HIPAA on healthcare practices (13). Lastly, researchers in healthcare information systems have adopted several quantitative methods including surveys, statistical modelling in the areas of patients’ privacy concern, public policy and security, fraud detection and control, risk management and its impact of health IT on medical errors (14). This paper focusses on identifying the various topics of research in the field of healthcare data security and privacy, studying the evolution of its research pattern over the last decade and predicting the trend of researches that will be performed in this field.

# Related Works

Several works of literature review have been performed surrounding the healthcare privacy concerns, malicious attacks and implementation of security providing systems to combat these data breaches. Ahmadi et al., reviewed and analyzed 60 research papers published between 2000 and 2016 regarding application of IoT in healthcare security. Their study employed a systematic review approach on the data that was collected (15). A similar study was conducted by Nazir et al., where again a systematic literature review of papers was performed. Researches published related to Mobile Computing in IoT for Healthcare were reviewed in their analysis (18). Mehraeen et al., performed a full text review of 210 articles published about security challenges in healthcare cloud computing technology and listed the information protection systems that are widely adopted to mitigate security incidents on healthcare applications (16). Ermakova et al., aimed to identify the state of research and determine potential areas of future work in the context of healthcare cloud computing (22). Ahmed et al., performed a question-based formalization to spot the major concerns in healthcare security, and had done a detailed review analysis of papers collected for their constructed search terms (17). Yao et al., studied the adoption and implementation of RFID in Healthcare through a systematic literature review of researches on RFID use in healthcare. They have followed a formal innovation-decision framework in their study (19). Wamba et al., highlighted the increased risk of RFID applications in healthcare and the potential issues that it could bring to the healthcare privacy (20). They developed a conceptual framework to perform an extant literature review in their study. Box and Pottas studied the gap between intent to use IT and actual compliance in Healthcare patient privacy by adopting an Appraisal Tendency Framework to review collected literature (21). Tandon et al., explored the application of blockchain in healthcare and then profiled & organized the intellectual capital in that domain. And integrated framework was developed during their research to illustrate thematic areas of future research (23). Despite numerous publications in the context of healthcare security and privacy there is no automated review of papers to decipher the current state of research so far in the domain. This paper aims to address this gap through a thematic analysis and identify the evolution pattern of research topics over the last decade in the healthcare privacy domain.

# Methodology and Results

NKUs Steely Library was considered as the source of information for articles related to healthcare security and privacy. Three keyword combination strings were used for this database search – Healthcare Security, Healthcare Privacy, Health IT. Papers searched were the ones published between Jan 2008 and March 2021. Results obtained for these terms were for the search performed in March 2021. These keywords were drawn from a review of prior studies (i.e. SLRs) in this field that used similar keywords, i.e. healthcare (or health\*), and IoT (or Internet of Things\*), and Data Security (or Data Privacy\*), etc. (19, 22, 23)

Table 1: Database Search Summary

|  |  |  |  |
| --- | --- | --- | --- |
| **Journal Database** | **Total hits appeared** | **Abstracts read** | **Full text downloaded** |
| **JISS** (Journal of Information Security Systems) | 18 | 16 | 16 |
| **IJIS** (International Journal of Information security) | 20 | 20 | 20 |
| Information Security Technology Report | 2 | 2 | 2 |
| Information & Security | 5 | 5 | 5 |
| **EURASIP JIS** (Journal of Information Security) | 6 | 5 | 5 |
| **IJISS** (International Journal of Information Security & Systems) | 5 | 4 | 4 |
| **IJCNIS** (International Journal of Computer Networks & Information security) | 176 | 165 | 162 |
| **JISIS** (Journal of Internet Services & Information Security) | 198 | 189 | 176 |
| **IJINS** (International Journal of Information and Network Security) | 1 | 1 | 1 |
| **JISA** (Journal of Information Security Applications) | 26 | 26 | 24 |

Note: Results include articles from multiple disciplines such as medicine, genomics, information science, banking etc. Multiple sources and document types were reflected in the search results including journals, trade magazines, books etc. Some papers appeared in the result list more than once and these were considered only once for further analysis.

