Healthcare Analytics 2 (2022) 100107

Contents lists available at ScienceDirect

Healthcare Analytics

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

Healthcare related event prediction from textual data with machine learning: A Systematic Literature Review

Oscar Hoekstra, William Hurst *, Joep Tummers

Information Technology Group, Wageningen University & Research, Hollandseweg 1, 6706 KN Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Systematic literature review
Event prediction
Natural Language Processing
Textual data
Machine Learning
Healthcare

ABSTRACT

In the field of healthcare, as well as many others, textual descriptions of events are logged. With the use of Natural Language Processing (NLP), these texts are used to train event prediction machine learning algorithms. In this review the aim was to assess the state-of-the-art within current literature concerning prediction of events on textual records. Thus, this study follows a standard Systematic Literature Review (SLR) process. Primary articles are selected from PubMed, IEEE and WebOfScience with a search query, and then exclusion and quality assessment criteria are used to select the articles that are relevant to this study. Published performance metrics for the prediction algorithms used in the studies were then extracted from the included articles and used to assess the different methods. The general-purpose neural network algorithms: Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) and Conditional Random Fields (CRF) demonstrate the highest F1-scores amongst all the methods in this review, of 98.5%, 98% and 90.13% respectively. The algorithms that were designed specifically for NLP such as word2vec and BERT also performed well with F1-scores of 88.93% and 91.50%. This review does not give a comparison between methods but gives an indication about which machine learning methods perform well according to the authors of the selected studies. Not enough performance results are published under comparable circumstances to give conclusive results about which methods perform the best. More research needs to be done in comparing algorithms on the same dataset to proof the performance of the methods.

1. Introduction

In the healthcare industry, patient information is typically stored as natural language text. In hospitals and other care institutions, descriptions and recommendations of the care given to patients is written down by caregivers, and subsequently stored in electronic healthcare records (EHR). These information sources are straightforward for human use, but the semi-structured nature of the content and the ad-hoc fashion in which it is constructed makes it a challenge for computers to process [1–3]. The field of computer science that focuses on making computers understand language, is referred to as natural language processing (NLP). In NLP, computational models are constructed to infer information and context from written or spoken text. Within the NLP research field, the last ten years have demonstrated a significant advancement towards the automated extraction and machine-driven interpretation of information stored in vast text datasets [4–6]. Because of the continuing research into NLP, the best performing methods are constantly changing. On top of that, there are many different tasks that NLP models are built for, and each model performs differently at each task.

In this study, an investigation is presented into the automated event prediction from medical textual data. Event prediction is the process of estimating the chance an event will happen in the future [7]. If events in the healthcare sector could be predicted; this could be used to influence care given and therefore improve overall healthcare. As text is an unstructured, the link between an event and the situation that led to that event can be a challenge to connect [8–11]. If this process is automated, there are core benefits in (for example) the healthcare industry, where caregivers report what has happened with their patients. With the use of automated event prediction, medical emergencies could possibly be predicted and, potentially, prevented.

Different machine learning (ML) methods are better in diverging situations, and within event prediction there is not one set method that can always be referred to as 'the best'. Thus, the goal of this study is to investigate the ML methods that are most employed in event prediction applications in the field of healthcare, and which can provide the optimal predictions consistently. To increase the size of this study, methods from other fields that could be used in healthcare are also considered. Therefore, the investigation extends to all available literature in all fields of study in order to research the state-of-the-art methods for predicting events from text by means of a Systematic

* Corresponding author.
E-mail addresses: oscar.hoekstra@wur.nl (O. Hoekstra), will.hurst@wur.nl (W. Hurst), joep.tummers@wur.nl (J. Tummers).

https://doi.org/10.1016/j.health.2022.100107
Received 18 July 2022; Received in revised form 25 August 2022; Accepted 2 September 2022
Available online xxxx

Literature Review (SLR). Most previous SLRs into similar subjects have focused on extraction and classification of events from text [12–15]. There have been some examples of event prediction literature reviews, but these focus on a single type of prediction or data source, such as stock market predictions from news articles [16]. The goal of this study is to discuss the optimal performing event prediction methods for textual records. The following research questions were constructed to be able to meet the objective of our research, as described above. (1)

RQ1: *What is the state of the Art within current literature concerning prediction of events based on textual records*? (2)

RQ1.1: *How is machine learning used to predict events from textual data*?

