International Journal of Population Data Science

Journal Website: www.ijpds.org

A scoping review of preprocessing methods for unstructured text data to assess data quality

Marcello Nesca[1,2], Alan Katz[1,2,3], Carson K. Leung[4], and Lisa M. Lix[1,2,5]

Submission History

Submitted:	02/05/2022
Accepted:	28/07/2022
Published:	05/10/2022

[1]Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada
[2]Manitoba Centre for Health Policy, University of Manitoba, Winnipeg, MB, Canada
[3]Department of Family Medicine, University of Manitoba, Winnipeg, MB, Canada
[4]Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada
[5]George & Fay Yee Centre for Healthcare Innovation, University of Manitoba, Winnipeg, MB, Canada

Abstract

Introduction
Unstructured text data (UTD) are increasingly found in many databases that were never intended to be used for research, including electronic medical record (EMR) databases. Data quality can impact the usefulness of UTD for research. UTD are typically prepared for analysis (i.e., preprocessed) and analyzed using natural language processing (NLP) techniques. Different NLP methods are used to preprocess UTD and may affect data quality.

Objective
Our objective was to systematically document current research and practices about NLP preprocessing methods to describe or improve the quality of UTD, including UTD found in EMR databases.

Methods
A scoping review was undertaken of peer-reviewed studies published between December 2002 and January 2021. Scopus, Web of Science, ProQuest, and EBSCOhost were searched for literature relevant to the study objective. Information extracted from the studies included article characteristics (i.e., year of publication, journal discipline), data characteristics, types of preprocessing methods, and data quality topics. Study data were presented using a narrative synthesis.

Results
A total of 41 articles were included in the scoping review; over 50% were published between 2016 and 2021. Almost 20% of the articles were published in health science journals. Common preprocessing methods included removal of extraneous text elements such as stop words, punctuation, and numbers, word tokenization, and parts of speech tagging. Data quality topics for articles about EMR data included misspelled words, security (i.e., de-identification), word variability, sources of noise, quality of annotations, and ambiguity of abbreviations.

Conclusions
Multiple NLP techniques have been proposed to preprocess UTD, with some differences in techniques applied to EMR data. There are similarities in the data quality dimensions used to characterize structured data and UTD. While a few general-purpose measures of data quality that do not require external data; most of these focus on the measurement of noise.

Keywords
review; data quality; natural language processing

*Corresponding Author:
Email Address: Lisa.Lix@umanitoba.ca (Lisa M Lix)

https://doi.org/10.23889/ijpds.v6i1.1757

Introduction

Routinely collected electronic health data are generated during the process of managing and monitoring the healthcare system [1, 2]. Unstructured text data (UTD) are common in electronic medical records (EMRs), which is one type of routinely collected electronic health data. Further examples of UTD found in other types of routinely collected electronic health data, are laboratory testing results and clinical registry files.

The quality of data has been defined as their fitness for use [3–5], that is, that data meets the needs of the user for a specific task or purpose, such as identifying individuals with a specific health condition. Given that routinely collected electronic health data are increasingly being used for research, it is important to consider their fitness for research, including epidemiologic studies or health services utilization studies. There are several consequences of poor data quality. For example, Kiefer noted that poor data quality can "slow down innovation processes" [6]. Poor data quality may also increase the time required to prepare a dataset for use, which can impact the timeliness of research outputs. Data quality is a multidimensional construct; it encompasses such dimensions as relevance, consistency, accuracy, comparability, timeliness, accessibility and usability [4, 5, 7, 8]. Most data quality frameworks and assessment methods have been developed for structured data. However, Kiefer [6] argued that most, if not all, data quality dimensions developed for structured data are also relevant for UTD, although she emphasized the importance of relevance, interpretability, and accuracy when assessing UTD fitness for use. Kiefer also noted that there has been little research about data quality dimensions and indicators of these dimensions for UTD [6].

The usability of UTD for research generally requires the application of natural language processing (NLP) techniques, including topic modeling, sentiment analysis, aspect mining (e.g., identifying different parts of speech), text summarization, and named entity recognition (e.g., identifying people, places, and other entities in unstructured data) [9–14]. To prepare UTD for one or more of these NLP techniques, preprocessing of the data is an essential step. Preprocessing of UTD includes such actions as removing stop words (i.e., common words in a language), removing punctuation, tagging (i.e., identifying or labelling) parts of speech, and transforming abbreviations into words or phrases so that they can be easily interpreted. Accordingly, some researchers have suggested that indicators that measure the outputs of data preprocessing steps, such as the number or percent of abbreviations and the number or percent of spelling errors, could be used to characterize UTD quality. Some types of NLP, such as named entity recognition, involve training classification models to learn to identify data entities, such as parts of speech, diseases, or geographic locations [10]. This requires the use of annotated databases as a "gold standard", which have been tagged (e.g., parts of speech have been documented or labelled) using manual or automated methods [10]. The quality of these annotated databases and the accuracy of classification models based on NLP applications involving annotated databases have also been proposed as indicators of UTD quality. Additionally, several unsupervised or supervised methods, which are used for both internal or external validation have been used for preprocessing and may be used to develop indicators for data quality. For example, in a study by Zennaki et al. [15], a recurrent neural network was used for unsupervised and semi-supervised parts of speech tagging in languages for which there are no labeled training data.

The data quality paradigm places a high value on the representation of truth from the perspective of the patient [16]. In a citizens' jury study by Ford et al. [16], a representative sample of citizens listened to subject matter experts about the sharing of UTD within EMRs for research. The jury then deliberated to reach a conclusion from questions they were asked. With respect to data quality, the jurors noted that that text data may contain information about patients, judgments and offhand comments that may be misinterpreted by the researcher [16]. The concern with the representation of truth is a form of external validation (i.e., assessing data veracity). A study by Pantazos et al. [17], discussed the preservation of medical correctness, readability and consistency after EMR records were de-identified, to ensure data quality from a representativeness perspective. At the same time, NLP technologies still have difficulties with context and understanding language [18, 19]. Preprocessing activities to prepare text data for research do not address the contextual concerns that the jurors raised. Thus, it should be noted that this study primarily focuses on assessing the goodness of fit of text data for analytics.

In summary, there are potentially many indicators that could be used to describe UTD quality, and these are primarily based on the use of NLP techniques to preprocess UTD. However, there is little relevant literature and few, if any, guidelines on the data quality indicators that might be recommended for inclusion in data quality frameworks for UTD, or that might be used to guide data preprocessing in studies that apply NLP methods to UTD. In addition, there have been few studies that have investigated the impact of UTD quality assessment on the performance of text analyses using NLP. The objective was to systematically document current research and practices about NLP preprocessing methods for UTD to describe or improve its quality.

Methods

To achieve the research objective, we undertook a scoping review of published literature about NLP preprocessing methods and data quality. The purpose of a scoping review is to map and describe the literature on a new topic or research area and identify key concepts, gaps in the research area, and types and sources of evidence to inform future research [20]. We adopted the Arksey and O'Malley framework [20] for scoping reviews, which has the following steps: 1) define the research question, 2) identify relevant studies, 3) select studies, 4) chart the data, and 5) collate, summarize, and report results.

Search strategy

The search strategy included the concepts of (1) data quality, (2) NLP, and (3) data preprocessing (see Figure 1). The selection of search terms was informed by a systematic review on extracting text from EMRs to improve case detection [21],

a scoping review about quality of routinely-collected electronic health data [22], and keywords related to preprocessing identified from an initial search of the literature [23–25]. We consulted a librarian who assisted in developing and refining the list of search terms. Our initial literature review revealed few articles that included NLP, data quality, and preprocessing in the health science discipline, thus we expanded our search to include relevant literature in all disciplines.