Results of the search were sorted for “relevance” before performing the Topic Modelling. The topic mining of Research Paper titles was done through the implementation of LDA algorithm. As part of this LDA model development and validation, following steps were performed.

1. Data Cleaning, Data Preparation and Data Loading for analysis
2. Exploratory Data Analysis
3. LDA Model Development and Training
4. Validation of LDA Model results

## Loading Data

The dataset collected in the initial phase of this research was used in this step. The dataset comprised of 467 unique articles related to healthcare security and privacy collected from NKUs Steely Library. The CSV data file contains information on different Papers that were published between Jan 2008 and Mar 2021 (13 years and 3 months of data). These papers discuss a wide variety of topics on protecting the security and privacy of PHI & PII data in health care field.

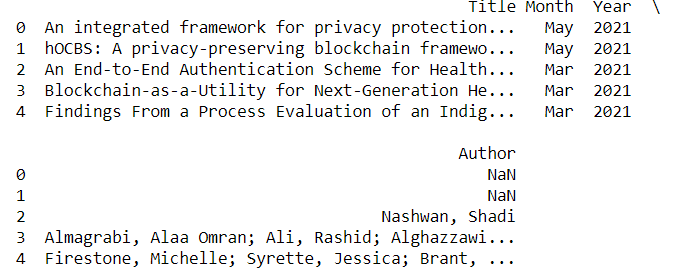


Figure : Dataset at a Glance

## Exploratory Data Analysis

At first, a time series chart of the number of articles published year-wise during the last 5 years was plotted to get an understanding of how the papers were distributed.

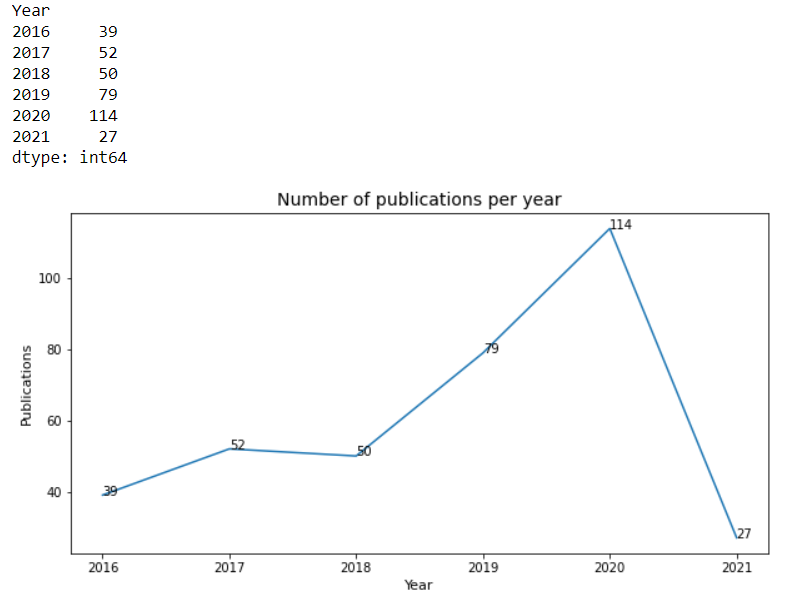


Figure : Papers published in each year

In this step, simple preprocessing of data was performed on the “Title” column of the dataset. As part of this stage, text analytics preprocessing steps such **as removal of punctuation, conversion to lowercase, lemmatization, extraction of nouns and stop word filtering** were performed.



