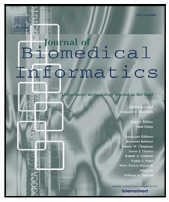
Journal of Biomedical Informatics 140 (2023) 104336

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Sentiment analysis of clinical narratives: A scoping review

Kerstin Denecke *, Daniel Reichenpfader

Bern University of Applied Sciences, Institute for Medical Informatics, Quellgasse 21, Biel/Bienne, 2502, Bern, Switzerland

ARTICLE INFO

Keywords:
Sentiment analysis
Clinical notes
Scoping review
NLP

ABSTRACT

A clinical sentiment is a judgment, thought or attitude promoted by an observation with respect to the health of an individual. Sentiment analysis has drawn attention in the healthcare domain for secondary use of data from clinical narratives, with a variety of applications including predicting the likelihood of emerging mental illnesses or clinical outcomes. The current state of research has not yet been summarized. This study presents results from a scoping review aiming at providing an overview of sentiment analysis of clinical narratives in order to summarize existing research and identify open research gaps. The scoping review was carried out in line with the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guideline. Studies were identified by searching 4 electronic databases (e.g., PubMed, IEEE Xplore) in addition to conducting backward and forward reference list checking of the included studies. We extracted information on use cases, methods and tools applied, used datasets and performance of the sentiment analysis approach. Of 1,200 citations retrieved, 29 unique studies were included in the review covering a period of 8 years. Most studies apply general domain tools (e.g. TextBlob) and sentiment lexicons (e.g. SentiWordNet) for realizing use cases such as prediction of clinical outcomes; others proposed new domain-specific sentiment analysis approaches based on machine learning. Accuracy values between 71.5–88.2% are reported. Data used for evaluation and test are often retrieved from MIMIC databases or i2b2 challenges. Latest developments related to artificial neural networks are not yet fully considered in this domain. We conclude that future research should focus on developing a gold standard sentiment lexicon, adapted to the specific characteristics of clinical narratives. Efforts have to be made to either augment existing or create new high-quality labeled data sets of clinical narratives. Last, the suitability of state-of-the-art machine learning methods for natural language processing and in particular transformer-based models should be investigated for their application for sentiment analysis of clinical narratives.

1. Introduction

Sentiment analysis studies opinions, sentiments, evaluations, attitudes and emotions as expressed in natural language text [1]. Research in this field became popular accompanied by the rise of social media in 2004 [2] with use cases such as spam detection in social media [3], sentiment analysis in political debates [4], or analysis of customer reviews [5]. Later on, applications in the medical domain were identified related to determining suicide ideation in social media [6] or recently, understanding sentiments expressed in social media related to pandemics [7].

Even though most research in the field of medical sentiment analysis considered social media data, interesting secondary use cases are emerging when analyzing sentiment expressed in clinical narratives. Clinical narratives are written documentations of patient encounters that describe the patient's history, symptoms, examination findings, diagnoses, treatment plans, and other relevant information. They also reflect perceptions of healthcare professionals on the patient's health status. They are used in the medical field to communicate important information about a patient's care and to document the treatment progress. Clinical narratives can be written by a variety of healthcare professionals, including doctors, nurses, and other clinicians, and they are typically included in a patient's medical record. Several types of documents exist that are generated at the various stages of the patient journey (e.g. patient's medical history, finding report, progress notes, nursing notes, discharge summaries). Clinical narratives are characterized by a specific use of language, including clinical terms or domain-specific abbreviations. The use of negations is prevalent, reflecting the process of clinical diagnosing, starting from a hypothesis and making examinations to confirm or reject hypotheses. Explicitly rejected hypotheses or negative findings are also documented in clinical narratives.

The focus of this paper is on sentiment analysis of clinical narratives. We aim at summarizing research in this field published in the last

* Corresponding author.
E-mail address: kerstin.denecke@bfh.ch (K. Denecke).

https://doi.org/10.1016/j.jbi.2023.104336
Received 6 January 2023; Received in revised form 6 March 2023; Accepted 10 March 2023
Available online 22 March 2023

8 years to identify open research gaps to be addressed in the future. A first overview and vision paper on medical sentiment analysis focusing on clinical narratives was published in 2015 [8]. After systematically comparing word usage and sentiment distribution between clinical narratives (nurse letters, discharge summary, and radiology reports) and medical social media (medical-related blogs, drug reviews), Denecke and Deng concluded that off-the-shelf sentiment analysis tools are not ideal for analyzing sentiment in clinical documents [8]. It was demonstrated that it is more complex to predict sentiment from clinical narratives than from social media data. This is because words can have different meanings based on the patient's medical history, and meanings of terms used in a clinical context can differ from meanings in general domains. Relating to this initial overview on medical sentiment analysis and considering the advances in the field of artificial intelligence during the last years, we are now interested in recent developments in this field.

Sentiment and opinion has been defined since the beginning of research in this field [9] as a quadruple comprising a *sentiment target*, a *sentiment* of the opinion about the target, an *opinion holder* and a *time* when the opinion was expressed. The *sentiment target* is the entity on which a sentiment has been expressed upon [9]. Medical sentiment can be expressed towards a diverse set of sentiment targets including anatomical structures, health status, symptoms, disease-specific risk factors or treatments. Concerning clinical narratives, the *opinion holder* expressing the opinion and *time* are only accessible from the overall document since it is normally written in passive voice.

A *medical sentiment* can be defined as an attitude, thought or judgment promoted by an observation with respect to the health of some individual [9]. *Rational sentiments* originate from "rational reasoning, tangible beliefs and utilitarian attitudes. They express no emotions". [9] (e.g. the phrase *"the tumor is malignant"* implies a rational sentiment). *Emotional sentiments* originate from "non-tangible and emotional responses to entities which go deep into people's psychological state of mind" [9]. Medical sentiment expresses judgments, vagueness, certainty etc. concerning a medical sentiment target (e.g. medical condition and its appearances and (health) consequences for an individual) [8].

Several reviews summarize use cases of sentiment analysis related to health and medicine when analyzing social media data: Babu et al. reviewed research on sentiment analysis from social media for depression detection [10]. Gohil et al. reviewed sentiment analysis from healthcare tweets [11]. Zunic studied approaches to sentiment analysis in health and well-being [12]. To the best of our knowledge, there is no review available summarizing research developments of sentiment analysis of clinical narratives. Given the linguistic and content peculiarities of clinical narratives and the resulting challenges, it is of relevance to study use cases, methods and quality of sentiment analysis of clinical narratives separately from social media.

More specifically, we are interested in answering the following research question: What is the state of research regarding sentiment analysis on clinical narratives? Associated to this, we will answer the following questions;

- RQ1: What are practical applications and outcomes?
- RQ2: What are the major sources of data used?
- RQ3: Which methods and features have been used?
- RQ4: What is the state-of-the-art performance?
- RQ5: What are open challenges in this field?

2. Material and methods

2.1. Study design

We conducted a scoping review to answer our research questions. The range of study designs currently used in the field of sentiment analysis of clinical narratives makes equitable risk of bias assessment difficult; therefore, we decided for a scoping review instead of a systematic literature review. Scoping reviews are generally accepted as appropriate when diversity of study designs is expected [13].

Table 1
Inclusion criteria.

I1	The study describes information extraction of input texts related to the healthcare sector.
I2	The input text's sentiment is analyzed automatically using natural language processing.
I3	The input text represents clinical narratives (e.g., nursing letter, discharge letter, reports).
I4	The input text is created by medical professionals (physicians, nurses, etc.).

The research methodology used in this scoping review adheres to specifications of the JBI Manual for Evidence Synthesis [14]. The PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guideline [15] was followed to carry out a transparent review. It is available in Appendix A. An internal review protocol was created and utilized to guide the research process, but not uploaded online.

