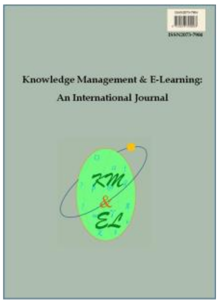
Knowledge Management & E-Learning, Vol.13, No.4. Dec 2021

Automated thematic analysis of health information technology (HIT) related incident reports

Yanyan Li
Casper Shyr
Elizabeth M. Borycki
Andre W. Kushniruk
University of Victoria, BC, Canada



Knowledge Management & E-Learning:
An International Journal

Knowledge Management & E-Learning: An International Journal (KM&EL)
ISSN 2073-7904

Recommended citation:
Li, Y., Shyr, C., Borycki, E. M., & Kushniruk, A. W. (2021). Automated thematic analysis of health information technology (HIT) related incident reports. *Knowledge Management & E-Learning, 13*(4), 408–420. https://doi.org/10.34105/j.kmel.2021.13.022

Knowledge Management & E-Learning, 13(4), 408–420

Automated thematic analysis of health information technology (HIT) related incident reports

Yanyan Li*

School of Health Information Science
University of Victoria, BC, Canada
E-mail: yayanli@uvic.ca

Casper Shyr

School of Health Information Science
University of Victoria, BC, Canada
E-mail: casshyr@hotmail.com

Elizabeth M. Borycki

School of Health Information Science
University of Victoria, BC, Canada
E-mail: emb@uvic.ca

Andre W. Kushniruk

School of Health Information Science
University of Victoria, BC, Canada
E-mail: andrek@uvic.ca

*Corresponding author

Abstract: In this paper, the authors describe a method for exploring the feasibility of using Natural Language Processing (NLP) and Machine Learning (ML) techniques to analyze patient safety incident database reports for themes. We developed a novel thematic analysis strategy to automatically detect keywords and latent themes that describe HIT-related patient safety incidents. The strategy was applied to patient safety reports to test the approach. The efforts by the automated strategy were compared to the efforts by analysts who manually reviewed and identified key words, topics, and themes for the same reports. The computer-based error themes were also compared to the human-determined themes for crosschecking. The manual thematic analysis took about 150 hours to complete on the patient safety reports. The semi-automated approach took only 10% of that time. 95% of the themes extracted from the automated method were aligned with the themes from the manual process. The findings underscore the utility of NLP and ML in identifying thematic patterns embedded in large numbers of unstructured data. The NLP-ML method therefore represents a valuable addition to the tools of detecting and understanding HIT-related errors.

Keywords: Patient safety incidents reporting; Natural language processing; Machine learning; Text mining

Biographical notes: Yanyan Li is a PhD student and researcher in the School of Health Information Science at the University of Victoria, Canada. She has over a decade of experience with health information technologies and over eight years of experience in improving the safety and quality of health information technology.

Dr. Casper Shyr is an adjunct assistant professor in the School of Health Information Science at the University of Victoria, Canada. He also directs the analytics and data science strategies within a large-scale healthcare organization. His key research focuses on the best practices to deliver healthcare analytics that drive impactful improvements to quality of patient care and efficiency in system operations. His work includes the development and deployment of data science modeling as well as structured change management integration for streamlined adoption into clinical practices.

Dr. Elizabeth Borycki (RN, PhD, FACMI, FCAHS, FIAHSI) is a professor in the School of Health Information Science at the University of Victoria, Canada, and a Michael Smith Foundation Health Research BC health professional investigator. She directs the Global Laboratory for Digital Health Innovation, where she leads a team of researchers who focus on health technology and safety. Dr. Borycki's health informatics research is in the areas of human factors, implementation science, and strategy involving health, technology, and safety.

Dr. Andre Kushniruk (PhD, FACMI, FCAHS, FIAHSI) is a professor and director of the School of Health Information Science at the University of Victoria, British Columbia, Canada. He has been published widely in health informatics and is known for his work in the usability of healthcare information systems. Dr. Kushniruk conducts research in a number of areas; he focuses on developing new methods for the evaluation of information technology in healthcare. Dr. Kushniruk has been a key researcher on several national and international collaborative projects. His work includes evaluation of systems for use by healthcare providers, patients, and citizens.