Figure : Text Preprocessing Steps

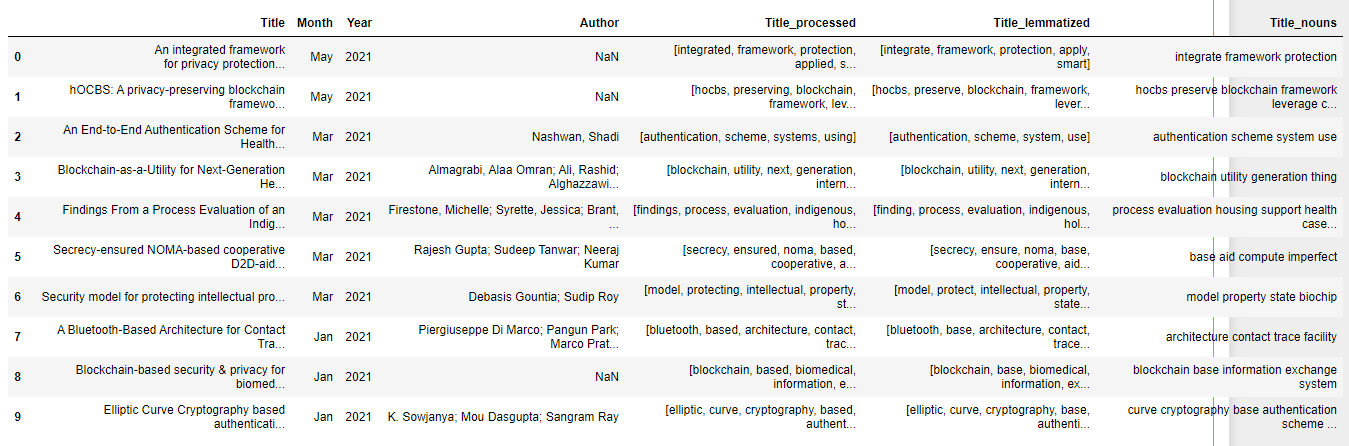


Figure : Dataset after text pre-processing

For the stop word filtering, a trial-and-error method was used to find the list of most repeating words that does not provide any meaning for the research. And these words were added to the standard set of English stop word filter available in the Word Cloud Python library. Once all these preprocessing steps were performed, a word cloud was produced to visualize the most frequently appearing words in the dataset.

This reveals that a lot of researchers were performed around studying the IoTs role in healthcare privacy, and a lot of survey-based studies were made on this domain. Several of the recent studies included in the dataset were about studying the effect of Covid-19 on healthcare security and privacy.

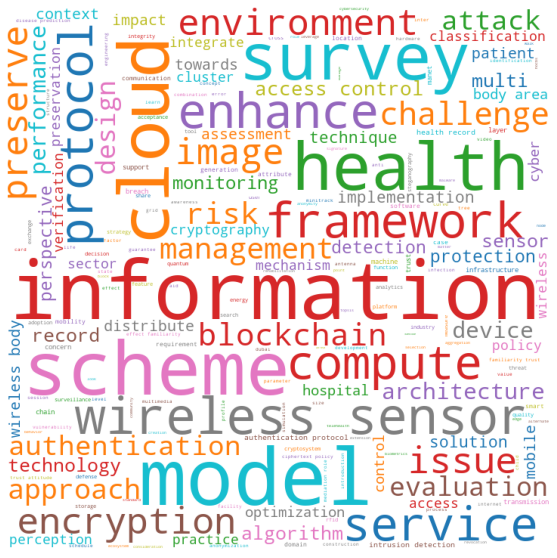


Figure : Wordcloud of the processed titles

Based on this, a bar plot was created to understand the most frequently appearing words and the number of times they appeared. Based on this analysis we were able to get the following plot.