2.2. Search strategy

First, we searched three databases for upcoming reviews that have been registered (Prospero, OSF and figshare) using the term "Sentiment analysis" to determine if a scoping or systematic review is currently conducted or planned. No relevant search results were found. Second, we iteratively formed the search string for this scoping review. Several variants of the search terms were used to optimize the number of search results, keeping the amount of results manageable while ensuring all relevant articles to be included. Starting with a first search term, which resulted out of a brainstorming process, the first query yielded more than 13,000 search results. Hence, we reduced the number of disjunctions and added one conjunction as well as an additional exclusion criterion. The final search string was: ("Sentiment classification" OR "sentiment analysis") AND (notes OR narrative OR document OR text OR report OR ehr OR "electronic health record") AND (medical OR clinical OR hospital OR healthcare) NOT Twitter NOT "Social Media". It reflects our interest in research on sentiment analysis of clinical narratives, explicitly excluding work related to social media. The following electronic databases were searched in the current review: PubMed, IEEE Xplore, ACM Digital Library and Web of Science. The search strings used for searching each electronic database are detailed in Appendix B.

2.3. Eligibility criteria

Moreover, we defined four inclusion criteria and five exclusion criteria to ensure the eligibility of potential sources of evidence, see Tables 1 and 2. We intentionally focused on the period of January 2015 to October 2022 to cover the period since the publication of the first overview paper on this topic published in 2015 by Denecke and Deng [8]. Since that paper gave an overview on sentiment analysis research in the medical domain before 2015, we were interested in the developments happening since then. Criterion E5 was used to ensure a certain quality of the papers, assuming that a publication with a minimum number of pages provides enough details to comprehend the addressed problem. Posters, study protocols or complete conference proceedings and reviews were excluded. In this review, KD and DR independently screened the titles and abstracts of all retrieved studies and decided for eligibility. Any disagreements were discussed between the reviewers.

Table 2
Exclusion criteria.

E1	The study was published before 1.1.2015.
E2	The study describes a literature review of any type.
E3	The study is not freely accessible to the authors.
E4	The study's language is not English or German.
E5	The study is shorter than five pages.

2.4. Data extraction and synthesis

To extract relevant data from the retrieved literature, we defined the following data items in addition to typical bibliographic properties such as the year of publication:

- Data used: Dataset, size, type of documents (progress notes, radiology report, discharge summaries),
- Methods and features used (Rule-based/Machine Learning/Hybrid),
- Tools and lexical resources used (e.g. sentiment lexicons, sentiment analysis tools),
- Use case related information: Objective of sentiment analysis, outcomes, key findings, open challenges mentioned.

The data extraction form is available in Appendix C. For certain aspects, categorical options were defined in order to facilitate the interpretation of the results. Data extraction was done independently by the two authors, except for six papers that were reviewed in duplicate. Data was extracted by both authors using the data extraction sheet.

The extracted data were summarized quantitatively where appropriate and qualitatively. Finally, we derived trends and research gaps in sentiment analysis of clinical narratives from the results.

3. Results

3.1. Search results

The final search was conducted on October 28, 2022. We queried the databases PubMed, IEEE Xplore, ACM Digital library and Web of Science, using the optimized search string. The search yielded a total of 1200 search results. They were imported into the tool Rayyan (https://www.rayyan.ai), which was used for the subsequent source selection process.

Upon completion of the import, 163 possible duplicates were automatically identified, hence 79 duplicates were removed after manual check. Next, a pilot test of the in- and exclusion criteria was conducted by the two authors, assessing the same 22 search results separately and blinded. This resulted in 18 concordant decisions and four conflicts, which were resolved by discussing each result and the applicable in- or exclusion criterion, eventually reaching accordance.

As a next step, 182 studies published before 2015 were excluded before starting the main source selection process which comprised 939 studies being marked as "included", "excluded" or "maybe" by each of the two reviewers, based on title and abstract. The resulting four conflicts and eight studies being marked as "maybe" were again resolved by discussion and reconsideration of the eligibility criteria. The source selection process resulted in 41 studies eligible for full-text review. These 41 results were exported from the review software and imported into the literature management tool Zotero. 40 of 41 full texts could be accessed. One study was not accessible and was therefore excluded. Four additional duplicates were removed manually after acquiring the full texts. Ten reports were dropped due to fulfilling exclusion criteria. Three additional reports were added based on bibliographic forward/backward search. Finally, a total of 29 full texts were used for data extraction, see Fig. 1.

3.2. Characteristics of included studies

Overall, between two and four papers were published each year since 2015. Interestingly, seven papers (24%) were published in 2021 forming a slight peak in the amount of publications published during these eight years. Most of the included studies were experimental, i.e. the sentiment analysis approach was not yet applied in daily use. One study reported on a text analysis where sentiment analysis was applied to understand and interpret manually the content of the texts. Nine studies were retrospective studies, one descriptive study was found, two studies dealt with the generation of a sentiment analysis corpus [16] and three studies reported on a comparison of sentiment analysis tools or methods.

3.3. Practical applications of sentiment analysis and outcomes (RQ1)

Overall, we can distinguish studies that propose a (new) sentiment analysis approach or generate a sentiment lexicon from those that apply sentiment analysis to clinical narratives and used the results for subsequent tasks like risk prediction or text analysis.

In the retrieved papers, a variety of secondary use cases were considered for medical sentiment analysis. Most research exploited sentiment analysis for predicting the in-hospital or 28-day-mortality risk [17–25]. Additional risk prediction use cases concerned prediction of venous thromboembolism [26,27], suicide risk assessment [28–30], or readmission risk prediction [20,31]. Other researchers used sentiment analysis to identify risk factors for specific diseases, comprising loneliness in patients [32] and analysis of sentimental risk factors [33] as well as identifying alterations in patient's attitudes and feelings [34]. We found a correlation analysis between healthcare provider sentiment and use of diagnostic imaging utilization [35], an application for analyzing treatment quality [36,37] and the use of sentiment analysis to identify bias in the clinical writings of physicians and nurses [38].

Even though a concrete use case in mind, some papers rather presented text analysis results such as coverage of sentiment lexicons when applied to the data relevant for the use case [28] or a corpus analysis [32,34].

None of the papers reported on results from a clinical trial where a concrete benefit for a group of patients could be achieved through the application of sentiment analysis. First indications that sentiment analysis can help to increase the prediction quality for mortality or readmission risk were reported [17–22]. Interesting conclusions were drawn related to the content of nursing notes: Sentiment scores measured in nursing notes are statistically significant predictors for mortality [23] and sepsis [17]. One study compared different state-of-the-art tools for sentiment analysis when applied to clinical narratives and confirmed that these tools fail in correctly identifying clinical sentiment [39]. This is due to a difference in word usage and differing meaning depending on the context [8].

3.4. Lexical and textual resources used (RQ2)

The datasets that have been used for medical sentiment analysis (see Table 3) include data that have been aggregated for scientific challenges (MIMIC databases, i2B2 challenges). Some research used data collected from the electronic health records of the involved hospitals. For this reason, for six (20.7%) papers, the used datasets are unavailable, for 16/29 (55.2%) they are available and for seven papers (24.1%) we were unable to judge the availability of the dataset.