The review included empirical research articles and review articles and was conducted over two time periods. An initial search was executed with an unrestricted minimum date criterion, with an end date of April 15, 2020. We updated the search to the end of May 15, 2021. We searched Scopus, Web of Science, EBSCOhost and ProQuest. In addition, the reference sections of the selected articles were hand searched to identify additional, relevant articles.

Inclusion and exclusion criteria

An article was selected for inclusion if it met one or both of the following criteria: (1) it described research about preprocessing methods for UTD or UTD quality measures or methods, or it was a review article that discussed preprocessing methods to restructure or reorganize UTD for analysis; (2) it was about methods or processes to create a gold standard (or reference) dataset to validate UTD.

An article was excluded if it met one or more of the following criteria: (1) it was about methods for sentiment analysis, ontologies, semantic models, geo-spatial analysis, or qualitative research (e.g., methods to analyze interview or focus group data); (2) it was about methods to construct lexicon databases, dictionaries, or language databases; (3) it focused on the creation of software programs or proprietary solutions for text analysis; (4) it was not available in English; (5) it was an article from the ProQuest database that was neither an empirical article nor a scholarly article.

Article screening

Title and abstract screening were conducted for all articles identified through the implementation of the search strategy, after duplicates were removed. Training was undertaken first for the title, and abstract screening was completed by two authors on a 10% sample of all articles identified from the application of the initial search strategy, to ensure consistency when applying the inclusion and exclusion criteria. Percent agreement and its 95% confidence interval (CI) were calculated. After training was completed, all remaining articles were screened by one author. Rayyan [26], a web application for systematic and scoping reviews, was used to manage and organize articles through the process of title and abstract screening. Differences of opinion on the title and abstract screening were resolved by consensus. Full text screening of all articles selected after title and abstract screening was then conducted, to identify the articles to retain in the scoping review.

Data extraction and analysis

The following types of information were extracted from each article: (1) characteristics of the article, (2) characteristics of the text data, and (3) characteristics of preprocessing methods to restructure or reorganize the UTD. The systematic review conducted by Hinds et al. [22] was used to inform the types of information extracted from the articles, such as the characteristics of articles and validation methods. The reviews conducted by Spasic et al. [6, 28] and Kiefer [27] provided guidance on the characteristics of text data and the preprocessing methods that were included in the data extraction form [6, 27, 28]. We extracted information about the specific dimensions (i.e., types) of data quality that were mentioned in each of the articles. These dimensions were identified from existing data quality frameworks, such as those developed by the Manitoba Centre for Health Policy [7]. Lastly, an inspection of the topics for data quality for UTD that used EMRs were explored. Information extracted included: (1) methods, (2) strengths and limitations, and (3) use cases.

Data extraction training was completed by two authors on a 10% sample of the articles selected for full data extraction. A data dictionary was created to ensure consistency of the data extraction methods. Percent agreement and its 95% CI were calculated. Variables with open-ended responses were excluded from this calculation. Any differences in agreement were resolved by consensus.

Results

The initial search (Figure 1, numbers not in parentheses), which encompassed the period up to April 15, 2020, identified a total of 1134 articles. The updated search (Figure 1, numbers in parentheses) identified an additional 154 articles. Thus, in total 1288 articles were retrieved from the search before duplicates were removed, and 1226 remained after duplicates were removed (Figure 2). Initial training for title and abstract screening yielded 83.3% (95% CI: 75.2%, 89.2%) agreement. After full text screening, a total of 41 articles remained for data extraction (Figure 2).

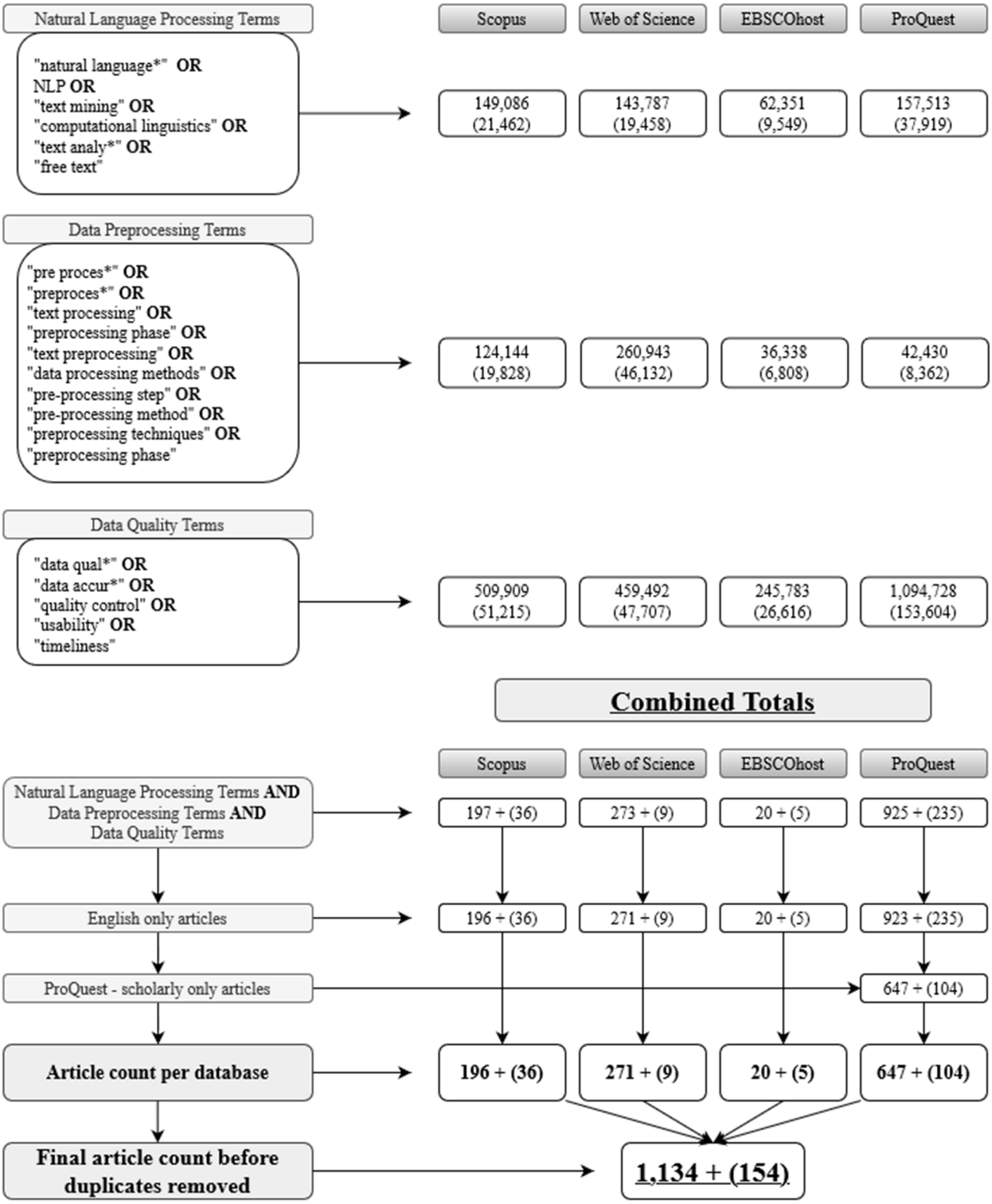
Data extraction results

Ten percent of articles from the initial search were selected for the calculation of agreement for full text extraction; the overall agreement was 94.6% (95% CI: 88.9%, 97.4%). Table 1 summarizes the characteristics of the articles selected for full data extraction.

In total, 90.2% of the articles reported the results of empirical research and another 7.3% ($n = 3$) were review articles. Only one article was classified as a case study. More than half (51.2%) of the articles were published between 2016 and 2021. No articles were published prior to 2002. In terms of disciplinary area, 61% of articles were deemed to be from computer science and engineering disciplines, while 39% were from the health sciences, social sciences, humanities, and business.