1. Introduction

In 1999, the seminal report "To Err is Human" was published by the Institute of Medicine (IOM). The paper recommended using health information technologies (HIT) to automate repetitive, time-consuming, and error-prone tasks (Kohn et al., 2000). While automation holds promising prospects for improving patient safety, a number of researchers have also reported the existence of several new and different types of technology-induced error (Ash et al., 2004; Kushniruk et al., 2005; Borycki, 2013). These researchers suggested that the use of HIT may lead to other latent types of errors that contribute to patient safety incidents and events.

Patient safety risks and incidents caused by problems associated with the use of HIT are now recognized as technology-induced errors (Borycki, 2013). Research has shown that incidents of error increase with high uptake of technologies in healthcare (Ash et al., 2004; Magrabi et al., 2010; Meeks et al., 2014). To reduce the errors and suggest methods for intervention, we need to know what and where those errors are occurring. An incident reporting database is a source to find and analyze errors.

The analysis of incident reports can uncover error patterns and serve as a cornerstone of patient safety improvement. Further, a systematic and automated approach to study correlations and dependencies or get a quantitative overview of frequently occurring events or incidents will enable the discovery of error patterns and HIT improvement opportunities. Such research is essential to develop an understanding of patient safety-related events and technology-induced errors, and is essential to improve the quality of healthcare, health professional provider experience, and patient safety.

While it is recognized that incident reports could support identifying risks and limitations of HIT, there is little technical support available to automate and optimize the analysis of record contents systematically or to retrieve key insights in a scalable way. Incident reports are manually reviewed to evaluate the level of harm to patients and to identify proper countermeasures. There has been some research on manually auditing and evaluating HIT-related events for identifying system breakdowns and opportunities from an optimization, practice, policy, and workflow perspective (e.g., Recsky et al., 2019; Williams, 2019). The analysis of many HIT-related reports at once to detect error themes and patterns automatically has not been done.

In this study, we employ natural language processing (NLP) and machine learning (ML) techniques to process free-text HIT-related descriptions contained in a large number of HIT-related incident reports. The data extracted for this study include structured and unstructured narrative data about patient safety events. The narratives include two description data elements: one is the description about the event, and the other is the essential information for evaluation of the causality and description of the HIT-related incident. This paper focuses on the methodological approach of the content analysis of the narrative data.

The objectives of the semiautomated approach included the following: reduce inherent analytics subjectivity in analyzing error themes, create new ways to observe the data, and increase the efficiency of thematic analysis. In particular, this would offer a unique opportunity to mathematically explore latent themes within a large number of reports as an exemplar to discuss how HIT error reports can be leveraged to address patient safety issues.

2. Methods approach

As mentioned above, the source of data used to test the approach consisted of incident data. The search strategy was designed to include all the reports where the unstructured field indicating that a computer contributed to the event was not null. Both structured and unstructured data elements were pulled for the study. The unstructured data included two narrative data elements: the field indicating that the computer system contributed to the error, and a description of the error. In this paper, we describe the methodological approach for the semiautomated analysis of the unstructured data.

The study included two parts: a manual thematic analysis conducted by analysts and a semiautomated thematic analysis using NLP and ML techniques. The manual thematic analysis for each report followed the common steps for thematic analysis: familiarization of the data, data coding, codes grouping, and themes generalization (Majumdar, 2018). The semiautomated approach included data preprocessing and transformation, keywords identification, keywords clustering, and topic modeling.

In the manual analysis, the analysts conducted a thematic analysis of the data. The description data was referenced when the analysts had disagreement or needed more information to understand the events. They first discussed and agreed on a guideline for identification of codes: a word or a phrase that best represented the HIT-related descriptions. For example, for a HIT-related description where the pharmacy did not transcribe onto a medication reconciliation record (MAR), the word MAR would be identified as a HIT function, so it would be selected as the code that represents this description. The analysts first identified the codes and themes individually. They then compared the results and reconciled inconsistencies with the codes and themes until an agreement was reached. As such, in the end, there was no discrepancy with codes or themes in the manual process.