69% of the studies (20/29) used clinical notes or nursing notes; 24% (7/29) developed an approach analyzing discharge summaries, one paper used radiology reports and for one paper it was not specified which clinical text type was used. The average size of the datasets is difficult to compare since the reported granularities differ (number of cases, number of sentences, number of documents). The smallest

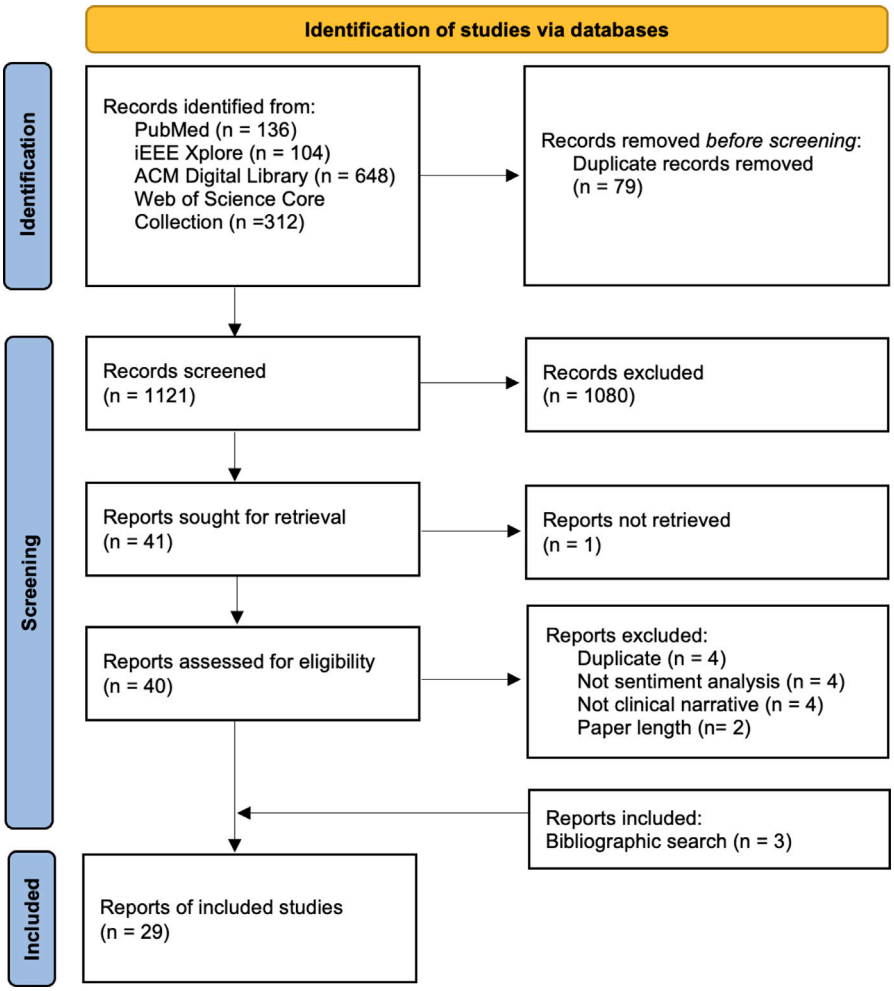


Fig. 1. PRISMA 2020 flow diagram of the source selection process.

datasets consisted of 100 documents; the largest dataset comprised 1,237,977 notes.

Annotations regarding sentiment as ground truth were only created by six papers. The annotations were made by health professionals (5150 sentences [39], 1400 documents [40], 150 documents [26], 25 documents [16]), the researchers themselves (6000 sentences [41]) or Ph.D students in linguistics (1212 documents [37]). All other studies included in the review were directly applying sentiment lexicons or sentiment analysis tools without constructing a ground truth for quality assessment of the sentiment analysis.

3.5. Applied sentiment lexicons and tools (RQ3)

A variety of sentiment lexicons and analysis tools have been used in the 29 papers, see Table 4. Three papers did not use a lexicon or did not report on it. The applied lexicons are well-known for sentiment analysis and include for example AFINN, SentiWordNet, EmoLex, or the subjectivity lexicon. Two papers reported on the development of their own sentiment lexicons. Some of the lexicons were integrated in a sentiment analysis tool, for example LIWC or VADER. The Python libraries TextBlob and Pattern were the most frequently used or tested tools. Neither the sentiment lexicons nor the sentiment analysis tools that have been applied were domain-specific, i.e. not specifically developed for the analysis of clinical narratives.

3.6. Features and methods used (RQ3)

We grouped the approaches used for realizing sentiment analysis into four categories: lexicon-based (n = 15) and machine learning-based (n = 2) approaches, papers combining both methodologies (n = 8) and comparison of different lexicons, tools or algorithms (n = 3). One paper did not report on sentiment analysis methods, as its intention was to rather create a corpus [16]. See Table 5 for an overview of approaches and corresponding papers. Several machine learning-based approaches were used, including Support Vector Machines (SVM) [41], Convolutional Neural Network (CNN) [40], Logistic Regression [29], or Random Forest classification [22,29]. Beyond, embedding algorithms and calculation of cosine similarity [38] were applied. Furthermore, we identified a newly developed approach named Extreme Learning Machine Autoencoder, combining unsupervised deep learning, statistical methods and clustering [36,37] or a machine learning algorithm called Absolute shrinkage and selection operator (LASSO) [30]. One source of evidence created a trained lexicon based on labeled data [42], another utilized word embeddings [43].

Depending on the chosen approach, different features were used to conduct the sentiment analysis: While lexicon-based approaches rely on provided features by the chosen lexicon or tool (e.g. binary or ternary sentiment, polarity, subjectivity), machine-learning approaches utilize word- and/or document embeddings to train a sentiment classifier. Most of the approaches used scores instead of frequency of polarity categories as features (see Table 6).

3.7. Performance of sentiment analysis (RQ4)

The included studies reported either on the quality of sentiment analysis or on the quality of the underlying use case (e.g. mortality

Table 3
Data sets used in the included research papers.

Name of dataset	Brief description	Used in
MIMIC-II	The MIMIC Corpus (Medical Information Mart for Intensive Care) comprises data from hospital admissions requiring ICU care at the Beth Israel Deaconess Medical Center in Boston. MIMIC-II contains data collected between 2001 and 2008 from a variety of intensive care units (medical, surgical, coronary care, and neonatal).	[8,16,41–43]
MIMIC-III	The MIMIC III database comprises additional MIMIC II data and additional data until 2012, in total from more than 40,000 patients.	[17–19,22–25, 35]
i2B2 Heart failure dataset	The dataset comprises 1304 de-identified longitudinal medical records describing 296 patients, selected to support research into the progression of coronary artery disease (CAD) in diabetic patients.	[26,27,33]
i2B2 Obesity dataset	The obesity challenge data consisted of 1237 discharge summaries. The data were taken from the discharge summaries of patients who had been hospitalized since December 1, 2004, for either obesity- or diabetes-related reasons.	[22,36–38]
Individually created dataset	Datasets were retrieved from specific hospitals.	[20,21,28–30,34, 39,40]
Not specified	No information on the origin of the data used were provided.	[31,32]

risk prediction). Only five papers (17.2%) studied the quality of the applied sentiment analysis method [28,29,33,41,42]. Reported accuracy values ranged from 71.5–88.2%, precision from 50–72%, recall from 4–60% and F1-measure from 50–65%. One paper studied the efficacy in sentiment representation by calculating the cosine similarity to the benchmark labeling [38]. 14 papers (48%) reported quality assessment results from the use cases, which were basically prediction use cases (e.g. in-hospital mortality). For these use cases, AUC values were reported [18,19,22,23,25,26,30,31] with values between 0.604–0.899 or AUC difference values [21]. Accuracy of risk prediction was reported with values between 64% and 93% [18,27,36,40]. None of these use case papers assessed the impact of the sentiment analysis quality to the quality of the downstream task.

3.8. Open challenges

We identified the following open challenges that were reported in the analyzed papers: Three papers pointed out that sentiment was only analyzed at patient level. Analyzing and comparing sentiment at the level of clinical narrative authors could yield additional insights [17,22,23]. Moreover, several studies pointed out the lack of missing medical sentiment lexicons: They are not domain specific [35], corpora are not representative [28] and there is no gold standard corpus of medical notes [20]. Existing lexicons are not suitable for suicide assessment [28] and clinical terms are missing in general-purpose lexicons like Pattern [39]. Another open challenge proves to be the source of available data sets; these are usually derived from one single healthcare institution or an association of multiple clinics [17,23,25]. The specific challenges of analyzing clinical narratives is also pointed out in several papers: when considering sentiment related to specific risk factors (e.g. mood of the patient), clinical narratives can contain less neutral content, leading to biased classifiers towards positive or negative polarity [39] and show discordant or opposite sense of polarity [24,26,33].