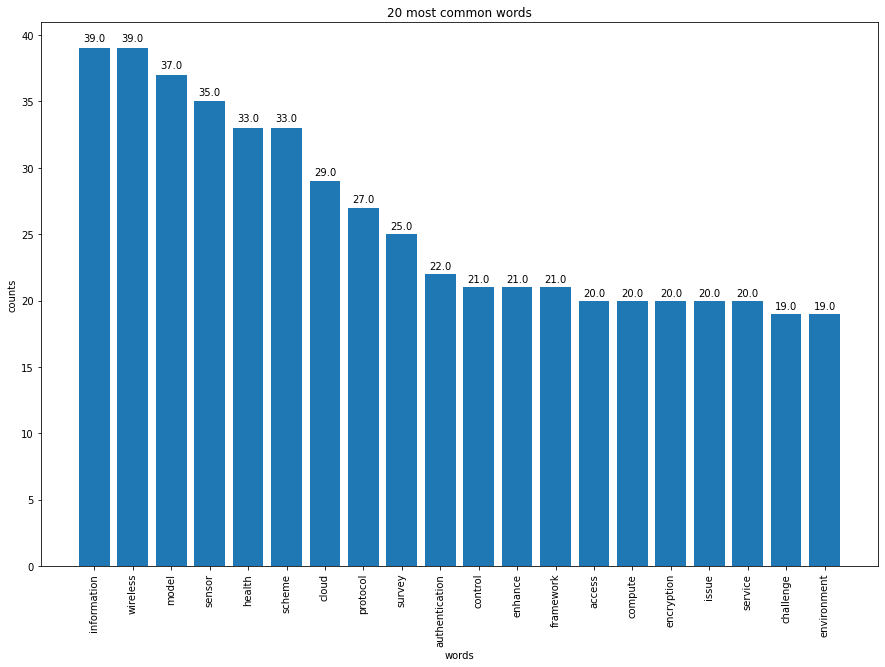


Figure : Top 20 most frequently occurring words

## LDA Model Development and Training

The LDA topic model algorithm requires a document word matrix as the main input. In order to generate this document word matrix, we use Count Vectorizer python module in sklearn library. This Document-Word Matrix was produced through the following method.

Text

Description automatically generated

Figure : Vectorization of dataset (Document Titles and Bag of Words)

Once this matrix was generated, its sparsity index was studied to see how many cells in this matrix will be non-zero. Sparsity is nothing but the percentage of non-zero datapoints in the document-word matrix, that is vectorized\_data.

Graphical user interface, text, application

Description automatically generated

Figure : Sparsity calculation

Then the LDA model was built using the SKLearn.decomposition Python Library through the LatentDirichletAllocation module. In the initial iteration, the number of topics (k-value) was set to 20. Thereafter the model with best number of topics and other hyper parameters was chosen through plotting the log-likelihood and perplexity factor.

Graphical user interface, text, application

Description automatically generated

Figure : Initial Topic Modelling via LDA

|  |  |
| --- | --- |
| **Topic Number** | **Keywords** |
| Topic #0: | service, assessment, device, evaluation, health, cloud, hospital, record, cross, model |
| Topic #1: | model, access, information, control, enhance, approach, management, towards, hospital, risk |
| Topic #2: | health, control, perspective, information, access, risk, cloud, environment, blockchain, protection |
| Topic #3: | framework, protocol, authentication, enhance, context, design, chain, integrate, blockchain, preserve |
| Topic #4: | compute, cloud, detection, policy, intrusion, attack, cryptography, information, requirement, model |
| Topic #5: | survey, health, information, image, record, preserve, classification, smart, storage, quantum |
| Topic #6: | solution, sector, evaluation, risk, transmission, cloud, model, record, multimedia, concept |
| Topic #7: | sensor, wireless, protocol, algorithm, survey, model, preserve, enhance, performance, distribute |
| Topic #8: | impact, trust, performance, communication, protocol, technology, perspective, concern, acceptance, anti |
| Topic #9: | authentication, scheme, management, compute, protocol, information, approach, trust, towards, profile |
| Topic #10: | service, function, insider, framework, encryption, context, layer, file, learn, threat |
| Topic #11: | attack, record, blockchain, approach, evaluation, health, platform, information, attribute, hardware |
| Topic #12: | issue, challenge, architecture, information, optimization, design, management, technology, infrastructure, algorithm |
| Topic #13: | environment, model, technique, child, service, detection, domain, infection, image, cross |
| Topic #14: | image, steganography, breach, technique, search, matter, frequency, scheme, quality, tool |
| Topic #15: | wireless, body, monitoring, area, preservation, scheme, preserve, authentication, health, enhance |
| Topic #16: | challenge, mobile, multi, introduction, minitrack, architecture, technology, redundancy, evolution, trace |
| Topic #17: | implementation, mobility, design, device, technique, concern, card, cyber, access, comparative |
| Topic #18: | perception, patient, effect, risk, cyber, trust, role, mediation, attitude, familiarity |
| Topic #19: | encryption, scheme, image, strategy, internet, performance, blockchain, protection, generation, mobile |

## Validation of LDA Model Results

The very first step in evaluating the model performance is to study the log-likelihood and perplexity score.