RQ1.2: *Which machine learning methods used for event prediction achieve the highest performance*?

As such, this article provides the following contributions. Firstly, a discussion is provided on the optimal performing event prediction methods for textual records, by means of an SLR; Secondly an investigation of related articles is put forward, from which (thirdly) a presentation of the findings in discussion and visual format is provided. The process of this SLR is conducted according to the guidelines set by Kitchenham et al. [17]. These guidelines are targeted at software engineering research and are adapted from the guidelines used in medical research set by Cochrane Collaboration [18]. The process of the review protocol will be described in the Section 2: Methods. After that the results will be shown in Section 3, followed by a discussion in Section 4 and a Conclusion in Section 5.

2. Methods

In adherence to the guidelines presented by Kitchenham et al. [17], this SLR was undertaken in 5 consecutive steps. These steps are shown in Fig. 1. As a first step, a search strategy, including a search query, was developed iteratively by reviewing related articles and their corresponding keywords. In the second step, a set of study exclusion criteria were constructed, and then used to filter out literature that was unusable or unrelated to the subject. Then, to exclude literature that is directly unrelated to the work, a quality assessment method was designed. The fourth step involved creating a data extraction form to be able to extracts all important information out of the remaining articles. The fifth step, a data synthesis, summarized the extracted information into the results shown in Section 3.

2.1. Search strategy

To select literature relevant for answering the research questions, a search was performed in three bibliographic databases: PubMed, IEEE Xplore and Web of Science. These databases are selected to obtain literature from a widespread source aligning to the project scope. PubMed was identified since it focuses predominately on health and medical science. IEEE Xplore was used as it focusses on computer science, of which machine learning is an applied field of technology. Web of Science was used as it is a more general database covering material potentially omitted from IEEE Xplore and PubMed.

The search query (1) was designed in such a way that it could be transferred between the different databases. No additional filtering in each of the search engines was applied. The search query was performed on All Metadata.

((Text (record OR data OR archive OR report OR dossier OR account)) AND (NLP OR "natural language processing") AND ((machine OR deep) learning) AND (((event OR occurrence OR incident) (predict* OR forecast OR foresee OR foretell OR anticipat* OR estimat*)) OR (time to event*) OR (survival analysis))) (1)*

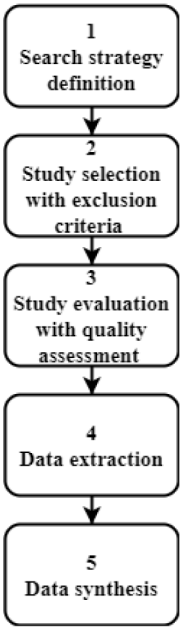

Fig. 1. Review protocol adapted from Kitchenham et al. [1].

2.2. Exclusion criteria

In line with the SLR methodology, to further select relevant studies, exclusion criteria are employed. The goal of the exclusion criteria was to exclude studies, of which, the subject was not relevant for the study objective. First, each abstract was read. If any of the criteria described in Table 1 were met for a study, it would be excluded from the SLR. To decrease the likelihood of bias, the selection criteria were established after 15 randomly selected trial articles were independently reviewed. After discussion of the selected exclusion criteria by each of the 3 authors, an interrater agreement of 80% was reached.

EC 4–6 were included to exclude articles that were selected from the query, but which were not according to the objective of this SLR. A significant volume of articles discuss machine learning, but often the technology is not used in the study and, therefore, not suitable for inclusion in this investigation. Similarly, textual data is often a subject covered in articles, but not the main focus. Additionally, the search query selected a proportion of studies that describe a method solely for identification or classification of events, without a clear aspect of prediction. For example, articles such as Kovačević et al. 2013 [19] and Jørgensen et al. 2020 [20] describe extraction and classification respectively. The goal of the study selection is to select articles that focus on a prediction method, such as the prediction of events (e.g., incidents or the result of a treatment) based on data of an earlier moment (e.g., the previous day or data recorded before or during the treatment).