The same dataset was then analyzed and processed by the semiautomated approach described in this paper. The process of the automated approach followed the common steps of the thematic analysis: data was first processed and transformed to identify keywords that were equivalent to codes in the manual process; the keywords were then assigned into different clusters. Themes that represented the clusters were manually interpreted by one of the authors (C.S.), who has a PhD in bioinformatics and more than ten years of experience in data analysis and qualitative research. Themes identified by the machine-based approach were compared to the themes obtained by human-based analysis for accuracy and efficiency.

As described earlier, the semiautomated approach included four steps. First, the data was preprocessed and transformed into mathematical representations using NLP techniques, such as tokenizing, punctuation, stop words removal, annotation, and word association using stemming and lemmatization. Second, the NLP techniques were applied to automatically identify keywords. These techniques included N-gram, Bag-of-Words, and Named Entity. Third, the ML models were used to cluster the records by key words. Topic Modeling was applied to analyze the attributes of clusters. Last, the themes within each cluster were manually identified by C.S. For evaluation purposes, the results of the automated analyses were cross-checked by comparing them to the manually analyzed themes. Fig. 1 is a visualization of this methodology.

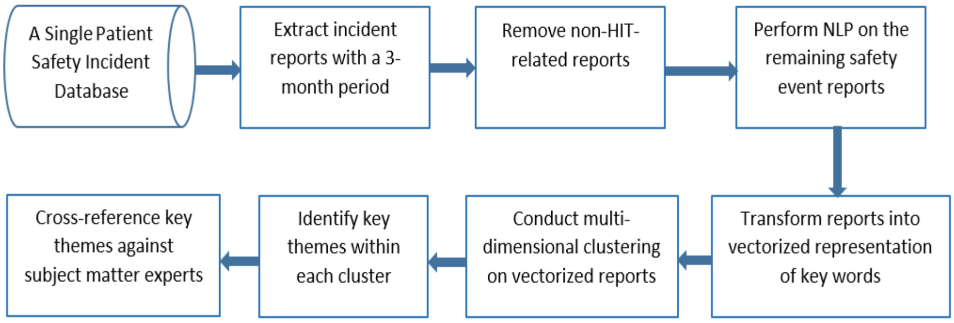

A Single Patient Safety Incident Database
Extract incident reports with a 3-month period
Remove non-HIT-related reports
Perform NLP on the remaining safety event reports
Transform reports into vectorized representation of key words
Conduct multi-dimensional clustering on vectorized reports
Identify key themes within each cluster
Cross-reference key themes against subject matter experts

Fig. 1. The stages of the methodology

3. Technology-driven solutions to incident analysis

There is a range of technical solutions that could be applied to automatically analyzing incident reports. In this section, we provide background to our methodology and delineate the steps involved.

3.1. Natural language processing (NLP)

NLP is a class of data science and ML techniques that can be used to analyze unstructured data. NLP can help get the right information and provide good suggestions by automatically analyzing massive amounts of data. It is an aspect of ML that helps computers understand, interpret, and use human language. It allows computers to communicate with people using a human language. It enables computers to read text, hear speech, and interpret text and speech. It can be used in an attempt to close the gap between human and computer communications (Beysolow, 2018b).

NLP draws from several disciplines, including linguistics and computer science. At a high level, the way NLP works is it breaks down language into shorter, more basic pieces called "tokens," attempts to understand the relationships between the tokens, explores how they fit together, and as a result provides insights and meaningful information from the given text. High-level NLP tasks include two steps: the first is cleaning and preprocessing and the second is language understanding and generation. The cleaning and preprocessing step involves tokenization, stemming, lemmatization, part-of-speech tagging, and chunking. The language understanding and generation steps involve the employment of Deep Learning algorithms (Taulli, 2019).

3.2. Machine learning (ML)

ML by name means "machine learning"; however, ML is not just about learning, but also about understanding and reasoning by being trained with data. Large amounts of data and algorithms can be used to train a machine and enable it to learn how to perform certain tasks. Generally speaking, machines perform better and have higher accuracy rates by training with more data and having better computer algorithms or models with faster computing power and cheaper memory. ML is used for artificial intelligence in general and as a technique within NLP (Panesar, 2021).