4. Discussion

4.1. Principal findings

We conducted this scoping review to revisit the current state-of-research regarding the application of sentiment analysis to clinical narratives. Specifically, we defined five research questions, encompassing data sources, practical applications, methods and features, performance and challenges of sentiment analysis. The review identified a total of 29 relevant studies, published in a period of eight years, which were used to support the findings that are summarized in the following. Sentiment analysis of clinical narratives since 2015 has focused on the application area of clinical outcome prediction, while most sources of evidence focused on risk prediction, in particular in-hospital or 28-day-mortality risk. Moreover, sentiment scores of nursing notes prove to be statistically significant predictors for mortality and sepsis.

The clinical text type that has been most frequently analyzed and used for medical sentiment analysis are clinical notes or nursing notes, respectively. Such notes bear valuable information on the health status of a patient. They include descriptions of the patient's physical condition, including vital signs, symptoms, and any treatments or interventions that have been administered. In addition, nursing notes may also include observations about the patient's behavior, mood or on the psychological status [39], and overall functioning, as well as recommendations for further care or treatment. This might be a reason why they are a more popular source for sentiment analysis compared to other clinical text types such as discharge summaries. Beyond, clinical notes are generated continuously during the course of a treatment and rather reflect the progress of health status. The existing datasets MIMIC or i2B2 that aggregate clinical narratives have been most frequently used for evaluating sentiment analysis-based approaches. It is worth mentioning that these datasets are not shared with annotated sentiment information (except for the i2b2 suicide dataset). Only one of the included studies released an annotated dataset derived from MIMIC-II [16].

Lexicon-based approaches were applied more often than machine learning-based or mixed approaches. A reason might be the unavailability of annotated data sources. We identified 20 sentiment lexicons and tools that were mentioned in one or more sources of evidence: Only general-purpose lexicons and tools not adapted to the medical domain were applied, of which SentiWordNet was used the most, followed by the Python modules TextBlob and Pattern as well as the AFINN and Liu's and Hu's Opinion sentiment lexicons. Older, but well-known lexicons, like the Harvard General Inquirer [62] or ANEW [63] as well as existing domain-specific sentiment lexicons (e.g., SentiHealth [64]) have not been used. Support vector machines, logistic regression, random forest classification or convolutional neural networks were applied as machine learning techniques.

Accuracy of the sentiment classification ranged from 71.5–88.2% and a F1-measure between 50–65%. While comparable to the average accuracy reported by Zunic et al. for sentiment analysis of medical social media data [12], this is well below accuracy achieved for sentiment analysis for movie reviews, which is typically larger than 90% [65].

4.2. Progress in medical sentiment analysis

Below, we will present the progress in sentiment analysis of clinical narratives in comparison to the results presented by Denecke and

Table 4
Sentiment lexicons and tools used in the included research papers.

Lexicon or tool	Description	Used in
AFINN	The AFINN lexicon is a lexicon of terms manually rated for emotional valence, derived from twitter postings [44].	[19,24,28,32,34]
SentiWordNet	This lexicon is the result of automatically annotating all WordNet synsets (groupings of synonymous words that express the same concept) according to their degrees of positivity, negativity, and neutrality [45]. WordNet is an English lexical database, containing meaningfully connected words and concepts [46].	[8,26–28,33,35–38,42]
Subjectivity Lexicon	This lexicon uses a phrase-level sentiment analysis approach, facilitating the identification of the contextual polarity of sentiment expressions. It is maintained as part of the Multi-Perspective Question Answering (MPQA) project [47].	[8]
NRC Word-Emotion Association Lexicon (EmoLex)	A list of manually annotated English words and their associations with eight basic emotions and two sentiments (negative and positive) [48].	[22,24,28,34]
Opinion Lexicon by Liu and Hu	A list of 6782 English words, annotated as having either a positive or negative sentiment [49].	[22,24,28,34,43]
Semantic Orientation Dictionaries V1.11[a] [50]	Lists with words and their semantic orientation on a scale (available in English and Spanish)	[34]
Negation detection from Vilares et al. [51]	Vilares et al. developed a negation detection algorithm for Spanish based on a list of negation terms (non, nunca, sin) and interpretation rules.	[34]
Generation of a new lexicon	Deng et al. developed a corpus and annotation scheme for clinical sentiment based on nurse letters. Holderness et al. created a sentiment lexicon adapted to psychiatric EHR.	[16,39]
No lexicon specified or used	In these papers no lexicon was mentioned or used.	[40,41,43]
Sentimentr	A dictionary-based sentiment analysis package for the software environment R that considers valence shifters (negators, amplifiers, de-amplifiers and adversative conjunctions) [52].	[24,31]
CoreNLP	CoreNLP is a Java-based set of tools to derive linguistic annotations for text, including sentiment. It is available for eight languages [53].	[24]
pattern	Pattern is a web mining module for Python that provides NLP-functionalities, including sentiment analysis [54].	[20–22,24,28]
VADER	The Valence Aware Dictionary and Sentiment Reasoner is a lexicon and rule-based sentiment analysis tool attuned to sentiment in microblog-like contexts [55].	[21]
TextBlob	TextBlob is a Python library for natural language processing, based on NLTK and pattern, including a sentiment analysis feature [56]	[17–19,22,23,25]
ABSApp	ABSApp provides weakly-supervised aspect-based sentiment extraction, without requiring labeled training data [57].	[29]
SEANCE	The Sentiment Analysis and Cognition Engine is a tool written in Python for text processing, using predefined word vectors from several source databases (including EmoLex and VADER, a.o.). It provides functionalities to analyze text regarding sentiment, cognition and social order [58].	[30]
LIWC	The Linguistic Inquiry and Word Count system is a commercial text analysis program for investigating different dimensions of texts [59].	[28]
SentiStrength	It is an algorithm to extract sentiment from informal texts, optimized for short social media texts [60].	[22]
Opinion Finder 2.0	It is a Java-based tool to process documents and identify subjective sentences and sentiment expressions, originally released in 2005. It is maintained as part of the MPQA project [61].	[22]

[a] https://github.com/sfu-discourse-lab/SO-CAL/tree/master/Resources/dictionaries.

Table 5
Chosen approaches to sentiment analysis.

Approach	Description	Used in
Lexicon-based	Using a sentiment lexicon or tool mentioned in Table 4	[8,17–21,23,25–27,31–35]
Machine learning-based	Labeled data is used to train a machine learning model to classify documents	[40,41]
Mixed approach	Combine lexicon- and machine learning-based methods	[22,29,30,36–38,42,43]
Method comparison	Compare several methods	[24,28,39]

Deng in 2015 [8]. Four tasks of medical sentiment analysis can be distinguished: polarity analysis, subjectivity analysis, emotion analysis and intensity analysis [8]. None of the studies included in this review reported on an approach to emotion analysis. However, the emotion lexicon EmoLex was used in some studies to identify features related to emotions. A reason for missing emotion analysis approaches might be the rather rational way in which clinical narratives are written. However, it could be studied whether nursing notes report on patient's emotions, which could be extracted as features for subsequent tasks.

In terms of methods applied or developed for conducting the sentiment analysis, we recognize that a majority of studies relied upon rule-based systems. Only two papers considered more recent developments around artificial neural networks and applied extreme learning machine autoencoder [36] and convolutional neural networks [40].

Table 6
Features used in the included research papers.