A variety of types of text data were represented in the selected articles including EMRs (i.e., clinical notes, progress notes, patient safety records [17, 30–36]), lexical documents (i.e., language treebanks which are bodies of text that have been parsed semantically and syntactically, WordNet database [37–43]), organizational

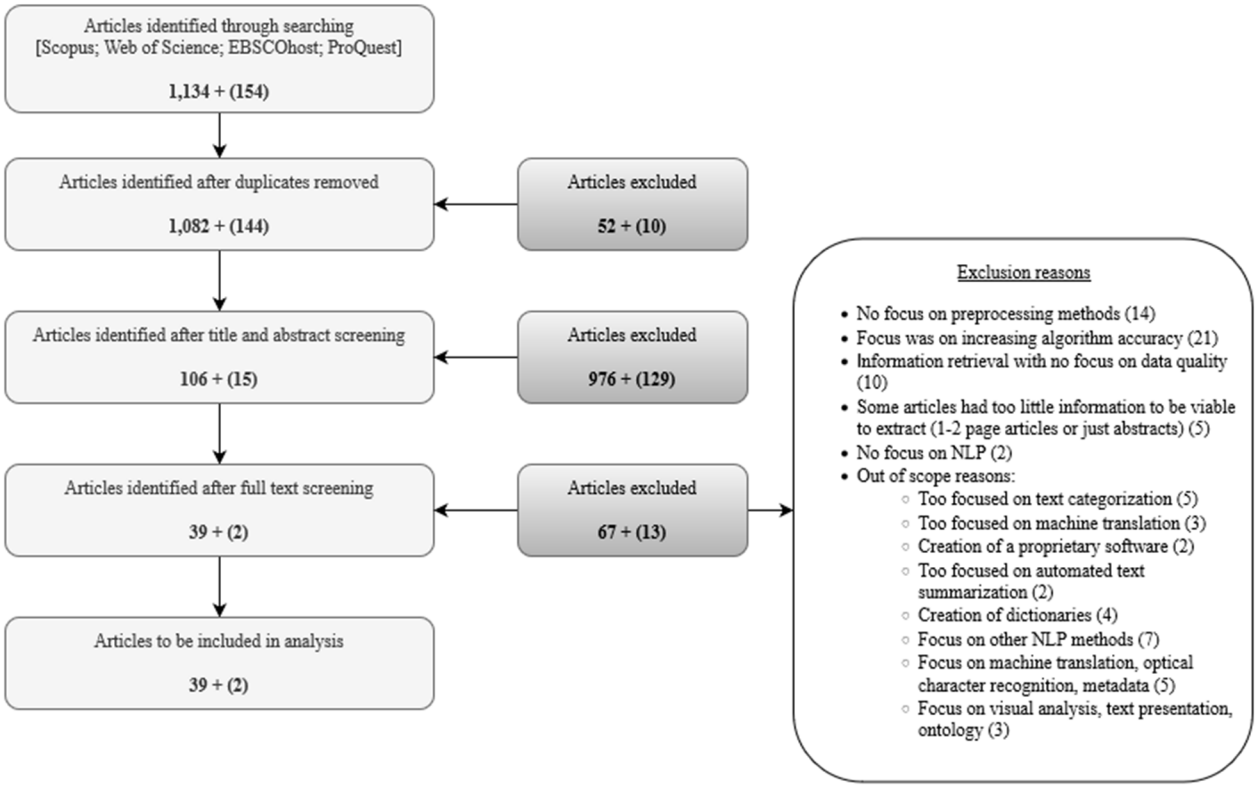
Figure 1: Scoping review search strategy results



Note: Numbers not in parentheses are for the search completed in April 2021; number in parentheses are for the search completed in May 2021.

documents (i.e., maintenance logs/data, accident reports, requirements documentation [44–47]), abstracts and scientific articles (i.e., PubMed and various engineering journals [29, 48–50]), various bodies of text (corpora) (i.e., non-language corpora, non-medical/medical/biomedical corpora, language corpus [50–53]), social media data (i.e., Twitter, meme tracker from various social media websites [54–56]), product reviews (i.e., general product, Chinese tourism, Amazon product [13, 57, 58]), and news articles (i.e., magazines, newswires, consumer reports [54, 59, 60]).

Figure 2: Scoping review PRISMA flowchart



Note: Numbers not in parentheses are for the search completed in April 2021; number in parentheses are for the search completed in May 2021.

Table 1: Characteristics of the scoping review articles ($n = 41$)

	n	%
Article type		
Empirical research	37	90.2
Review article	3	7.3
Case study	1	2.4
Publication year		
2002–2009	7	17.1
2010–2015	13	31.7
2016–2021	21	51.2
Disciplinary area		
Computer Science and Engineering	25	61.0
Health Sciences	8	19.5
Social Sciences, Humanities, Business	8	19.5
Type of data*		
Electronic Medical Records	8	19.5
Lexical (e.g., language treebanks)	8	19.5
Organizational Documents	6	14.6
Product Reviews	4	9.8
News Articles	4	9.8
Corpora (e.g., biomedical text of gene entities)	4	9.8
Abstracts and Articles from Scientific Journals	4	9.8
Social Media	3	7.3
Administrative	1	2.4
Other	5	12.2

*Categories are not mutually exclusive.

Almost all empirical articles (85.4%) described preprocessing methods to improve NLP algorithm performance. However, one article [55] offered an empirical approach to compare a new preprocessing methodology to an existing (i.e., baseline)

Table 2: Characteristics of text data and quality assessment in the scoping review articles ($n = 41$)

	n	%
Measures used to describe data size*		
Documents	22	53.7
Words	19	46.3
Phrases (sentences)	8	19.5
Rows (records)	6	14.6
Features	4	9.8
Data annotation*		
Manual	20	48.8
Automated	17	41.5
No annotation	15	36.6
Preprocessing methods*		
Reorganizing or restructuring methods	35	85.4
Internal validation	23	56.1
External validation	11	26.8
Other	4	9.8
Data quality dimensions mentioned*		
Accuracy	28	68.3
Relevance	14	34.1
Comparability	13	31.7
Usability	7	17.1
Completeness	7	17.1
Validity	7	17.1
Readability	6	14.6
Accessibility	3	7.3
Timeliness	3	7.3

*Categories are not mutually exclusive.

preprocessing method on short text similarity measures; results showed that the newly-proposed method outperformed the existing baseline method [55]. Other empirical articles discussed methods for manual or automated annotation. For example, Westpfahl et al. [53] discussed the creation of a gold standard corpus for teaching and research about spoken German. In this article, the corpus was manually annotated with part of speech tagging (i.e., identifying and annotating words that are nouns, adverbs, verbs, and other parts of speech) and lemmatization, a form of word stemming, where the morphological base of a word is returned [53].