3.3. NLP models applied and steps

This section describes the steps and stages for the automated approach. As well, trade-offs in selection of approaches used are described.

3.3.1. Text pre-processing

The first step of the automated approach was to transform the text into a state that a computer could understand or parse. This process included the following steps: converting text to lower- or uppercase, removing numbers or converting them into words, removing punctuation, accent marks, and other diacritics, removing white spaces, expanding abbreviations, removing stop words, removing sparse terms (e.g., terms occurring twice in a document) or particular words, stemming, lemmatization, and text canonicalization (Vasiliev, 2020).

3.3.2. Tokenization

In this step, the texts of each report were broken down into pieces such as words, sentences, or phrases that could be used for tokenization. For example, a sentence or paragraph is broken down into words based on the delimiter "space." While in sentence tokenization, a paragraph is broken down into sentences based on "period mark."

3.3.3. Stemming and lemmatization

Stemming is a process of normalizing words into their base form or root form. For example, the stem of the three words "going," "goes," and "gone" is "go." Another example is the stem of the words "intelligence," "intelligent," and "intelligently": "intelligen." The problem of stemming is that the produced intermediate representation of some words may not have any meaning (e.g., intelligen, fina, etc.).

Lemmatization is an alternative way of stemming that is meant to get away from the problem of stemming. Lemmatization takes into consideration the morphological analysis of a word. It is the same as stemming, but the intermediate representation or root form has a meaning. For example, from the lemmatization process, the representation of the words "intelligence," "intelligent," and "intelligently" is "intelligent." However, lemmatization takes more time than stemming. Lemmatization is used when the meaning of words is important for the purpose of analysis; for example, automated question-answering applications like chatbots that use semiautomated techniques to automatically answer questions via a business messenger (Vasiliev, 2020).

In our method, we explored stemming in our initial analysis of the data. It did not work as well as lemmatization, so it was not incorporated in the final data science pipeline. It is not a surprise that lemmatization gave better results than stemming. Stemming is an easy step to try, and it is an accepted practice within the data science community to use it as a baseline for performance.

We also introduced a step in which certain healthcare vocabularies were programmed into the data science pipeline to improve the NLP performance. This included understanding clinical acronyms (for example, EMR: electronic medical record), or how vendors' EMR names should be interpreted along the same lines as "EMR." This step involved exploratory analysis of the free text data to get a sense of the range of vocabularies that needed this customized manipulation.

3.3.4. Stop words removal

We used the stop words dictionary in the NLK package in Python to remove words that had no meaning or were not able to express any special meaning based on some specific context, such as "to," "the," "this," "you," etc.

3.3.5. Bag-of-words (BoW), TF-IDF, and n-grams

Models The BoW, TF-IDF, and N-Grams models were applied in order to identify the features of the reports. The BoW model is a method to create a vocabulary of unique words extracted from a document, and then a vector is created that contains the frequency of the unique words in the corresponding document, disregarding its semantic, syntactic, or order of words (in other words, the importance of words) information. This model is only concerned with whether known words occur in a document, not where in the document they occur. With a bag-of-words text file, the counts of each of the words in the text file can be calculated and visualized to extract features from the text file (Beysolow, 2018a).

TF-IDF is short for "term frequency-inverse document frequency." It is another technique to visualize text. Compared to the BoW model, the TF-IDF model preserves some semantic information because uncommon words are considered more important than common words. For example, for the sentence "she is beautiful," the word "beautiful" will be given more importance than the other two words "she" and "is" in a TF-IDF table. The challenge of the TF-IDF matrix is its high dimension (very sparse) and noise as it still includes many low-frequency words. As such, this model is often used with dimension reduction models that will be talked about in the advanced NL models' section.

Bag of n-grams is a natural extension of bag of words. In NLP, n-grams are used for a variety of things. Some examples include auto-completion of sentences (such as we see in Outlook or Gmail), auto spellcheck, and grammar check in a given sentence. An n-gram is simply any sequence of n-tokens (words) and involves analyzing a group of words together instead of one word at a time. Bags of n-grams can be more informative than bags of words because they capture more context around each word (e.g., "love this dress" is more informative than just "dress"). Fig. 2 below is an example of performing n-gram analysis.