Features	Used in
Binary sentiment (positive, negative)	[18,19,28,30,34,41,43]
Ternary sentiment (positive, negative, neutral)	[16,21,39]
Score (reflecting polarity, intensity or subjectivity)	[17,19–27,29,31–38,40–42]
Word- and/or document-embeddings	[36–40,43]
Word count	[8,22,32]
n-grams	[22,41]
Specific features provided by lexicon or tool (ABSApp, SEANCE, VADER, EmoLex)	[21,22,29,30]

While in 2015, standard methods were Naïve Bayes and Support vector machines, these techniques are still applied; in the included studies sentiment analysis was also realized with logistic regression. Deep learning algorithms are rising techniques in sentiment analysis in other domains [66]. Especially, Recurrent Neural Networks and LSTM are increasing in popularity for these tasks. This trend could not be recognized for the medical domain. A reason for the limited application of machine learning and even deep learning approaches for realizing sentiment analysis of clinical narratives might be the limited availability of annotated data. Neural network-based approaches are data-intensive, requiring large amounts of data for training. This can be considered a potential research gap to be addressed in the future.

First approaches started using semantic features and use medical ontologies within sentiment analysis [31,42]. This is a recent trend, suggested by Denecke and Deng [8] and should be followed-up since it seems to improve the quality of sentiment analysis. Features used for sentiment analysis did not change much since the paper in 2015. Word or document embeddings are upcoming representations for sentiment analysis of clinical narratives.

Sentiment lexicons that existed already in 2015 are still used. Some of these lexicons are continuously updated, however, none of them is domain-specific. There exist first domain-specific lexicons which remained unused in the reviewed studies. SentiHealth [67] was developed using the scores from SentiWordNet and a domain-specific strategy for assigning scores to the terms in the lexicon. More specifically, bootstrapping was applied to acquire a set of opinion words from manually compiled seed lists of medical terms. It comprises 1520 words (40% positive, 45% negative and 15% neutral). WordNet for Medical Events (WME) is a resource comprising medical concepts together with their linguistic and semantic features [68,69]. The conventional WordNet and an English medical dictionary were used as a basis. Future research should try to exploit these domain-specific lexicons.

The Python library TextBlob and Pattern seem to be popular tools in this domain. However, they are neither domain-specific nor based on domain-specific resources. The papers also resisted on analyzing the quality of such tools, even though it was already reported that off-the-shelf tools fail in analyzing the medical sentiment correctly. Since the quality of feature collection methods impacts on the performance of the subsequent higher-level tasks (such as risk prediction) it is recommended to ensure a high quality of sentiment analysis methods when applied for extracting features.

Denecke and Deng envisioned application areas for sentiment analysis of clinical narratives that were addressed in recent years. The use case of health status aggregation was implemented for the psychological sentiment by Holderness et al. [39]. They considered the sentiment related to different psychological risk factors, i.e. a patient's prognosis with regard to seven readmission risk factor domains (appearance, mood, interpersonal relations, substance use, thought content, thought process and occupation). This application area was enriched with additional use cases related to risk prediction of occurrence of specific symptoms or clinical events (e.g. loneliness, suicide, venous thromboembolism).

Outcome research is the most prevalent use case in the reviewed papers (mortality risk, readmission risk analysis). Interestingly, other application areas came up related to quality assessment (e.g. bias of nurses and physicians, or use of imaging facilities depending on sentiment in clinical notes).

In summary, the research challenges identified in 2015 by Denecke and Deng [8]:

1. modeling of implicit clinical context and determining implicit sentiment,
2. building upon a domain-specific sentiment lexicon,
3. determining sentiment depending on the context,
4. modeling different aspects of the patient's status,

were only partially addressed by researchers during the last eight years. The suggested idea of building a domain-specific sentiment resource using the UMLS [8] was taken up by two groups. Only two studies recognized the importance of use case-specific definitions of medical sentiment [16,39]. This demonstrates that there are still open research gaps regarding clinical sentiment analysis resulting from this comparison that are summarized in the next section.

4.3. Implications for research in sentiment analysis of clinical narratives

Towards domain-specific resources and datasets. There are several options to address the problem of unavailability of annotated datasets for medical sentiment analysis. One option is to manually annotate a dataset using human annotators, i.e. domain experts will have to label a set of clinical narratives regarding sentiment or emotions. This is time-consuming and resource-intensive, but it can yield high-quality annotated data. For human annotation, it is crucial to clearly define the clinical sentiment under consideration, since the definition might differ depending on the medical condition considered.

Another option is to use transfer learning, which involves training a model on a large, general-purpose dataset and then fine-tuning it on a smaller, domain-specific dataset. Creating an annotated dataset with transfer learning can help reduce the amount of (manually) annotated data needed for training. Weak supervision techniques allow using large amounts of unannotated or partially annotated data to train a machine learning model. For example, a rule-based classifier could automatically label a large dataset, which could then be used to train a machine learning model. We have seen that there are some annotated datasets available for medical sentiment analysis, such as those derived from the i2b2 or MIMIC datasets. While these datasets may not be ideal for every research problem, they can offer a useful starting point. As exemplified, they could be used to label automatically another dataset using a trained machine learning model. Research started to study the bias of MIMIC data [70] which resulted already in bias towards ethnicity and gender. Bias analysis regarding the sentiment expressed in clinical narratives could be another upcoming research field.

A community effort could also help to create and sharing a large, anonymized dataset which would support in benchmarking existing methods and could help in exploring new approaches.

Towards high-quality sentiment analysis. There are several options to address the domain-specific challenges. First, research has to be conducted to find technical solutions to the linguistic challenges to ensure a correct interpretation of detected sentiments (e.g. analysis of coordinated structures, negation processing). These challenges were merely neglected in the research of the last years. Domain-specific

dictionaries and lexicons could support in adapting machine learning models and in understanding medical terminology and jargon. Furthermore, incorporating domain knowledge into a machine learning model could help in better understanding the context and meaning of medical text. First attempts into this direction were made and would have to be continued. An option would be designing custom features or using domain-specific embeddings. Research on aspect-based medical sentiment analysis is still rare. Aspect-based analysis would allow for a more fine-grained analysis of expressed sentiments which might be important for clinical use cases.

Attention mechanism (self attention [71]), transformer-based models and gated multiplication (Gated CNN [72]) are widely used to realize sentiment analysis in the general domain. However, those newly emerged methods have not yet been tested with clinical narratives. Further, representation learning should be evaluated for sentiment analysis of clinical narratives.

For the conscious use of off-the-shelf tools like TextBlob or Pattern, their quality for sentiment analysis of clinical narratives has to be assessed. A benchmark and a standard evaluation procedure could hold together with an annotated dataset to test and compare the quality of such tools.

Demonstrate benefit for patient care. As outlined before, medical sentiment is not yet integrated in clinical practice. Besides their integration into clinical decision support systems, its potentials within concrete clinical use cases have to be demonstrated [22,39]. Some work has been done in the context of analyzing clinical notes or other clinical narratives for predicting re-admission risk or outcomes. However, this research is still in its beginnings and more convincing results have to be provided.

The performance of medical sentiment analysis has to be demonstrated on real word data sets, which is related to the generation of diverse annotated datasets. There is a need for more research on cross-cultural sentiment analysis. Medical sentiment analysis technologies must be trained on diverse patient datasets and must be rigorously evaluated for the various forms of bias [73].

Towards understandable and ethical sentiment analysis. Algorithmovigilance refers to the monitoring and evaluation of algorithms and artificial intelligence systems to ensure that their operation is correct and ethical [74]. This involves assessing the performance and bias of the algorithm, as well as ensuring that it is being used responsibly and in accordance with relevant regulations and guidelines. As soon as machine learning models are more frequently applied for clinical sentiment analysis, methods for making these models explicable or interpretable have to be considered [75]. Simple models, such as rule-based classifiers or linear models, are often more interpretable than complex models such as deep neural networks. Feature importance or feature relevance measures can help identify the most important features for a given prediction, which in turn provides insights into the aspects of a text that were most influential in determining the sentiment. There are several tools and libraries available that can help interpret the decisions made by complex machine learning models, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) [76]. Visualizations, such as attention maps or decision trees, can make the decisions made by a model more understandable.