Log-likelihood is the logarithmic value of each probability in Document-Term Vector. In general, the Probability of each word in a Document-Term vector is denoted by p(wd). The log-likelihood is denoted as log P(wd). A model with higher log-likelihood and lower perplexity is considered to be good.

Perplexity Score is given by the formula,

Text

Description automatically generated

The measures of log-likelihood and perplexity for our model comes to be as follows.

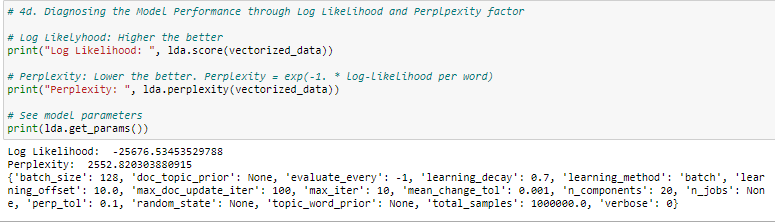


Figure : Log-likelihood and Perplexity Score Derivation

Once this is found, our next interest is to find the best LDA model that would fit our analysis. To find this, GridSearch was performed. The most important tuning parameter for LDA models is n\_components (number of topics). The grid search constructs multiple LDA models for all possible combinations of param values in the param\_grid dictionary (which in our case contains only the n\_components hyperparameter).

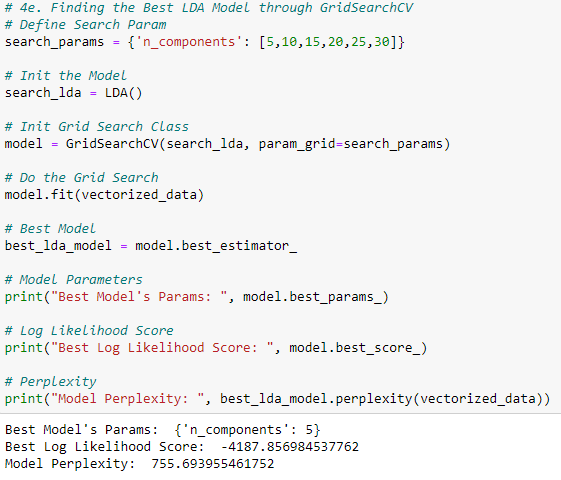


Figure : GridSearch to find the best fit LDA

The best model was found to be the one with number of topics as 5. Plotting the log-likelihood scores against num\_topics, clearly shows number of topics = 5 has better scores. To tune this even further, a finer grid search for n\_topics between 5 & 10 can be done. But this is considered for future scope of work. Bottomline is that a lower number of distinct topics is reasonable for this dataset.

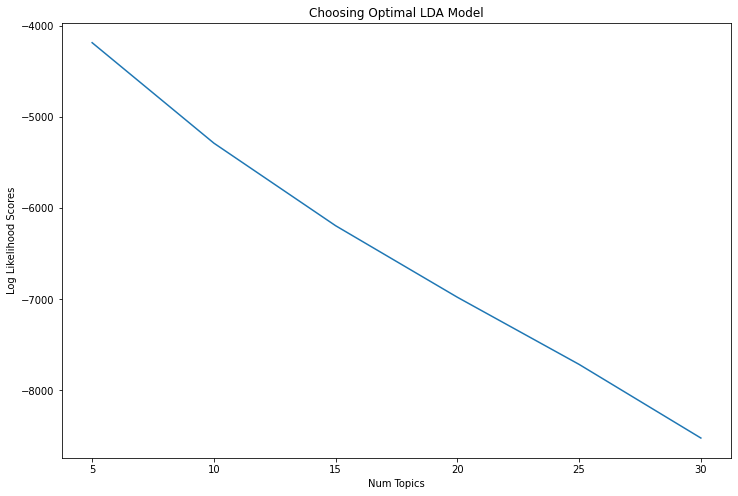


Figure : Line Plot of Log-Likelihood Score vs Number of Topics in each LDA Model

With this analysis, the 5 topics produced by our best fit LDA model was listed. Table below shows this information.