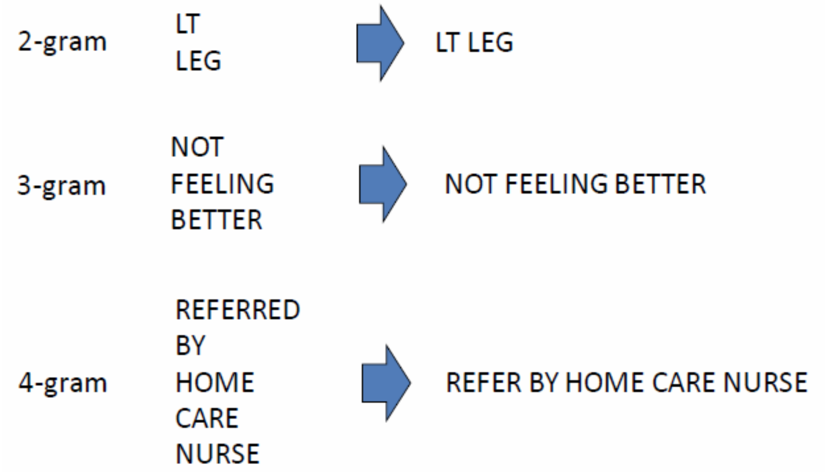


Fig. 2. Bag of n-grams example

3.4. Advanced NLP techniques

The advanced NLP techniques that we applied included topic modeling and clustering.

3.4.1. Topic modelling (TM)

TM is technique used to identify thematic patterns or latent topics in large quantities of text. A topic model is a type of probabilistic model that helps to examine massive amounts of documents, cluster similar groups of documents together, and identify what the topics might be (Jelodar et al., 2019). The input for the TM is a document-term matrix that can be produced by either the BoW or a TF-IDF model. The output of data modeling is a list of topics with associated clusters of words.

There are many approaches to topic modeling. For the exploratory nature of this research, only two approaches are used in this paper, which include the determinant Latent Semantic Analysis (LSA) and probabilistic Latent Dirichlet Allocation (LDA). Both LSA and LDA share a fundamental assumption about the latent semantic structure of the corpora, but they use different mathematical frameworks: the matrix algebra process for LSA and the probabilistic approach for LDA (Rizun et al., 2017).

In our study, we tried LSA and compared it against LDA. LDA gave a better performance, so the final data science pipeline only included LDA. It is generally considered a good practice to try different NLP approaches. The performance of each methodology can vary depending on the dataset that is worked with. There is no golden rule as to which method is absolutely the best. There is a lot of dependency on the type of data that one works with, such as the quality of text, the complexity of the content, the length of sentences, the sample size, and so on.

3.4.2. Latent semantic analysis (LSA)

LSA is the most common method for topic modeling. The idea of LSA is that words will occur in similar pieces of text if they have similar meanings. It is used to analyze relationships between a set of documents and the terms contained in them. The input of this model is the output of the TF-IDF matrix. Because the TF-IDF matrix is very sparse or has high dimension, LSA uses truncated singular value decomposition (SVD), a mathematical technique to reduce the dimensions of the matrix. A limitation of the LSA method is that it cannot distinguish multiple meanings of words (e.g., regression in the statistics vs. regression testing). It is less accurate than LDA (Rizun et al., 2017).

3.4.3. Latent dirichlet allocation (LDA)

LDA is the most popular TM model. The idea of LDA is to look at the text in terms of probability distributions and find hidden topics from the distribution (Blei et al., 2003). The input of the LDA model is the output of the BoW model. LDA is guided by two principles. The first principle is that LDA treats each text or document as a mixture of topics. For example, in a two-topic model, we could say that document A has 60% of topic A and 40% of topic B; document B has 20% of topic A and 80% of topic B.