4.4. Strengths and limitations of this work

This scoping review was conducted adhering to the JBI Manual for Evidence Synthesis and the PRISMA-ScR guideline for scoping reviews. Therefore, we assure the traceability and objectivity of the results presented. Additionally, the source selection procedure was carried out concurrently by the two authors and produced just twelve sources that required additional explanation, demonstrating the clarity of the previously established review procedures. However, we did not publish our review protocol to any online database. Moreover, data charting was not conducted in parallel, meaning that each selected source of evidence was only examined by one reviewer. During result synthesis, several sources of evidence were additionally examined by the second reviewer, as charted data was not detailed enough for synthesis. Our results show that sentiment of clinical narratives is often used as input for regression algorithms to improve prediction of clinical outcomes. As the focus of this review lies on sentiment analysis itself, we decided not to analyze these different regression algorithms in depth. Some papers did not describe explicitly the sentiment analysis approach — for these papers we were unable to extract the corresponding characteristics, which was clearly described in the result section.

5. Conclusion

With this scoping review, we present an overview of current research on sentiment analysis applied to clinical narratives. Since 2015, there still does not exist a gold standard lexicon for sentiment analysis within the medical domain, and domain-specific adaptations of sentiment lexicons remain unused. While machine learning-based methods could make such lexicons for the medical domain obsolete, we only identified a limited number of sources of evidence that used such a machine learning approach for sentiment analysis. This, in turn, could be caused by the limited availability of labeled open-source data sets of clinical narratives. Our research shows that sentiment analysis improves predictions on clinical outcomes like mortality and readmission risk.

We conclude that future research should focus on one of the following three areas of action: First, focus should be put on developing a gold standard sentiment lexicon, adapted to the specific characteristics of clinical narratives (e.g. little neutral content, discordant polarity). Second, in order to facilitate the application of machine learning-based approaches for sentiment analysis, effort needs to be taken to either augment existing or create new high-quality labeled data sets of clinical narratives. Last, the suitability of state-of-the-art machine learning methods for natural language processing and in particular transformer-based models should be investigated for their application for sentiment analysis.

CRediT authorship contribution statement

Kerstin Denecke: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Writing – review & editing, Project administration. **Daniel Reichenpfader:** Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. PRISMA checklist

Appendix B. Search strategy

See Table B.7.

Appendix C. Data extraction form

See Table C.8.

Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	1
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	2-4
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	4
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	4
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	6
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	7
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	5, Appendix B
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	5
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	6
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	6
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	-
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	7

St. Michael's
Inspired Care.
Inspiring Science.

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	Figure 1, 8
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	7-9
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	-
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	Appendix
Synthesis of results	18	Summarize and/or present the charting results as they relate to the review questions and objectives.	9-17
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	17-22
Limitations	20	Discuss the limitations of the scoping review process.	22-23
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	23
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	23

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.
* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.
† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).
‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.
§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/or qualitative research, expert opinion, and policy document).

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. doi: 10.7326/M18-0850.

St. Michael's
Inspired Care.
Inspiring Science.

Table B.7
Query variations for each database searched.

Database	Query string	Results
PubMed	(“Sentiment classification” OR “sentiment analysis”) AND (notes OR narrative OR document OR text OR report OR ehr OR “electronic health record”) AND (medical OR clinical OR hospital OR healthcare) NOT Twitter NOT “Social Media”	136
IEEE Xplore	(“All Metadata”:sentiment classification OR “All Metadata”:sentiment analysis) AND (“All Metadata”:note* OR “All Metadata”:narrative OR “All Metadata”:document OR “All Metadata”:text OR “All Metadata”:report OR “All Metadata”:ehr OR “All Metadata”:“electronic health record”) NOT (“All Metadata”:review) NOT (“All Metadata”:social media) NOT (“All Metadata”:twitter) AND (“All Metadata”:medical OR “All Metadata”:clinical OR “All Metadata”:hospital OR “All Metadata”:healthcare)	104
Web of science	((“Sentiment classification” OR “sentiment analysis”) AND (notes OR narrative OR document OR text OR report OR ehr OR “electronic health record”) AND (medical OR clinical OR hospital OR healthcare)) NOT Twitter NOT “Social Media” (All Fields)	312
ACM digital library	[[All: “sentiment classification”] OR [All: “sentiment analysis”]] AND [[All: notes] OR [All: narrative] OR [All: document] OR [All: text] OR [All: report] OR [All: ehr] OR [All: “electronic health record”]] AND [[All: medical] OR [All: clinical] OR [All: hospital] OR [All: healthcare]] AND NOT [All: twitter] AND NOT [All: “social media”]	648
Total		1200

Table C.8
Scoping review sentiment analysis: Aspects to be extracted.

Aspect	Dimensions						
Type of clinical narrative	Nursing notes	Radiology reports	EHR entries	Other			
Origin of dataset	Online dataset	Clinical data	Other				
Dataset size							
NLP methods used	Heuristic/Lexicon/Rule-based	Machine-learning based	Classification	Language Model (e.g. BERT)	Other		
Machine learning methods used	Naive Bayes	SVM	CNN	Multiple	Other		
Features used	Language Model (e.g. BERT)	Sentiment scores	Parts of speech	Multiple	Other		
Performance measures	F1-Score	Precision	Recall	Accuracy	AUC	Multiple	Other
Study-design	Method comparison	Retrospective					
Other NLP-tasks	Information Extraction	Named Entity Recognition	Relationship Extraction	Other			
Used embeddings	Word2Vec	Doc2Vec	Fasttext	Custom	Other		
Clinical outcome	Mortality Prediction	Thrombosis	etc…	Other			
SA tools used (e.g. TextBlob)							
SA lexicon used (SentiWordNet)							
Status (actual implementation, test phase)	Conceptualized	Tested	Implemented	Other			
Reported challenges							
Dataset available							

Appendix D. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2023.104336.