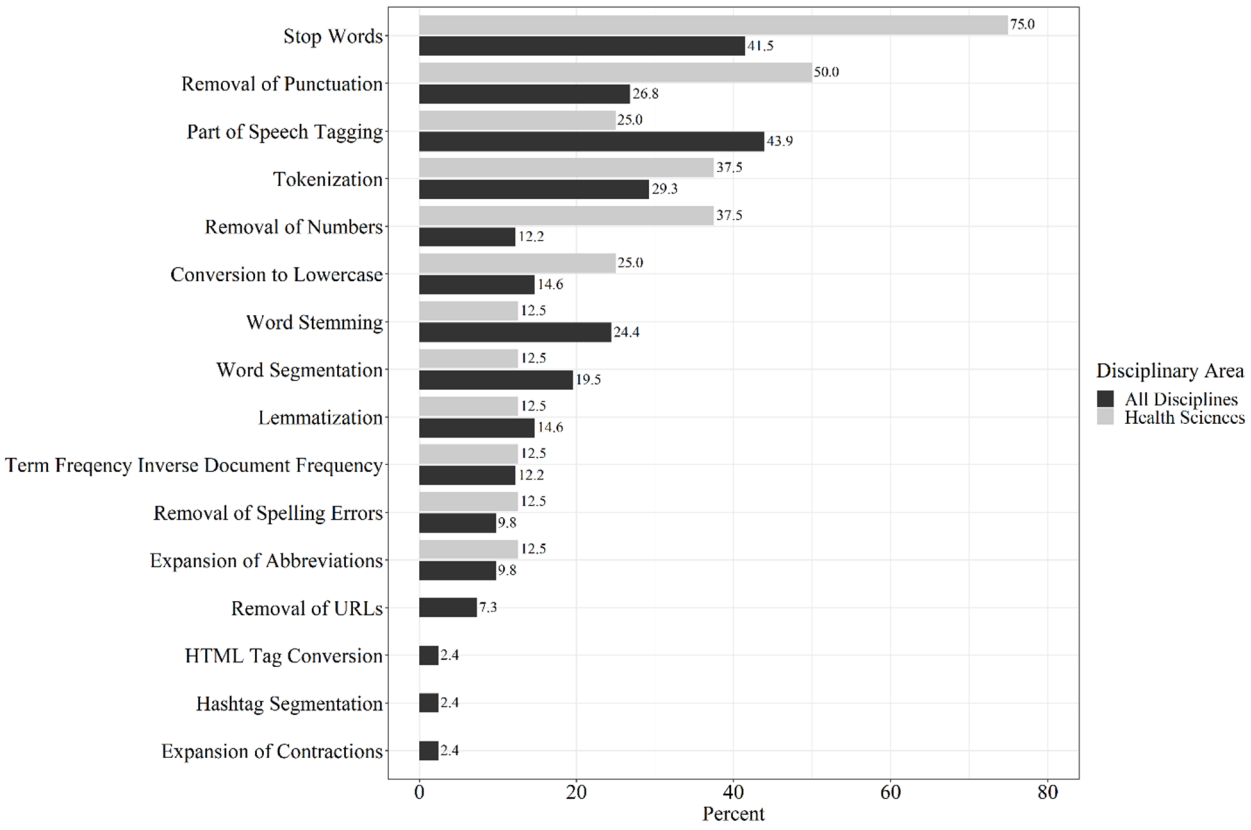
In terms of data size and terms used to describe data size, the UTD described in the articles were characterized in many ways (Table 2). Documents and words were mentioned most frequently (e.g., citing how many documents were used or how many words were found in a type of UTD). More than half (53.7%) of the articles mentioned the volume of documents, and 46.3% of the articles mentioned the count of words in each document. Elements of size often used to describe the size of structured data such as "how many rows (records)" or "how many columns (features)" were among the least common ways to describe UTD size.

Overall, over one third (36.6%) of the articles did not employ text annotation, while 48.8% of the articles discussed manual annotation (e.g., an experienced coder assigning codes using the International Classification of Diseases - Clinical Modification (ICD-9-CM) [61]), and 41.5% of articles discussed automated annotation (e.g., words were automatically annotated for part of speech analyses, grammar tagging, and assistance with manual annotation processes [12, 29, 37, 39, 52, 53, 55, 62]).

Almost all (85.4%) of the articles discussed using restructuring and reorganizing methods to prepare UTD for analysis (e.g., removing stop words, punctuation, removing URLs). Figure 3 describes the types of restructuring and reorganizing methods that were used for all articles and for the subset of articles from health science disciplines.

We identified the articles that explicitly mentioned a data quality dimension (i.e., we counted the words pertaining to data quality). There was no prior determination of the quality dimension criteria when capturing words in relation to data quality. The three data quality dimensions that were most frequently mentioned were: accuracy (68.3%), relevance (34.1%), and comparability (31.7%). Furthermore, "data quality" or "quality" as terms were described or referenced in several ways among the 41 articles. Several articles discussed quality either from the perspective of data quality (or information quality), or using terminology from data or information quality dimensions (e.g., accuracy, correctness, interpretability) [13, 17, 47, 56, 58]. Other articles discussed enhancing data quality by focusing on utilizing or improving preprocessing methods [31, 34–37, 40, 42, 46, 50, 52, 54–56, 63, 64]. Several of these articles only mentioned "data quality" in passing; the main focus of these articles were

Figure 3: Comparison of restructuring and reorganizing methodsfor all articles and for health science articles



the preprocessing methods. Other articles discussed quality from an "annotation" perspective whereby tagging documents or words was intended to enhance the quality of algorithms either through the creation of treebanks or manual annotation using expert opinion [13, 30, 32, 33, 38, 39, 47–49, 51, 53, 59, 65, 66]. Lastly, some articles discussed quality through utilizing preprocessing methods that were primarily focused on achieving a more "accurate or relevant" outcome for algorithmic model performance [29, 41, 43–45, 57, 58, 60, 62, 67, 68]. Overall, while articles mentioned terminology such as "accuracy" or "relevance", data quality as a concept (or its individual dimensions), was referenced or described in multiple ways. Articles that focused exclusively on data quality dimensions or their measurement were rare.

Overall, many of the articles described an improvement of algorithm outcome performance through a variety of preprocessing techniques to enhance UTD quality. However, there were no articles that reported on indicators of UTD quality before preprocessing. While data quality dimensions were not mapped to specific preprocessing methods, "accuracy" was used to evaluate outcomes in a confusion matrix, and it was also used as a descriptor for utilising preprocessing methods to enable unsupervised and supervised algorithms outcomes to be more accurate. That is, all preprocessing methods were closely tied to the data quality dimension of accuracy.

UTD quality topics for EMR data

Seven data quality topics were discussed specifically in reference to EMR data. The topic areas were: (1) misspelled words, (2) security/de-identification, (3) reducing word variability, (4) sources of noise, (5) quality of annotations, (6) ambiguous abbreviations, and (7) reducing manual annotations. Further details such as the methods used, strengths and limitations of the methods, and the data used (i.e., use case) are provided in Supplementary Appendix 1.

Several of these quality topics focused on the reduction of possible error or variability (e.g., correct misspelled words, reduction of word variability, addressing sources of noise, or addressing ambiguity in abbreviations). Two articles reported on the assessment of misspelled words. Some of the methods utilized were a combination of rule-based approaches or machine learning approaches such as dictionaries, regular expressions, a string-to-string edit distance known as the Levenshtein-Damerau distance, and word sense disambiguation for words with similar parts of speech (e.g., distinguishing between two similar words that are classified as verbs). One article by Assale et al. [31], was about reducing word variability in typographical corrections, where typographical errors in one word can create variability in how the word is spelled. Rule-based approaches were used in this article such as preprocessing methods to restructure and reorganize UTD (e.g., removal of stop words), counting word frequencies, and utilizing the Levenshtein-Damerau distance metric. One article addressed sources of noise where the removal of noise in text involves preprocessing methods that restructure and reorganize UTD. Some of these methods include (but are not limited to): tokenization, converting uppercase letters to lowercase letters, and removal of stop words. Lastly, ambiguity in abbreviations was addressed in one article where the authors used deep learning methods

(i.e., convolutional neural networks); no feature engineering or preprocessing was required. The convolutional neural network was trained on word embeddings, which are representations of words in a list. These word embeddings were extracted from journal articles found in PubMed.

Other topics focused on the evaluation or reduction of manual efforts such as assessing the quality of annotations or solutions involving reducing manual annotation of text data. Two articles focused on the quality of annotations. One focused on manual annotation of clinical notes for fall-related injuries. In the second article, tags for parts of speech and named entities were applied by annotators with backgrounds in computational linguistics, while physicians in training solved any disagreements between annotators.

Discussion

The scoping review has documented practices for preprocessing UTD to describe or improve its quality. Few articles in the scoping review discussed the quality of UTD before analysis or preprocessing was initiated. The main topics raised in the selected studies were about the challenges of defining data quality, the choice of data quality assessments for UTD and how this is influenced by the context of the text data, and differences in the data quality challenges associated with EMR UTD when compared to other types of UTD.