The second principle is to view each topic as a mixture of words; for example, a two-topic model of “food” and “animals.” The most common words in the food topic include peach, broccoli, and milk, while the animal topic includes cats, owls, and chickens. LDA does not separate documents into discrete groups; instead, it allows documents to have overlap in terms of content. It is noted that in a classical LDA model, the number of topics is initially fixed and predetermined by users (Rizun et al., 2017).

3.4.4. Identification of themes within each cluster

In the study, after the topics were generated by either topic modeling or clustering methods, themes were assigned to the collections of words. Without semantic interpretation of what the word clusters meant or symbolized, the output would simply be a collection of words. During the interpretation phase, there is always potential for biases to influence the meaning of the topics. As such, validity checks were done to bolster interpretation.

4. Unsupervised machine learning - Clustering

Both clustering and topic modeling are methods of organizing the collection of documents. The difference is the output of the topic modeling is a list of topics with associated clusters of words, while the output of clustering is a list of clusters with every document showing up in one of the clusters.

Clustering is an unsupervised machine learning technique that attempts to partition documents into different groups based on some suitable similarity measure. The input of clustering is also a TF-IDF matrix. In this pilot study, clustering topics and modeling were used to identify the themes on the patient safety report datasets. The model used in this pilot study was the Principal Component Analysis.

4.1. Identification of themes within each cluster

As described earlier in this paper, the meaning of the collections of words were manually assigned after the clusters and topics were generated. Without semantic interpretation of what the word clusters meant or symbolized, the output was simply a collection of words. During the interpretation phase, there was potential for biases to influence the meaning of the topics. As such, cross-checks were done to bolster interpretation.

5. Results

To evaluate the accuracy and efficiency of the automated approach, the amount of effort spent by the manual and machine-based approaches were compared. As shown in Table I below, the manual process took about 150 hours to complete the thematic analysis, while the automated approach took about 15 hours. A total of 135 hours were saved by the automated approach.

Table 1
Comparison of the total hours spent by manual process vs. semiautomated approach

	Manual Analysis	Semiautomated Approach
Hours Spent	150 (5 hours per day, 2 work-weeks in total, by 3 analysts)	15 hours (3 workdays)

The linguistic and semantic characteristics of the safety event reports were identified during the data familiarization process in the development of the data science pipeline. As such, based on the linguistic and semantic characteristics, certain healthcare vocabularies were programmed into the data science pipeline to improve the NLP performance.

The syntax characteristics included the following:

- Complex sentences: when the NLP techniques were applied, complex sentences were split with efforts to keep their original meanings
- Misspellings
- Incomplete sentences, phrases, and/or keywords (e.g., “see above,” “meds,” “pt,” “Er”)
- Numeric expressions (e.g., “Q5 minutes”)
- Information summarized in a few words, resulting in compound phrases (e.g., “state worsening”)
- Clinical texts are error-prone because they are written under time pressure (Denecke, 2008)

The semantic characteristics included the following:

- Incomplete information. Because there were two fields for the description of a report, users often entered the main description (mostly contextual information) into the description of the event and minimal information into the HIT-related incident description field. In the manual process, the event description field was cross-referenced when the information in the HIT field was either too short or incomplete. In the semiautomated approach, both the event description and HIT-How computer system contributed description fields were analyzed.
- AM vs. am: AM here is used to refer to morning; "am" is a stop word in the NLK stop words dictionary. As such, this word was not removed by the NLP model.
- Involved persons (e.g., "I received a call from [doctor's name]," "MOA," "PCC"). Such data were manually removed during the process of familiarization of the data.
- Drug names
- Names of technical devices (e.g., "oxygenator," "cardiac monitor")
- Specific locations
- Abbreviations for facility names
- Clinical procedures
- Software names
- Acronyms (e.g., "iv" for "intravenous," "MAR")
- Host language and medical terms ("the length of time it takes to enter a TST")
- Subjective and not factual information (e.g., "If so why is the patient not on it," "I am unsure of the cause of this incident but I suspect the change in system …")

6. Discussion

This paper has described an NLP and ML-based approach to analyzing the themes emerging from the HIT-related patient safety reports. The goal of the work was to explore and trial a novel approach that could automate the identification of error themes from a large number of patient safety reports at once and on a regular basis.