References

[1] B. Liu, Sentiment analysis and opinion mining, Synth. Lect. Hum. Lang. Technol. 5 (1) (2012) 1–167.
[2] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168–177.
[3] N. Jindal, B. Liu, Opinion spam and analysis, in: Proceedings of the 2008 International Conference on Web Search and Data Mining, 2008, pp. 219–230.
[4] O. Onyimadu, K. Nakata, T. Wilson, D. Macken, K. Liu, Towards sentiment analysis on parliamentary debates in hansard, in: Joint International Semantic Technology Conference, Springer, 2013, pp. 48–50.
[5] L. Garcia-Moya, H. Anaya-Sánchez, R. Berlanga-Llavori, Retrieving product features and opinions from customer reviews, IEEE Intell. Syst. 28 (3) (2013) 19–27.
[6] J.P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K.B. Cohen, J. Hurdle, C. Brew, Sentiment analysis of suicide notes: A shared task, Biomed. Inform. Insights 5 (2012) BII–S9042.
[7] A.H. Alamoodi, B.B. Zaidan, A.A. Zaidan, O.S. Albahri, K. Mohammed, R.Q. Malik, E.M. Almahdi, M.A. Chyad, Z. Tareq, A.S. Albahri, et al., Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review, Expert Syst. Appl. 167 (2021) 114155.
[8] K. Denecke, Y. Deng, Sentiment analysis in medical settings: New opportunities and challenges, Artif. Intell. Med. 64 (1) (2015) 17–27, http://dx.doi.org/10.1016/j.artmed.2015.03.006, URL https://www.sciencedirect.com/science/article/pii/S0933365715000299.
[9] B. Liu, in: E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco (Eds.), Many Facets of Sentiment Analysis, Vol. 5, Springer International Publishing, Cham, 2017, pp. 11–39.
[10] N.V. Babu, E. Kanaga, Sentiment analysis in social media data for depression detection using artificial intelligence: A review, SN Comput. Sci. 3 (1) (2022) 1–20.
[11] S. Gohil, S. Vuik, A. Darzi, et al., Sentiment analysis of health care tweets: review of the methods used, JMIR Public Health Surv. 4 (2) (2018) e5789.
[12] A. Zunic, P. Corcoran, I. Spasic, et al., Sentiment analysis in health and well-being: systematic review, JMIR Med. Inform. 8 (1) (2020) e16023.
[13] H. Arksey, L. O’Malley, Scoping studies: towards a methodological framework, Int. J. Soc. Res. Methodol. 8 (1) (2005) 19–32.
[14] M. Peters, C. Godfrey, P. McInerney, Z. Munn, A. Trico, H. Khalil, Chapter 11: Scoping reviews, in: E. Aromataris, Z. Munn (Eds.), JBI Manual for Evidence Synthesis, JBI, 2020, pp. 406–451, http://dx.doi.org/10.46658/JBIMES-20-12.
[15] A.C. Tricco, E. Lillie, W. Zarin, K.K. O’Brien, H. Colquhoun, D. Levac, D. Moher, M.D. Peters, T. Horsley, L. Weeks, S. Hempel, E.A. Akl, C. Chang, J. McGowan, L. Stewart, L. Hartling, A. Aldcroft, M.G. Wilson, C. Garritty, S. Lewin, C.M. Godfrey, M.T. Macdonald, E.V. Langlois, K. Soares-Weiser, J. Moriarty, T. Clifford, Ö. Tunçalp, S.E. Straus, PRISMA extension for scoping reviews (PRISMA-scr): Checklist and explanation, Ann. Internal Med. 169 (7) (2018) 467–473, http://dx.doi.org/10.7326/M18-0850.
[16] Y. Deng, T. Declerck, P. Lendvai, K. Denecke, The generation of a corpus for clinical sentiment analysis, in: H. Sack, G. Rizzo, N. Steinmetz, D. Mladenic, S. Auer, C. Lange (Eds.), Semantic Web, ESWC 2016, Vol. 9989, 2016, pp. 311–324, http://dx.doi.org/10.1007/978-3-319-47602-5_46.
[17] Q. Gao, D. Wang, P. Sun, X. Luan, W. Wang, Sentiment analysis based on the nursing notes on in-hospital 28-day mortality of sepsis patients utilizing the MIMIC-III database, Comput. Math. Methods Med. 2021 (2021) 3440778, http://dx.doi.org/10.1155/2021/3440778.
[18] V. Kumar, R. Bajpai, R.B. Roy, Clinical notes mining for post discharge mortality prediction, IETE Tech. Rev. (2021) http://dx.doi.org/10.1080/02564602.2021.1936224.

[19] Z. Liu, Y. Yang, H. Song, J. Luo, A prediction model with measured sentiment scores for the risk of in-hospital mortality in acute pancreatitis: a retrospective cohort study, Ann. Transl. Med. 10 (12) (2022) http://dx.doi.org/10.21037/atm-22-1613.

[20] T.H. McCoy, V.M. Castro, A. Cagan, A.M. Roberson, I.S. Kohane, R.H. Perlis, Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: An electronic health record study, PLoS One 10 (8) (2015) e0136341, http://dx.doi.org/10.1371/journal.pone.0136341, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0136341. Publisher: Public Library of Science.

[21] J.Y. Nakayama, V. Hertzberg, J.C. Ho, Making sense of abbreviations in nursing notes: A case study on mortality prediction, in: AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science, Vol. 2019, 2019, pp. 275–284.

[22] N. Tran, J. Lee, Using multiple sentiment dimensions of nursing notes to predict mortality in the intensive care unit - 2018 IEEE EMBS international conference on biomedical & health informatics (BHI), in: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2018, pp. 283–286, http://dx.doi.org/10.1109/BHI.2018.8333424.

[23] I.E.R. Waudby-Smith, N. Tran, J.A. Dubin, J. Lee, Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients, PLoS One 13 (6) (2018) e0198687, http://dx.doi.org/10.1371/journal.pone.0198687, URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0198687. Publisher: Public Library of Science.

[24] G.E. Weissman, L.H. Ungar, M.O. Harhay, K.R. Courtright, S.D. Halpern, Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness, J. Biomed. Inform. 89 (2019) 114–121, http://dx.doi.org/10.1016/j.jbi.2018.12.001.

[25] Y. Zou, J. Wang, Z. Lei, Y. Zhang, W. Wang, Sentiment analysis for necessary preview of 30-day mortality in sepsis patients and the control strategies, J. Healthc. Eng. 2021 (2021) 1713363, http://dx.doi.org/10.1155/2021/1713363.

[26] S. Sabra, K. Mahmood Malik, M. Alobaidi, Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives, Comput. Biol. Med. 94 (2018) 1–10, http://dx.doi.org/10.1016/j.compbiomed.2017.12.026, URL https://www.sciencedirect.com/science/article/pii/S0010482517304237.

[27] S. Sabra, K.M. Malik, M. Afzal, V. Sabeeh, A.C. Eddine, A hybrid knowledge and ensemble classification approach for prediction of venous thromboembolism, Expert Syst. 37 (1) (2020) http://dx.doi.org/10.1111/exsy.12388.

[28] A. Bittar, S. Velupillai, A. Roberts, R. Dutta, Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: Corpus-based analysis., JMIR Med. Inform. 9 (4) (2021) e22397, http://dx.doi.org/10.2196/22397.

[29] A. George, D. Johnson, G. Carenini, A. Eslami, R. Ng, E. Portales-Casamar, Applications of aspect-based sentiment analysis on psychiatric clinical notes to study suicide in youth, in: AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science, Vol. 2021, 2021, pp. 229–237.

[30] M. Levis, C. Leonard Westgate, J. Gui, B.V. Watts, B. Shiner, Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models, Psychol. Med. (2020) 1–10, http://dx.doi.org/10.1017/S0033291720000173.

[31] A. Zubillaga, P. Laccourreye, J. Kerexeta, N. Larburu, E. Alonso, D.J. Gómez, F. Martínez, M. Alonso-Arce, Hospital readmission prediction via keyword extraction and sentiment analysis on clinical notes, Stud. Health Technol. Inform. 295 (2022) 339–342, http://dx.doi.org/10.3233/SHTI220732.

[32] P. Bearse, O. Manejwala, A.F. Mohammad, I.R.I. Haque, An initial feasibility study to identify loneliness among mental health patients from clinical notes, in: 2020 3rd International Conference on Information and Computer Technologies (ICICT 2020), 2020, pp. 68–77, http://dx.doi.org/10.1109/ICICT50521.2020.00019.

[33] S. Sabra, K. Mahmood, M. Alobaidi, A semantic extraction and sentimental assessment of risk factors (SESARF): An NLP approach for precision medicine: A medical decision support tool for early diagnosis from clinical notes - 2017 IEEE 41st annual computer software and applications conference (COMPSAC), in: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), Vol. 2, 2017, pp. 131–136, http://dx.doi.org/10.1109/COMPSAC.2017.34.

[34] J.N. Cuenca-Zaldívar, M. Torrente-Regidor, L. Martín-Losada, C. Fernández-De-Las-Peñas, L.L. Florencio, P.A. Sousa, re, D. Palacios-Ceña, Exploring sentiment and care management of hospitalized patients during the first wave of the COVID-19 pandemic using electronic nursing health records: Descriptive study, JMIR Med. Inform. 10 (5) (2022) e38308, http://dx.doi.org/10.2196/38308.

[35] M.M. Ghassemi, T. Al-Hanai, J.D. Raffa, R.G. Mark, S. Nemati, F.H. Chokshi, How is the doctor feeling? ICU provider sentiment is associated with diagnostic imaging utilization - 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 4058–4064, http://dx.doi.org/10.1109/EMBC.2018.8513325.