The scoping review reveals the following key points: 1) The most common preprocessing methods used in health science articles were different from the most common preprocessing methods used in all disciplines combined. To elaborate, the most common preprocessing methods for health science articles included removal of stop words, removal of punctuation, removal of numbers, tokenization, parts of speech tagging, and converting characters to lowercase. 2) Few dimensions of data quality were considered in assessed UTD. Accuracy, relevance, and comparability were the most commonly-reported dimensions. 3) Quality indicator topics addressed potential challenges in preprocessing, such as the quality of annotations, presence of spelling errors, and presence of ambiguous abbreviations.

One difficulty with describing the quality of UTD is the lack of standardized terminology. Strong et al. [69] differentiated "information quality" from "data quality" in terms of its specific goals; information quality is about assessing the needs of information users while data quality refers to the fitness of data for its intended use. However, Chen and Tseng [58] did not make that distinction; their indicators of information quality are similar to those found in existing data quality frameworks.

Amongst the data quality measures identified from the scoping review, some were tailored specifically for the data being assessed. This emphasizes that quality of UTD depends on the context or type of data that is being assessed. Language usage must be contextual to the environment (i.e., vernacular used in product reviews differ from vernacular used in EMR notes). For example, several of the data quality measures for UTD that Chen and Tseng's article reference do not necessarily apply to data in EMRs such as the data quality dimension for "objectivity", and assessments of whether a product review is an opinion rather than factual [58].

Some of the challenges with quality of UTD in EMRs are different than the challenges associated with UTD found in social media data or organizational reports. One of the biggest challenges of the former is ensuring privacy, anonymity, and confidentiality of patient data. Pantazos et al. [17] stated that as UTD in EMRs increase in usage, it must achieve two goals (1) that it is de-identified and anonymized and (2) EMRs that are de-identified must contain accurate information about the un-identified patient and be coherent to the reader. The scoping review revealed that high-quality UTD within EMRs must be readable, correct, and consistent with a patient's record.

This research has some limitations. The scoping review was restricted to English language articles and grey literature was not searched. Accordingly, the scoping review may not represent all published articles about UTD quality. Furthermore, articles that discussed preprocessing methods to improve algorithmic modelling outcomes were not included. It was not feasible to address all different modelling techniques for text data (i.e., collect all articles that conducted a sentiment analysis or other methods). The articles selected were those that focused on quality of text data, or that focused on preprocessing methods and also mentioned data quality. Choosing key words for the scoping review search strategy was challenging. This was due in large part to the lack of standardized terminology and the diverse terminology within the NLP and data quality literatures and may have resulted in some articles being missed.

Despite these limitations, the major strength of this research is that it used a systematic approach to examine preprocessing methods to describe data quality for UTD. Data quality for UTD is an important area for research in multiple fields, including in the health sciences. Written language in EMRs and other patient-related documents is different from other types of text data found in textbooks or social media due to its nature in short text and point form.

Several opportunities for future research exist. First, a scoping review could be conducted to identify operational definitions for data quality dimensions specific to text data from routinely collected electronic health data. Akin to Weiskopf et al.'s scoping review to discover methods and dimensions [70], the scoping review could be used to develop definitions for dimensions of data quality for UTD. This research has shown that there are unique characteristics of text data that are not present in numerical data (e.g., grammar rules or punctuation). Thus, operational data quality definitions that specifically address text data properties is an important step towards structuring a quality framework for text data. Second, a documentation project that maps operational definitions for data quality dimensions to the data quality indicators for routinely collected electronic health data should be explored [70]. Third, a case study could be undertaken for several text data quality indicator topics identified from this research (e.g., disambiguation of abbreviations). While there have been studies to disambiguate abbreviations [34] and correct spelling/typographical errors [31, 35], it would also be beneficial to identify the impact of word variability on algorithm outputs in text data. Lastly, additional scoping or systematic reviews could be conducted. In particular, quality

indicators for short text documents such as social media posts, reviews, and EMRs, could be explored [13, 54–56, 58, 71]. Since EMRs are characterized by short texts, it would be interesting to examine other text quality indicators appropriate for these types of data.

Conclusion

Data quality is a multidimensional construct that depends on the context in which the data will be used. However, there are many similarities between the dimensions of data quality for structured and unstructured text and methods to assess data quality. Assessing data quality in UTD often requires access to specialized gold standard datasets or dictionaries. However, there are a few general-purpose measures of data quality that do not require external data; most of these focus on the measurement of noise in the data.

Acknowledgements

MN was supported by the Visual and Automated Disease Analytics program during his research program. LML is supported by a Tier I Canada Research Chair in Methods for Electronic Health Data Quality. This research was supported by the Canada Research Chairs program.

Statement on conflicts of interest

The authors declare that they have no conflicts to report.

Ethics statement

Ethical approval was not required because this research did not use patient data.

Supplementary appendices

Supplementary Appendix 1 present data quality topics for UTD found in articles that used EMR data. The following information was collected: 1) the quality topic, 2) author of publication, 3) the methods used, 4) the strengths and limitations of the method, and 5) the use case that describes what data the authors used.

References

1. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med [Internet]. 2015;12(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4595218/.

2. Nicholls SG, Langan SM, Benchimol EI. Routinely collected data: the importance of high-quality diagnostic coding to research. Can Med Assoc J [Internet]. 2017;189(33):E1054–5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5566604/.

3. Smith M, Lix LM, Azimaee M, Enns JE, Orr J, Hong S, Roos LL. Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. J Am Med Inform Assoc [Internet]. 2018 [cited 2018 Nov 18];25(3):224–9. Available from: http://academic.oup.com/jamia/article/25/3/224/4102320.

4. Agboola Y, Camara Y, Brown A, Zhao H, Kendall O. The PHAC Data Quality Framework: A useful tool to serve surveillance programs. In Canada: Public Health Agency of Canada; 2011. Available from: https://www.researchgate.net/publication/305469008_The_PHAC_Data_Quality_Framework_A_Useful_Tool_to_Serve_Surveillance_Programs.

5. Canadian Institute for Health Information. CIHI's information quality framework [Internet]. Canadian Institute for Health Information; 2017. Available from: https://www.cihi.ca/sites/default/files/document/iqf-summary-july-26-2017-en-web_0.pdf.

6. Kiefer C. Assessing the quality of unstructured data: An initial overview. In: LWDA [Internet]. 2016. p. 62–73. Available from: https://www.semanticscholar.org/paper/Assessing-the-Quality-of-Unstructured-Data%3A-An-Kiefer/a7d31b09498a16201f7044eb77ff14d27c1c559b.

7. Azimaee M, Smith M, Lix L, Ostapyk T, Burchill C, Orr J. MCHP data quality framework [Internet]. Winnipeg Manitoba Canada University of Manitoba MCHP; 2018. Available from: http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/protocol/media/Data_Quality_Framework.pdf.

8. Government of Canada SC. Statistics Canada's quality assurance framework, 2017 [Internet]. Statistics Canada; 2017. Available from: https://www150.statcan.gc.ca/n1/pub/12-586-x/12-586-x2017001-eng.pdf.

9. Sonntag D. Assessing the quality of natural language text data. In Gesellschaft für Informatik e.V.; 2004. p. 259–63. Available from: https://dl.gi.de/handle/20.500.12116/28866.

10. Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: an experimental study. In: Proceedings of the 2011 conference on empirical methods in natural language processing [Internet]. Association for Computational Linguistics; 2011. p. 1524–34. Available from: https://aclanthology.org/D11-1141.

11. Kee YH, Li C, Kong L, Jieyi Tang C, Chuang KL. Scoping review of mindfulness research: a topic modelling approach. Mindfulness [Internet].

2019;10:1474–88. Available from: https://link-springer-com.uml.idm.oclc.org/article/10.1007%2Fs12671-019-01136-4.