The results of this work demonstrate the potential for the application of the same data science pipeline to scale and analyze a broader set of patient safety data. Manual thematic analysis was found to be labor-intensive and time-consuming. The data science pipeline produced 95% alignment between the themes for the NLP and ML approach of extraction versus what the analysts manually extracted. Further, the NLP and ML approach holds the advantage of being able to process data in real-time. Rather than periodically extracting and analyzing patient safety event data, the data science pipeline can be set up to analyze live safety event data as the data source updates. This will provide up-to-date information for informed decision-making, rather than having to rely on information that may be 12 months out of date.

Once the NLP and ML pipeline is set up, it can easily scale to any larger dataset of the same type. With an anticipated volume of more than 400 patient safety reports to analyze per year, the resource requirements and time complexity scale exponentially under the traditional manual thematic analysis. With NLP and ML, the same data science pipeline that has already been set up for the data analysis can easily be set up for 12 months of data. The manual interpretation of the output from the NLP and ML pipeline at the end still requires some minimal labor, but the data will have already been cleaned, processed, and grouped by the data science pipeline.

Another finding was the data characteristics of two narrative descriptions fields of a patient safety event: the event description and the description about how the HIT contributed to an event. The first one contained insights of the events from a broader clinical perspective, and the other contained the description of how HIT contributed to the event specifically. How the computer system contributed data was used as the core driver behind the semiautomated analysis. In the majority of cases, this field provided the context around how computer systems played a role in an event. In order to have high accuracy, the NLP and ML pipeline analyzed both fields. However, in the manual thematic analysis, only how the computer contributed to an event was coded and used for theme generation. The HIT field was referenced to gain more insights when the analysts had disagreements or when the HIT field was too short and required more information to be understood. Both the description and the computer system contribution fields were found to be critical in human and machine-based approaches.

Data quality problems unique to the incident reports were found. Extra efforts were put in the preprocessing step. Both narrative description (i.e., “Description” and HIT-related description) data sources contained problems, such as mislabeled incidents, clinical-based contextual information, clinical-based incident categories, domain-specific language, acronyms and abbreviations, colloquialisms (e.g., informal phrases, expressions, idioms), and errors (e.g., typos, wrong information, network issues mistaken for application issues) that ranged widely across different settings and contexts.

Another challenge of the semiautomated approach was that the error classification fell short with respect to HIT-related incidents. The impact of these errors on patient safety is less well understood, and the source of risk has yet to be examined. An error classification system for patient safety events could form the basis for semantic annotation of the reports. The system is necessary for adapting inference functionalities to various situations and application scenarios; it could make unstructured text automatically processed to generate useful information.

Lack of annotated data and a standard error classification system pose challenges for the application of the advanced NLP and ML techniques. As a result, this study used unsupervised machine learning techniques and did not consider ML techniques for classification or prediction. More review efforts from the community of HIT professionals would be useful in order to fully identify the features of HIT reports and develop a full taxonomy of patient safety errors. More patient safety professionals contributing to the incident report annotation tasks would help to create a data playground to apply advanced ML or Deep Learning methods.

In conclusion, we developed an NLP and ML-based approach to providing optimal, cost-effective, and efficient processes for identifying the themes from a patient incident database. The semiautomated approach took only 10% of the time of the manual approach in analyzing the records. Ninety-five percent of the themes generated from the semiautomated approach were consistent with the themes from the manual approach. We anticipate that the NLP and ML pipeline could easily scale to analyze any larger dataset of the same type, with an increasing number of patient safety reports to analyze per year.

As next steps, the semiautomated approach could be adapted and used in the education or training programs in health care. The approach will contribute to building the basis of HIT and health professional training programs about HIT-induced errors.

Health and HIT professionals could use the approach to identify how HIT contribute to patient safety events and as a result, they could develop targeted strategies to prevent or mitigate such errors (Borycki, 2015; Mattingly et al., 2012).

Author Statement

The authors declare that there is no conflict of interest.