2.3. Quality assessment

The next step in the review protocol was the quality assessment. The goal is to further assess the included studies and remove studies that did not fit with the goal of this review or which did not describe the details that are important for this review in enough detail. With this goal in mind the assessment criteria described in Table 2 were established. For this step the articles were read and special attention was given to the assessment criteria questions. For each of the quality assessment criteria described in Tables 2, 1 point could be obtained if the criterion is met, or half a point if it is partially met. This gave a quality score between 0 and 8 for each study. Articles with a score above 4.0 are included in this review.

Table 1
Study exclusion criteria.

No.	Exclusion criteria description
EC1	Papers not available in English or Dutch
EC2	Papers without full text available
EC3	Duplicate publication from multiple sources
EC4	Papers that do not describe the use of ML in the abstract
EC5	The ML application described does not focus on textual data
EC6	No event prediction, occurrence prediction, or result of treatment prediction
EC7	Papers that are literature reviews.

Table 2
Study quality assessment criteria adapted from Kitchenham et al. [17].

Nr.	Quality assessment question
vQ1	Are the aims of the study clearly stated?
Q2	Is the dataset used in the study clearly described?
Q3	Is the underlying mechanism of the method clearly described?
Q4	Is the method reproducible?
Q5	Does the conclusion describe the main findings?
Q6	Are limitations of the approach mentioned?
Q7	Are accuracies related to the methods and results mentioned?
Q8	Are negative findings presented?

2.4. Data extraction

To extract the relevant information from the articles that remain after selection and quality assessment, a data extraction form was constructed. This was used as a guide during the SLR and supported the extraction of the data necessary to answer the research question.

This process was performed in two separate steps. First a general information extraction (Table 6 in Appendix), and second an extraction of quantitative performance metrics (Table 7 in Appendix). In the general information extraction, details relating to the article, as well as information about the performed research that is mentioned in the text of the articles, is collected. In the extraction of the algorithm performance results, all articles were reviewed for all machine learning algorithms that were used. All the given performance metrics that could be compared of these methods were extracted. This produced a considerable variety of different metrics and situations, in which the methods were tested. 7 out of the 35 articles [21–27] did not give any clear or comparable results for the described algorithms. These were excluded from this part of the data extraction and further synthesis. The published performance metrics, described in the lower part of Table 7, were normalized to be between 0 and 1 to allow for comparison. Some of the results which could not be normalized in a logical manner were left out of this review.

2.5. Data synthesis

In the data synthesis, the data extracted from the primary studies was collated and summarized. To provide an overview of all the different methods that are used to predict events from text, the results of all the studies selected after the quality assessment were employed.

To compare the performances of different methods or algorithms for the task of event prediction from text, the results are visualized from the 28 out of 35 articles that did have comparable performance metrics for the described algorithms, as further shown and described in Section 3.2. To allow for wider comparison, the machine learning algorithms applied by the primary studies were grouped by their method. The 89 identified algorithms were synthesized into 32 classes (detailed in Table 8 in Appendix). Not all methods could be assigned to a class as their approach was too specific or unclear. For example, in Feller et al. [28], baseline and a unigrams model performance results are compared to an LDA model performance, but the method behind the baseline and unigram models are not explained clearly. These methods were not included in the algorithm performance comparison results of this study in Section 3.2.