12. Briesch D, Hobbs R, Jaja C, Kjersten B, Voss C. Training and evaluating a statistical part of speech tagger for natural language applications using Kepler Workflows. Procedia Comput Sci [Internet]. 2012 [cited 2020 Jun 5];9:1588–94. Available from: http://www.sciencedirect.com/science/article/pii/S1877050912002955.

13. Yang HL, Chao AFY. Sentiment annotations for reviews: an information quality perspective. Online Inf Rev [Internet]. 2018 [cited 2020 Jun 5];42(5):579–94. Available from: https://doi.org/10.1108/OIR-04-2017-0114.

14. Subha R, Palaniswami S. Quality factor assessment and text summarization of unambiguous natural language requirements. In: Unnikrishnan S, Surve S, Bhoir D, editors. Advances in Computing, Communication, and Control [Internet]. Berlin, Heidelberg: Springer; 2013. p. 131–46. Available from: https://doi-org.uml.idm.oclc.org/10.1007/978-3-642-36321-4_12.

15. Zennaki O, Semmar N, Besacier L. Unsupervised and lightly supervised part-of-speech tagging using recurrent neural networks. In: 29th Pacific Asia Conference on Language, Information and Computation (PACLIC) [Internet]. Shangai, China; 2015 [cited 2021 Jul 18]. Available from: https://hal.archives-ouvertes.fr/hal-01350113.

16. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. Journal of Medical Ethics [Internet]. 2020 Jun 1 [cited 2022 Jul 5];46(6):367–77. Available from: https://jme.bmj.com/content/46/6/367.

17. Pantazos K, Lauesen S, Lippert S. Preserving medical correctness, readability and consistency in de-identified health records. Health Informatics J [Internet]. 2017 Dec 1 [cited 2020 Jun 6];23(4):291–303. Available from: https://doi.org/10.1177/1460458216647760.

18. Baclic O, Tunis M, Young K, Doan C, Swerdfeger H, Schonfeld J. Challenges and opportunities for public health made possible by advances in natural language processing. Can Commun Dis Rep [Internet]. 2020 Jun 4 [cited 2022 Jul 5];46(6):161–8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7343054/.

19. van der Aa H, Carmona Vargas J, Leopold H, Mendling J, Padró L. Challenges and opportunities of applying natural language processing in business process management. In Association for Computational Linguistics; 2018 [cited 2022 Jul 5]. p. 2791–801. Available from: https://upcommons.upc.edu/handle/2117/121682.

20. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol [Internet]. 2005;8(1):19–32. Available from: https://doi.org/10.1080/1364557032000119616.

21. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc [Internet]. 2016 [cited 2019 May 21];23(5):1007–15. Available from: http://academic.oup.com/jamia/article/23/5/1007/2379833.

22. Hinds A, Lix LM, Smith M, Quan H, Sanmartin C. Quality of administrative health databases in Canada: A scoping review. Can J Public Health [Internet]. 2016;107(1):e56-61. Available from: https://link-springer-com.uml.idm.oclc.org/article/10.17269/cjph.107.5244.

23. Dařena F, Žižka J. Interdependence of text mining quality and the input data preprocessing. In: Silhavy R, Senkerik R, Oplatkova ZK, Prokopova Z, Silhavy P, editors. Artificial Intelligence Perspectives and Applications [Internet]. Cham: Springer International Publishing; 2015. p. 141–50. (Advances in Intelligent Systems and Computing). Available from: https://link-springer-com.uml.idm.oclc.org/chapter/10.1007/978-3-319-18476-0_15#citeas.

24. Vijayarani M. Preprocessing techniques for text mining - an overview. Int J Comput Netw Commun [Internet]. 2015;5(1):7–16. Available from: https://www.researchgate.net/profile/Vijayarani-Mohan/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview/links/5e57a0f7299bf1bdb83e7505/Preprocessing-Techniques-for-Text-Mining-An-Overview.pdf.

25. Malak P. Text preprocessing: a tool of information visualization and digital humanities [Internet]. Information Visualization Techniques in the Social Sciences and Humanities. IGI Global; 2018 [cited 2020 Jun 6]. p. 86–104. Available from: http://www.igi-global.com/chapter/text-preprocessing/201305.

26. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews [Internet]. 2016;5(1):210. Available from: https://doi.org/10.1186/s13643-016-0384-4.

27. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform [Internet]. 2020;8(3):e17984. Available from: https://medinform.jmir.org/2020/3/e17984/.

28. Kiefer C. Quality indicators for text data. In: BTW 2019–Workshopband [Internet]. Gesellschaft für Informatik, Bonn; 2019. p. 145–54. Available from: https://dl.gi.de/handle/20.500.12116/21801.

29. Gero Z, Ho J. PMCVec: Distributed phrase representation for biomedical text processing. J Biomed Inform [Internet]. 2019 [cited 2020 Jun 5];100:100047. Available from:

http://www.sciencedirect.com/science/article/pii/S2590177X19300460.

30. Berndt DJ, McCart JA, Finch DK, Luther SL. A case study of data quality in text mining clinical progress notes. ACM Trans Manag Inf Syst [Internet]. 2015 [cited 2020 Jun 5];6(1):1:1-1:21. Available from: https://doi.org/10.1145/2669368.

31. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The revival of the notes field: leveraging the unstructured content in electronic health records. Front Med [Internet]. 2019;6:66. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2019.00066/full.

32. He B, Dong B, Guan Y, Yang J, Jiang Z, Yu Q, Cheng J, Qu C. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. J Biomed Inform [Internet]. 2017 [cited 2020 Jun 5];69:203–17. Available from: http://www.sciencedirect.com/science/article/pii/S153204641730076X.

33. Liang C, Gong Y, Christian N., Kuziemsky C.E., Kushniruk A.W., Borycki E.M. Enhancing patient safety event reporting by K-nearest neighbor classifier. Stud Health Technol Informatics [Internet]. 2015;218:93–9. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84951923066&doi=10.3233%2f978-1-61499-574-6-93&partnerID=40&md5=9dd7ea3d83936147533292dec0f1447b.

34. Joopudi V, Dandala B, Devarakonda M. A convolutional route to abbreviation disambiguation in clinical text. J Biomed Inform. 2018;86:71–8. Available from: https://www-sciencedirect-com.uml.idm.oclc.org/science/article/pii/S1532046418301552.

35. Lai KH, Topaz M, Goss FR, Zhou L. Automated misspelling detection and correction in clinical free-text records. J Biomed Inform [Internet]. 2015 [cited 2019 Aug 3];55:188–95. Available from: http://www.sciencedirect.com/science/article/pii/S1532046415000751.

36. Ruch P, Baud R, Geissbühler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artificial Intelligence in Medicine [Internet]. 2003 Sep 1 [cited 2020 Jun 6];29(1):169–84. Available from: http://www.sciencedirect.com/science/article/pii/S0933365703000526.

37. Hoste V, Hendrickx I, Daelemans W, Bosch AVD. Parameter optimization for machine-learning of word sense disambiguation. Natural Language Engineering [Internet]. 2002 Dec [cited 2020 Jun 6];8(4):311–25. Available from: http://www.cambridge.org/core/journals/natural-language-engineering/article/parameter-optimization-for-machinelearning-of-word-sense-disambiguation/ED5DAA7F32FEFBC5D1C81B22DEA5F43B#.

38. Nguyen QT, Miyao Y, Le HTT, Nguyen NTH. Ensuring annotation consistency and accuracy for Vietnamese treebank. Lang Resour Eval [Internet]. 2018 [cited 2020 Jun 6];52(1):269–315. Available from: https://doi-org.uml.idm.oclc.org/10.1007/s10579-017-9398-3.