Table : Topics in Best Fit LDA Model and its keywords

|  |  |  |
| --- | --- | --- |
| **Topic Number** | **Keywords** | **No. of Articles** |
| Topic #0: | survey, approach, challenge, information, detection, device, architecture, model, enhance, practice | 101 |
| Topic #1: | cloud, record, health, preserve, scheme, management, encryption, monitoring, compute, service | 93 |
| Topic #2: | wireless, information, model, sensor, performance, issue, body, evaluation, scheme, area | 131 |
| Topic #3: | sensor, protocol, wireless, scheme, authentication, image, cluster, challenge, cyber, survey | 74 |
| Topic #4: | access, control, framework, design, model, risk, health, trust, context, chain | 68 |

The number of documents in each topic was plotted for the best fit LDA model. The bar plot shows a maximum number of papers towards Topic 0 and Topic 2.

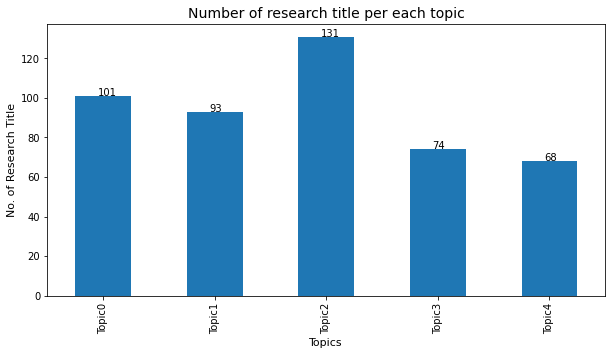


Figure : Count of Research Papers across Topics in best fit LDA

Our next interest is to see the dominant topics in each document. To classify a document as belonging to a particular topic, a logical approach is to see which topic has the highest contribution to that document and assign it. In the table below, all major topics in a document are greened out and the most dominant topic are assigned in its own column. Table below shows a detailed look at the Topic classification for first 15 papers on our dataset. Likewise, the topic classification for all 542 papers in our dataset was made.



Figure : Document-Topic Probability Distribution

Once all these are determined, our next interest would be to validate that the model we derived through grid search is in fact the best fit model. This can be checked through clustering analysis of topics derived in the best fit LDA. The pyLDAvis.sklearn.pepare() method accepts a Document-Term vector and a multi-dimensional scaling parameter – which in turn tries to scale the topics (topic bubbles) in two-dimensional space instead of n-terms in the BoW (Bag of Words). The general assumption is that the topics are distributed in a n-dimensional space, where n is the total number of terms in BoW (features in Document-Term vector). The centers of the default topic circles are laid out in two dimensions according to a multidimensional scaling (MDS) algorithm that is run on the inter-topic distance matrix (24).

The bubbles on the left shows the topics derived through our best fit LDA model. The size of the bubble corresponds to number of terms from BoW that belong to the corresponding topic. The axes are derived through a tSNE (t-Stochastic Neighbor Estimation) Multi-Dimensional Scaling algorithm. We are using a tSNE here because we know that the topics are non-linearly distributed in the document-word vector space.

By measuring the intra-cluster and inter-cluster distances, we will be able to tell that the produced model is a best fit one. Intra-cluster topic distance is the measure of how close each topic within each cluster is to every other topic in its cluster. Inter-cluster distance is the measure of how close each cluster of topics is to other clusters. Models that produce relatively small intra-cluster distances and relatively large inter-cluster distances evaluate favorably because they appear to be doing a good job of grouping like topics with discrete characteristics.