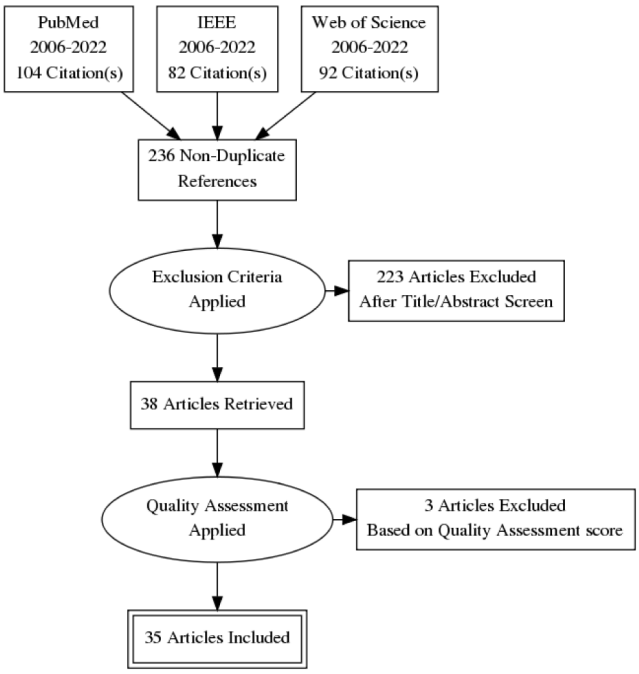

Fig. 2. PRISMA statement flow diagram.

This review does not give a direct comparison between algorithms, which can be used to prove that one method performs better than another. The goal of the data synthesis is to show the machine learning algorithms that give the highest scores for different performance metrics as reported by the primary studies.

3. Results

3.1. Findings

The search strategy yielded a total of 236 non-duplicate articles. Each of the databases returned a similar number of results, as can be seen in the PRISMA diagram in Fig. 2.

Fig. 3 shows the year of publication of the studies before the study exclusion criteria were applied. It can be seen that the majority of these studies were undertaken within the past five years, with ~80% having been published after 2017.

After applying the exclusion criteria, 38 articles remained. These results of the exclusion criteria are shown in Table 3. Fig. 6 in Appendix shows the number of articles that are in- and excluded, separated by year. Inclusion rates of the years with more than 10 articles ranged between 5% and 25%.

In Fig. 4 the results of the quality assessment are shown. With a cut-off value of 4.0, four studies are excluded based on quality. The average quality of literature in our SLR according to our quality assessment criteria, after use of the exclusion criteria, has a 94% pass and without significant variability, as there are no outliers in the quality scores.

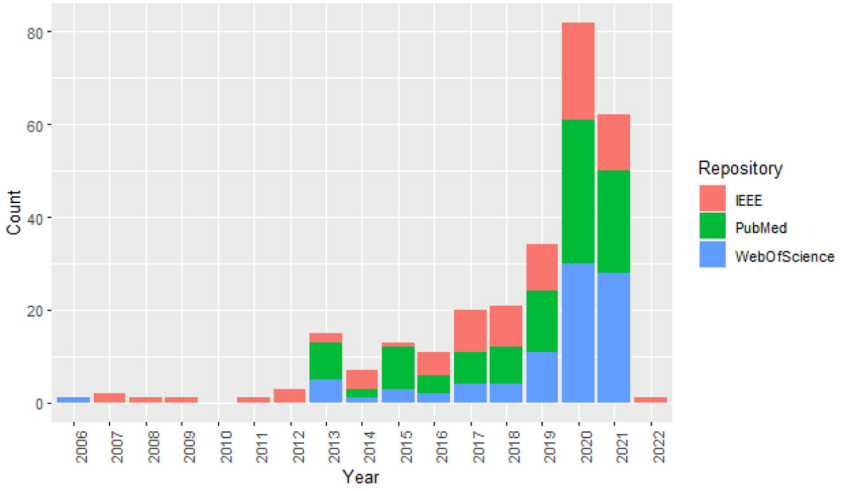

Fig. 3. Results of the search query used, shown separated by year of publication and coloured by the database source. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

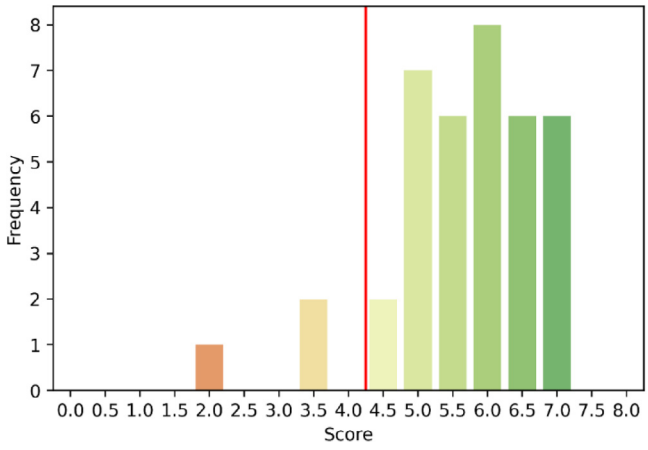

Fig. 4. Quality assessment scores by frequency. Cut-off score = 4.0.