39. Nguyen PT, Le AC, Ho TB, Nguyen VH. Vietnamese treebank construction and entropy-based error detection. Lang Resour Eval [Internet]. 2015 [cited 2020 Jun 6];49(3):487–519. Available from: https://doi-org.uml.idm.oclc.org/10.1007/s10579-015-9308-5.

40. Qin kan, Yujiu Yang, Wenhuang Liu, Xiaodong Liu. An integrated approach for detecting approximate duplicate records. In: 2009 Asia-Pacific Conference on Computational Intelligence and Industrial Applications (PACIIA). 2009. p. 381–4. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/5406409.

41. Uejima H, Miura T, Shioya I. Improving text categorization by resolving semantic ambiguity. In: 2003 IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PACRIM 2003) (Cat No03CH37490) [Internet]. IEEE; 2003. p. 796–9. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/abstract/document/1235901.

42. Zurini M. Word sense disambiguation using aggregated similarity based on WordNet graph representation. Inform econ [Internet]. 2013 [cited 2020 Jun 6];17(3):169–80. Available from: https://ideas.repec.org/a/aes/infoec/v17y2013i3p169-180.html.

43. Šnajder J. DerivBase.hr: A high-coverage derivational morphology resource for croatian. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) [Internet]. Reykjavik, Iceland: European Language Resources Association (ELRA); 2014 [cited 2020 Jun 6]. p. 3371–7. Available from: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1090_Paper.pdf.

44. Edwards B, Zatorsky M, Nayak R. Clustering and classification of maintenance logs using text data mining. In: Conf Res Pract Inf Technol Ser [Internet]. CRC for Integrated Engineering Asset Management, Faculty of Information Technology, Queensland University of Technology, PO Box 2434, Brisbane 4001, QLD, Australia: AusDM; 2008. p. 193–9. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84870484659&partnerID=40&md5=f34d58eba076a357439884fae90dc336.

45. Marzec M, Uhl T, Michalak D. Verification of text mining techniques accuracy when dealing with urban buses maintenance data. Diagnostyka [Internet]. 2017

Dec 20 [cited 2020 Jun 6];15(3):51–7. Available from: http://www.diagnostyka.net.pl/Verification-of-text-mining-techniques-accuracy-when-dealing-with-urban-buses-maintenance,81413,0,2.html.

46. Stewart M, Liu W, Cardell-Oliver R, Wang R. Short-Text Lexical Normalisation on Industrial Log Data. In: 2018 IEEE International Conference on Big Knowledge (ICBK). 2018. p. 113–22. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8588782.

47. Subha R, Palaniswami S. Quality Factor Assessment and Text Summarization of Unambiguous Natural Language Requirements. In: Unnikrishnan S, Surve S, Bhoir D, editors. Advances in Computing, Communication, and Control. Berlin, Heidelberg: Springer; 2013. p. 131–46. Available from: https://doi-org.uml.idm.oclc.org/10.1007/978-3-642-36321-4_12.

48. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform [Internet]. 2014;47:1–10. Available from: https://www-sciencedirect-com.uml.idm.oclc.org/science/article/pii/S1532046413001974?via%3Dihub.

49. Jung Y. A semantic annotation framework for scientific publications. Qual Quant [Internet]. 2017 [cited 2020 Jun 6];51(3):1009–25. Available from: https://doi.org/10.1007/s11135-016-0369-3.

50. Kang N, van Mulligen EM, Kors JA. Comparing and combining chunkers of biomedical text. Journal of Biomedical Informatics [Internet]. 2011 Apr 1 [cited 2020 Jun 6];44(2):354–60. Available from: http://www.sciencedirect.com/science/article/pii/S1532046410001577.

51. Tissot H, Del Fabro MD, Derczynski L, Roberts A. Normalisation of imprecise temporal expressions extracted from text. Knowl Inf Syst [Internet]. 2019 Dec 1 [cited 2020 Jun 6];61(3):1361–94. Available from: https://doi.org/10.1007/s10115-019-01338-1.

52. Vítovec P, Kléma J. Gene interaction extraction from biomedical texts by sentence skeletonization. CEUR Workshop Proc [Internet]. 2011;802:NA. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84891772107&partnerID=40&md5=4c64e3778af84417d329e274327ffff7.

53. Westpfahl S, Schmidt T. FOLK-Gold - A Gold Standard for Part-of-Speech-Tagging of Spoken German. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) [Internet]. Portorož, Slovenia: European Language Resources Association (ELRA); 2016 [cited 2020 Jun 6]. p. 1493–9. Available from: https://www.aclweb.org/anthology/L16-1237.

54. Allen TT, Sui Z, Akbari K. Exploratory text data analysis for quality hypothesis generation. Qual Eng [Internet]. 2018 [cited 2020 Jun 5];30(4):701–12. Available from: https://doi.org/10.1080/08982112.2018.1481216.

55. Alnajran N, Crockett K, McLean D, Latham A. A heuristic based pre-processing methodology for short text similarity measures in microblogs. In: Proc - Int Conf High Perform Comput Commun, Int Conf Smart City Int Conf Data Sci Syst, HPCC/SmartCity/DSS [Internet]. Institute of Electrical and Electronics Engineers Inc.; 2019. p. 1627–33. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8623003.

56. Christen P, Gayler RW, Tran KN, Fisher J, Vatsalan D. Automatic discovery of abnormal values in large textual databases. J Data Inf Qual [Internet]. 2016;7(1–2):1–31. Available from: https://doi.org/10.1145/2889311.

57. Gharatkar S, Ingle A, Naik T, Save A. Review preprocessing using data cleaning and stemming technique. In: 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). 2017. p. 1–4. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8276011.

58. Chen CC, Tseng YD. Quality evaluation of product reviews using an information quality framework. Decis Support Syst [Internet]. 2011 [cited 2020 Jun 5];50(4):755–68. Available from: http://www.sciencedirect.com/science/article/pii/S0167923610001478.

59. Xue N, Xia F, Chiou F dong, Palmer M. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Nat Lang Eng [Internet]. 2005 Jun 1 [cited 2020 Jun 6];11(2):207–38. Available from: https://doi.org/10.1017/S135132490400364X.

60. Scharkow M. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Qual Quant [Internet]. 2013 [cited 2020 Jun 6];47(2):761–73. Available from: https://doi.org/10.1007/s11135-011-9545-7.

61. Lauría EJM, March AD. Effect of dirty data on free text discharge diagnoses used for automated ICD-9-CM coding. In: AMCIS [Internet]. 2006. Available from: https://aisel.aisnet.org/amcis2006/188.

62. Abad ZSH, Karras O, Ghazi P, Glinz M, Ruhe G, Schneider K. What works better? A study of classifying requirements. In: 2017 IEEE 25th International Requirements Engineering Conference [Internet]. 2017. p. 496–501. Available from: https://arxiv.org/abs/1707.02358.

63. HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. PLOS ONE [Internet]. 2020 May 1 [cited 2021 Feb 21];15(5):e0232525. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232525.

64. Song JW, Chung KC. Observational Studies: Cohort and Case-Control Studies. Plast Reconstr Surg [Internet]. 2010 Dec [cited 2019 Nov 5];126(6):2234–42. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2998589/.

65. Moreno I, Boldrini E, Moreda P, Romá-Ferri MT. DrugSemantics: A corpus for named entity recognition in Spanish summaries of product characteristics. J Biomed Inform [Internet]. 2017 [cited 2020 Jun 6];72:8–22. Available from: https://doi.org/10.1016/j.jbi.2017.06.013.

66. Song B, Wu P, Zhang Q, Chai B, Gao Y, Yang H. Intelligent assessment of 95598 speech transcription text quality based on topic model. IOP Conf Ser: Mater Sci Eng [Internet]. 2019 [cited 2020 Jun 6];563(4):2001. Available from: https://doi.org/10.1088%2F1757-899x%2F563%2F4%2F042001.