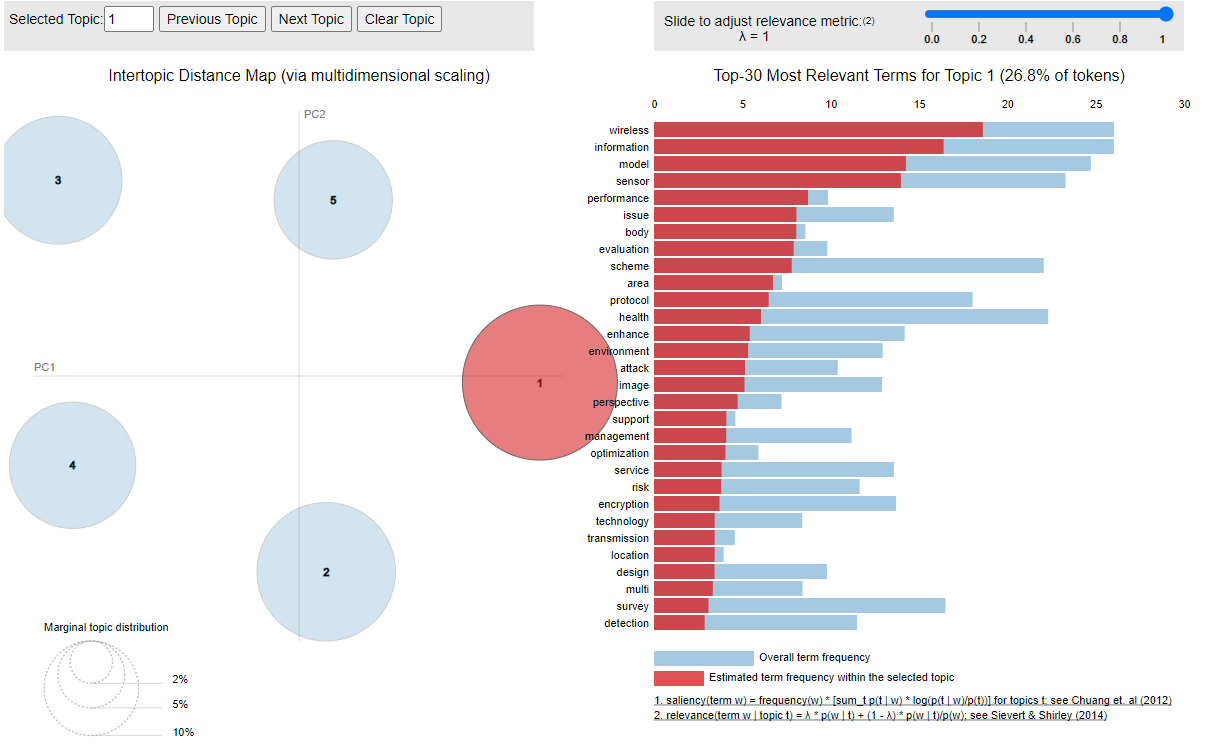


Figure : Visualization of Topics in a 2 dimensional space through Dimensionality Reduction

On the right side of the visualization, a bar chart shows the top 30 keywords from the topic-term distribution matrix. The keywords that appear on the plot, changes with selection of each topic. For each keyword, its overall frequency in the BoW is shown as Blue bar (referred to as Saliency). And the relative frequency that the keyword appears in a particular (selected) topic is shown as Red bar (referred to as Relevance).

# Conclusion

Based on our research, it can be interpreted that several of the studies were performed in analyzing and evaluating the performance of security models implemented using wireless sensors. And a vast majority of papers were contributed towards surveying the security and privacy challenges of health care IT systems.

The future study of this research would continue with focus on development of a prediction model for predicting topics of unknown text input. And thereby optimizing the topic modelling algorithm that has been developed by passing various different datasets.