ORCID

Elizabeth M. Borycki https://orcid.org/0000-0003-0928-8867

Andre W. Kushniruk https://orcid.org/0000-0002-2557-9288

References

Ash, J. S., Berg, M., & Coiera, E. (2004). Some unintended consequences of information technology in health care: The nature of patient care information system-related errors. *Journal of the American Medical Informatics Association, 11*(2), 104–112.

Beysolow II, T. (2018a). What is natural language processing? In T. Beysolow II (Ed.), *Applied Natural Language Processing with Python* (pp. 1–12). Apress. doi: 10.1007/978-1-4842-3733-5_1

Beysolow II, T. (2018b). Working with raw text. In T. Beysolow II (Ed.), *Applied Natural Language Processing with Python* (pp. 43–75). Apress. doi:10.1007/978-1-4842-3733-5_3

Blei, M. D., Ng, Y. A., & Jordan, I. M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Borycki, E. M. (2013). Technology-induced errors: Where do they come from and what can we do about them? *Studies in Health Technology and Informatics, 194*, 20–26.

Borycki, E. M. (2015). Towards a framework for teaching about information technology risk in health care: Simulating threats to health data and patient safety. *Knowledge Management & E-Learning, 7*(3), 480– 488

Denecke, K. (2008). Semantic structuring of and information extraction from medical documents using the UMLS. *Methods of Information in Medicine, 47*(5), 425–434. doi: 10.3414/ME0508

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications, 78*(11), 15169–15211. doi: 10.1007/s11042-018-6894-4

Kohn, L., T., Corrigan, J. M., & Donaldson, M. S. (2000). *To err is human: Building a safer health system*. National Academies Press.

Kushniruk, A., Triola, M. M., Borycki, E. M., Stein, B., & Kannry, J. L. (2005). Technology induced error and usability: The relationship between usability problems and prescription errors when using a handheld application. *International Journal of Medical Informatics, 74*(7/8), 519–526.

Magrabi, F., Ong, M. S., Runciman, W., & Coiera, E. (2010). An analysis of computer-related patient safety incidents to inform the development of a classification. *Journal of the American Medical Informatics Association, 17*(6), 663–670.

Majumdar, A. (2018). Thematic analysis in qualitative research. In M. Gupta, M.

Shaheen, & K. P. Reddy (Eds.), *Qualitative Techniques for Workplace Data Analysis* (pp. 197–220). IGI Global.

Mattingly, K. D., Rice, M. C., & Berge, Z. L. (2012). Learning analytics as a tool for closing the assessment loop in higher education. *Knowledge Management & E-Learning, 4*(3), 236–247.

Meeks, D. W., Smith, M. W., Taylor, L., Sittig, D. F., Scott, J. M., & Singh, H. (2014). An analysis of electronic health record-related patient safety concerns. *Journal of the American Medical Informatics Association*, *21*(6), 1053–1059.

Panesar, A. (2021). What is machine learning? In A. Panesar (Ed.), *Machine Learning and AI for Healthcare* (pp. 63–83). Apress. doi: 10.1007/978-1-4842-6537-6_3

Recsky, C., Blackburn, L., Muniak, A., Rush, K., MacPhee, M., & Currie, L. (2019). *Technology-mediated adverse events in primary and community care*. Retrieved from http://bcpslscentral.ca/wp-content/uploads/2020/02/RecskyAMIA_19-11-13_v2-1.pdf

Rizun, N., Taranenko, Y., & Waloszek, W. (2017). The algorithm of modelling and analysis of latent semantic relations: Linear algebra vs. probabilistic topic models. In *Proceedings of the International Conference on Knowledge Engineering and the Semantic Web* (pp. 53–68). Szczecin, Poland.

Taulli, T. (2019). Natural language processing (NLP). In T. Taulli (Ed.), *Artificial Intelligence Basics* (pp. 103–124). Apress.

Vasiliev, Y. (2020). The text-processing pipeline. In Y. Vasiliev (Ed.), *Natural Language Processing with Python and SpaCy* (pp. 15–30). No Starch Press.

Williams, A. (2019). Nursing informaticians address patient safety to improve usability of health information technologies. *Studies in Health Technology and Informatics, 257*, 501–507.