67. Rianto, Mutiara Achmad Benny, Wibowo Eri Prasetyo, Santosa PI. Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. J Big Data [Internet]. 2021;8(1):26. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00413-1.

68. Lauría EJM, March AD. Combining bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis. ACM J Data Inf Qual [Internet]. 2011 [cited 2020 Jun 6];2(3):13:1-13:22. Available from: https://doi.org/10.1145/2063504.2063506.

69. Strong DM. Information quality: managing information as a product. In: LIU L, ÖZSU MT, editors. Encyclopedia of Database Systems [Internet]. Boston, MA: Springer US; 2009 [cited 2021 Jun 12]. p. 1502–8. Available from: https://doi.org/10.1007/978-0-387-39940-9_497.

70. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc [Internet]. 2013;20(1):144–51. Available from: https://doi-org.uml.idm.oclc.org/10.1136/amiajnl-2011-000681.

71. Pantazos K, Lauesen S, Lippert S. De-identifying an EHR database-Anonymity, correctness and readability of the medical record [Internet]. Stud. Health Technol. Informatics. IOS Press; 2011. 862–866 p. (Health Technol. Informatics; vol. 169). Available from: https://ebooks.iospress.nl/publication/14293.

Abbreviations

UTD:	Unstructured Text Data
EMR:	Electronic Medical Records
CI:	Confidence Interval
NLP:	Natural Language Processing
PRISMA:	Preferred Reporting Items for Systematic Reviews and Meta-Analysis

Supplementary material

Supplementary Appendix 1: UTD restructuring and reorganizing data quality topics described in scoping review articles about EMR data

Data quality topic	Source	Methods	Strengths/Limitations	Use case
Misspelled words	Lai et al	The authors measured spelling errors for clinical and general terms. A clinical term was a word that was not a name or was not present in the Aspells [31] default dictionary. A general term was any other word that was not a clinical term. These methods were used to identify both clinical and general terms: 1) used many dictionaries for misspelling detection 2) used Named Entity Recognition (NER) to avoid misclassification of person names as misspellings 3) used regular expressions to correct emails, URLs, and commonly misspelled words	**Strengths:** A combination of previous methods that worked well. **Limitations:** 1) Since it's a composite of previous methods so its incomparable to previous methods. 2) Authors did not attempt to investigate whether misspellings & misspelling corrections impacted the quality of the information extracted from medical texts.	Applied to: 1) clinical notes from primary care clinics 2) free-text allergy entries 3) free-text medication orders
	Ruch et al	Methods used to improve spelling correction: module 1) string to string edit distance known as Damerau-Levenshtein module 2) syntactic correction – to address word order and agreement problems module 3) processing words with the same parts of speech by applying contextual word sense disambiguation	**Strengths:** NLP tools such as NER and lexical disambiguation can reduce spelling corrections, this can allow for automated processes. **Limitations:** Authors did not attempt to investigate whether misspellings & misspelling corrections impacted on the quality of the information extracted from medical texts.	Applied to electronic patient records
Security	Pantazos et al	The methods used for de-identification has to meet the criteria of: 1) Medical correctness - each health record must show a true medical picture of a patient 2) Anonymity - it's not possible to see who the real patient is 3) Readability - the health record has to represent reality 4) Consistency - the patient's identifiers have to be consistent with the entire medical picture Each method utilized a mapping table to replace existing identifiers with new identifiers. To treat ambiguity, when the de-identification program meets an ambiguous word it: 1) gets deleted if it's a rare word (occurs less than 200 times) 2) gets left if it occurs more than 200 times permutation tables - that mapped existing identifiers with new ones - ensured readability and consistency distorted identifier tables - mapped existing civil registration numbers to a scrambled one - ensured readability and correctness	**Strengths:** While de-identifying EMR data, the authors tried to preserve essential components of the note to mimic the original note. **Limitations:** 1) Did not address spelling errors. 2) Overlooked pharmaceutical names. 3) Missed several clinical abbreviations. 4) Cannot do analysis on geography due to scrambled zip codes. 5) For statistical purposes de-identification should not impact readability or consistency.	Applied to de-identify an existing EMR database
Reducing word variability	Assale et al	1) Tokenization, removal of stop words, numbers and non-ASCII symbols 2) Counted frequency of words - words occurring above an 80% percentage were considered correct, less frequent words were checked if they were present in an Italian dictionary or a medical dictionary, whatever is left over was considered typographical. 3) Used distance metric between strings - "Levenshtein distance". Search in the 80% of most frequent words that had a "distance 1" from the typos. Distance 1 signifies that typological words differ from 1 letter insertion/deletion/substitution from the original word.	**Strengths:** Proposed a method for reducing word variability **Limitations:** 1) Number of false positives is high. 2) Cannot guarantee to correct all errors.	Applied to anamnestic summaries of endocrinology and rheumatology

Continued

Supplementary Appendix 1: Continued

Data quality topic	Source	Methods	Strengths/Limitations	Use case
		4) Also take into account "Damerau-Levenshtein distance" metric where it also takes into account inversions between letters - because its common to invert two adjacent letters (distance 2). 5) Manually inspected to verify no association errors. Ambiguous associations were discarded. 6) Multi-associated words (i.e., words with the same meaning but varied in spelling) were replaced with the most frequent one.		
Sources of noise	Berndt et al	1) converting text to lowercase 2) tokenization 3) removal of tokens with less than three characters or no alphabetical characters 4) normalizing terms 5) removal of stop words 6) removing tokens that only occur once	**Strengths:** Authors addressed text noise by introducing preprocessing methods to reduce noise **Limitations:** The study only used one dataset	Applied to clinical progress notes
Quality of annotations	He et al	Annotation methods included: 1) word segmentation 2) part of speech tagging (with shallow and full parsing of parts of speech tags) 3) named entity tagging 4) relational tagging (i.e., finding relationships among named entities) Annotation quality was evaluated using F1 measure, precision, and recall	**Strengths:** Authors have built a concept of data quality into the creation of a corpus by assessing the quality of annotations **Limitations:** Since there was a limit of annotation resources, the corpus created only covered two departments of a hospital, thus lacking medical terminology in other departments	Applied to: 1) discharge summaries 2) clinical progress notes
	Berndt et al	Quality of Annotations include: 1) "Fall" or "Not fall" was given an operational definition 2) Manual annotation evaluated with Cohen's kappa for interrater agreement	**Strengths:** Generally manual annotations are used as a reference standard to be used for machine learning classification algorithms **Limitations:** Manual annotations are time consuming, costly, and can lead to manual error	Applied to clinical progress notes
Ambiguous abbreviations	Joopudi et al	Utilized a convolutional neural network (CNN) that: 1) Was trained on word embeddings (i.e., a representation of words in a vector) located on journal articles in PubMed 2) Was trained on parts of speech tags 3) Used clinical notes meta information such as author	**Strengths:** Using deep learning allows the user to bypass feature engineering tasks which can be time consuming, furthermore the CNN model worked well on disambiguating abbreviations **Limitations:** While the authors have shown that methods used to disambiguate abbreviations works well, there is no investigation if the corrections have had an impact on the quality of information.	Applied to: 1) de-identified longitudinal patient records with clinical notes 3) publicly available dataset created by University of Minnesota
Reducing manual annotation	Liang et al	Utilized KNN classifier (supervised method) to predict the type of documents (i.e., a document is either "diagnostic errors" or "device related complications"). The process was as follows: 1) Noise Removal: removal of punctuations, words were set to lowercase, white space was removed, stop words were removed 2) Document was converted to a document term matrix 3) Documents were pre-annotated as either "diagnostic errors" or "device related complications" to create a gold standard comparison 4) Evaluated using F measure and Accuracy	**Strengths:** Demonstrated a process that enhances automatic annotation processes that include data quality elements. **Limitations:** The sample size in the study was small and this method was not demonstrated on other types of data such as EMR clinical notes.	Applied to publicly available patient safety documents from WebM&M