[36] S.A. Waheeb, N.A. Khan, X. Shang, An efficient sentiment analysis based deep learning classification model to evaluate treatment quality, Malays. J. Comput. Sci. 35 (1) (2022) 1–20, http://dx.doi.org/10.22452/mjcs.vol35no1.1.

[37] S.A. Waheeb, N.A. Khan, B. Chen, X. Shang, Machine learning based sentiment text classification for evaluating treatment quality of discharge summary, Information 11 (5) (2020) http://dx.doi.org/10.3390/info11050281.

[38] Q. Chen, M. Sokolova, Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in analysis of scientific and medical texts., SN Comput. Sci. 2 (5) (2021) 414, http://dx.doi.org/10.1007/s42979-021-00807-1.

[39] E. Holderness, P. Cawkwell, K. Bolton, J. Pustejovsky, M.-H. Hall, Distinguishing clinical sentiment: The importance of domain adaptation in psychiatric patient health records, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 117–123, http://dx.doi.org/10.18653/v1/W19-1915, URL https://aclanthology.org/W19-1915.

[40] B. Shin, F.H. Chokshi, T. Lee, J.D. Choi, Classification of radiology reports using neural attention models - 2017 international joint conference on neural networks (IJCNN), in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 4363–4370, http://dx.doi.org/10.1109/IJCNN.2017.7966408.

[41] T.-T. Dang, T.-B. Ho, Mixture of language models utilization in score-based sentiment classification on clinical narratives, in: H. Fujita, M. Ali, A. Selamat, J. Sasaki, M. Kurematsu (Eds.), Trends in Applied Knowledge-Based Systems and Data Science, in: Lecture Notes in Artificial Intelligence, vol. 9799, 2016, pp. 255–268, http://dx.doi.org/10.1007/978-3-319-42007-3_22.

[42] N. Sanglerdsinlapachai, A. Plangprasopchok, T.B. Ho, E. Nantajeewarawat, Improving sentiment analysis on clinical narratives by exploiting UMLS semantic types, Artif. Intell. Med. 113 (2021) 102033, http://dx.doi.org/10.1016/j.artmed.2021.102033.

[43] M.M. Ghassemi, R.G. Mark, S. Nemati, A visualization of evolving clinical sentiment using vector representations of clinical notes, Comput. Cardiol. 2015 (2015) 629–632, http://dx.doi.org/10.1109/CIC.2015.7410989, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5070922/.

[44] F.Å. Nielsen, A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, in: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages, arXiv, 2011, http://dx.doi.org/10.48550/ARXIV.1103.2903.

[45] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, 2010.

[46] G.A. Miller, WordNet: A lexical database for english, Commun. ACM 38 (11) (1995) 39–41, http://dx.doi.org/10.1145/219717.219748.

[47] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proc. of HLT-EMNLP-2005, 2005.

[48] S.M. Mohammad, P.D. Turney, Crowdsourcing a word-emotion association lexicon, 2013, arXiv:1308.6297.

[49] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '04, Association for Computing Machinery, New York, NY, USA, 2004, pp. 168–177, http://dx.doi.org/10.1145/1014052.1014073.

[50] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) 267–307.

[51] D. Vilares, M. Alonso Pardo, C. Gómez-Rodríguez, Polarity classification of opinionated spanish texts using dependency parsing, Process. Leng. Nat. 50 (2013) 13–20.

[52] T. Rinker, V. Spinu, Trinker/Sentimentr: Version 0.4.0, Zenodo, 2016, http://dx.doi.org/10.5281/zenodo.222103.

[53] Stanford NLP Group, Stanford CoreNLP, 2023, Stanford NLP.

[54] T.D. Smedt, W. Daelemans, Pattern for python, J. Mach. Learn. Res. 13 (66) (2012) 2063–2067.

[55] C. Hutto, E. Gilbert, VADER: a parsimonious rule-based model for sentiment analysis of social media text, Proc. Int. AAAI Conf. Web Soc. Media 8 (1) (2014) 216–225, http://dx.doi.org/10.1609/icwsm.v8i1.14550.

[56] S. Loria, TextBlob: simplified text processing, 2023.

[57] O. Pereg, D. Korat, M. Wasserblat, J. Mamou, I. Dagan, ABSApp: A portable weakly-supervised aspect-based sentiment extraction system, 2019, http://dx.doi.org/10.48550/arXiv.1909.05608, arXiv:1909.05608.

[58] S.A. Crossley, K. Kyle, D.S. McNamara, Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis, Behav. Res. Methods 49 (3) (2017) 803–821, http://dx.doi.org/10.3758/s13428-016-0743-z.

[59] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, J. Lang. Soc. Psychol. 29 (1) (2010) 24–54, http://dx.doi.org/10.1177/0261927X09351676.

[60] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, J. Am. Soc. Inf. Sci. Technol. 61 (2010) 2544–2558, http://dx.doi.org/10.1002/asi.21416.

[61] Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, Identifying sources of opinions with conditional random fields and extraction patterns, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 2005, pp. 355–362.

[62] P.J. Stone, D. Dunphy, M.S. Smith, D.M. Ogilvie, The general inquirer, 1966.

[63] M.M. Bradley, P.J. Lang, Affective norms for english words (ANEW): Instruction manual and affective ratings, 1999.
[64] D.M. Asghar, S. Ahmad, M. Qasim, R. Zahra, F. Kundi, SentiHealth: Creating health-related sentiment lexicon using hybrid approach, SpringerPlus 5 (2016) http://dx.doi.org/10.1186/s40064-016-2809-x.
[65] P. Nandwani, R. Verma, A review on sentiment analysis and emotion detection from text, Soc. Netw. Anal. Min. 11 (1) (2021) 1–19.
[66] A. Ligthart, C. Catal, B. Tekinerdogan, Systematic reviews in sentiment analysis: a tertiary study, Artif. Intell. Rev. 54 (7) (2021) 4997–5053.
[67] M.Z. Asghar, S. Ahmad, M. Qasim, S.R. Zahra, F.M. Kundi, Sentihealth: creating health-related sentiment lexicon using hybrid approach, SpringerPlus 5 (2016).
[68] A. Mondal, D. Das, E. Cambria, S. Bandyopadhyay, WME: Sense, polarity and affinity based concept resource for medical events, in: Proceedings of the 8th Global WordNet Conference, GWC, Global Wordnet Association, Bucharest, Romania, 2016, pp. 243–248, URL https://aclanthology.org/2016.gwc-1.35.
[69] A. Mondal, D. Das, E. Cambria, S. Bandyopadhyay, WME 3.0: An enhanced and validated lexicon of medical concepts, in: Proceedings of the 9th Global Wordnet Conference, Global Wordnet Association, Nanyang Technological University (NTU), Singapore, 2018, pp. 10–16, URL https://aclanthology.org/2018.gwc-1.2.
[70] E. Röösli, S. Bozkurt, T. Hernandez-Boussard, Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model, Sci. Data 9 (1) (2022) 24.
[71] S. Lin, W. Su, P. Chien, M. Tsai, C. Wang, Self-attentive sentimental sentence embedding for sentiment analysis, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2020, pp. 1678–1682.
[72] W. Xue, T. Li, Aspect based sentiment analysis with gated convolutional networks, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 2514–2523, http://dx.doi.org/10.18653/v1/P18-1234, URL https://www.aclweb.org/anthology/P18-1234.
[73] I. Straw, Ethical implications of emotion mining in medicine, Health Policy Technol. 10 (1) (2021) 191–195.
[74] P.J. Embi, Algorithmovigilance-advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity, JAMA Netw. Open 4 (4) (2021) e214622.
[75] C. Zucco, H. Liang, G. Di Fatta, M. Cannataro, Explainable sentiment analysis with applications in medicine, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2018, pp. 1740–1747.
[76] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable AI methods-a brief overview, in: International Workshop on Extending Explainable AI beyond Deep Models and Classifiers, Springer, 2022, pp. 13–38.