# References

1. Health IT Dashboard - <https://dashboard.healthit.gov/quickstats/quickstats.php>
2. Raman, A. (2007, November). Enforcing privacy through security in remote patient monitoring ecosystems. In 2007 6th International Special Topic Conference on Information Technology Applications in Biomedicine (pp. 298-301). IEEE.
3. Appari, A., & Johnson, M. E. (2010). Information security and privacy in healthcare: current state of research. International journal of Internet and enterprise management, 6(4), 279-314.
4. Kripalani, S., LeFevre, F., Phillips, C. O., Williams, M. V., Basaviah, P., & Baker, D. W. (2007). Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care. Jama, 297(8), 831-841.
5. Warkentin, M., Johnston, A. C., & Shropshire, J. (2011). The influence of the informal social learning environment on information privacy policy compliance efficacy and intention. European Journal of Information Systems, 20(3), 267-284.
6. Vaast, E. (2007). Danger is in the eye of the beholders: social representations of information systems security in healthcare. The Journal of Strategic Information Systems, 16(2), 130-152.
7. Munson, S. A., Cavusoglu, H., Frisch, L., & Fels, S. (2013). Sociotechnical challenges and progress in using social media for health. Journal of medical Internet research, 15(10), e226.
8. Liginlal, D., Sim, I., Khansa, L., & Fearn, P. (2009). Human Error and Privacy Breaches in Healthcare Organizations: Causes and Management Strategies. AMCIS 2009 Proceedings, 406.
9. Office of the National Coordinator for Health Information Technology. 'Individuals' Perceptions of the Privacy and Security of Medical Records and Health Information Exchange,' Health IT Quick-Stat #58. dashboard.healthit.gov/quickstats/pages/consumers-privacy-security-medical-record-information-exchange.php. June 2019. <https://dashboard.healthit.gov/quickstats/pages/consumers-privacy-security-medical-record-information-exchange.php>
10. Office of the National Coordinator for Health Information Technology. 'Individuals Use of Technology to Track Health Care Charges and Costs,' Health IT Quick-Stat #57. dashboard.healthit.gov/quickstats/pages/consumers-health-care-charges-costs-online.php. April 2018
11. Al Ameen, M., Liu, J., & Kwak, K. (2012). Security and privacy issues in wireless sensor networks for healthcare applications. Journal of medical systems, 36(1), 93-101.
12. <https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/#13compareldamodelperformancescores>
13. Ahmadi, H., Arji, G., Shahmoradi, L., Safdari, R., Nilashi, M., & Alizadeh, M. (2019). The application of internet of things in healthcare: a systematic literature review and classification. Universal Access in the Information Society, 18(4), 837-869.
14. Mehraeen, E., Ghazisaeedi, M., Farzi, J., & Mirshekari, S. (2017). Security challenges in healthcare cloud computing: a systematic review. Global Journal of Health Science, 9(3), 157-157.
15. Ahmed, A., Latif, R., Latif, S., Abbas, H., & Khan, F. A. (2018). Malicious insiders attack in IoT based multi-cloud e-healthcare environment: A systematic literature review. Multimedia Tools and Applications, 77(17), 21947-21965.
16. Nazir, S., Ali, Y., Ullah, N., & García-Magariño, I. (2019). Internet of things for healthcare using effects of mobile computing: a systematic literature review. Wireless Communications and Mobile Computing, 2019.
17. Yao, W., Chu, C. H., & Li, Z. (2012). The adoption and implementation of RFID technologies in healthcare: a literature review. Journal of medical systems, 36(6), 3507-3525.
18. Wamba, S. F., Anand, A., & Carter, L. (2013). A literature review of RFID-enabled healthcare applications and issues. International Journal of Information Management, 33(5), 875-891.
19. Box, D., & Pottas, D. (2013). Improving information security behaviour in the healthcare context. Procedia Technology, 9, 1093-1103.
20. Ermakova, T., Huenges, J., Erek, K., & Zarnekow, R. (2013). Cloud Computing in Healthcare–A literature review on current state of research.
21. Tandon, A., Dhir, A., Islam, N., & Mäntymäki, M. (2020). Blockchain in healthcare: A systematic literature review, synthesizing framework and future research agenda. Computers in Industry, 122, 103290.
22. Behera, R. K., Bala, P. K., & Dhir, A. (2019). The emerging role of cognitive computing in healthcare: A systematic literature review. International journal of medical informatics, 129, 154-166.
23. Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics, 114, 57-65.
24. <https://cran.r-project.org/web/packages/LDAvis/vignettes/details.pdf>