Table 3
Results of study exclusion by database source.

	PubMed	IEEE	WebOfScience
Exclude	81	73	86
Include	23	9	7
Percentage (Include)	22%	11%	8%

28 of the 35 studies that passed the quality assessment had numerical results describing the performance of machine learning algorithms that could be compared to similar results from other methodologies. For example, Zhang et al. (2021) [29] and Sakarkar et al. (2021) [30] both report the F1-score, precision, sensitivity, and accuracy for their LSTM models that perform different tasks. Articles often describe the performance of multiple machine learning methods. The mean number of different methods compared in an article is 4, and 7 articles only show the performance of a single algorithm. Additionally, one or more algorithms are also regularly tested on different datasets or under different circumstances. 11 of the 28 articles show the performance results of the described methods on multiple datasets or with different settings. When either the method, the data or the settings are different, we process the results as a different situation. In these 28 articles, the performance metrics of 121 situations were documented with one or more of the performance metrics that could be adopted to compare algorithms. Table 4 shows the metrics compared and the number of times they are used in the selected studies. Table 5 shows how often the methods were used in separate selected studies and the total number of uses among all selected studies.

3.2. Algorithm performances

As the F1-score combines both precision and recall into a single metric, it makes it a suitable overall performance metric to compare methods. It is also one of the most widely used metrics in the primary studies, with it being used in 14 of the 28 articles, only falling behind the recall/sensitivity which was used in 17 of the studies. The F1-score also had highest number of total uses at 88. In Fig. 5, it can be seen that F1-scores in most studies range between 60% and 100% percent, with a few exceptions in LexRank, NRT and DCA. These 3 methods all came from a single article [31]. It was not made clear in what format their performance results were published, and they could not logically be compared to with methods from other studies as they were neither fraction or percentages. This meant that no data for these algorithms was available for this review.

The algorithms that were designed specifically for NLP, such as word2vec [32] and BERT [33], resulted in high F1-scores in Sun et al. [34], with for example 88.93% and 91.50% for word2vec and BERT respectively. These are, however, singular results from a single article [34]. The general multi-purpose machine learning methods seem to perform to a similar level and sometimes even higher. For example, the Recurrent Neural Networks (RNNs) [35] reported in the article by Sakarkar et al. (2021) [30] produced the highest F1-scores in this SLR study. Convolutional Neural Networks (CNNs) [36], Long Short-Term Memory (LSTM) [37] and Conditional Random Fields (CRF) [38] demonstrate the highest F1-scores of 0.985 [39], 98% [30] and 90.13% [40], among multiple different studies.

These types of algorithms mentioned above also perform with high scores in the other performance metrics, which can be seen in Figs. 6, 7, 8 and Appendix (Figs. 10–14). One of the other methods, which also perform well for these metrics such as the precision, specificity and accuracy, are simple tree based methods as in articles [41–43].

The Receiver Operating Characteristics Area Under the Curve (ROC AUC) is also one of the most commonly used methods, in the selected studies, for judging a model's performance. The articles that used LASSO Logistic Regression (LR) [44,45] reported the highest ROC-AUC scores compared to other methods. Of these 2 primary studies that included LASSO LR models, ROC AUC is the only shared metric.

The metrics NPV [42,46], MAE [47,48] and PR-AUC [28,49] were each used in only 2 of the selected studies. This makes them unsuitable as a comparison metric for the purpose of this study. For this reason, the approach did not involve factoring them into further consideration when comparing machine learning methods.

Table 4
Number of times each performance metric was used in the selected studies. The first row describes in how many separate articles each performance metric is used at least once. The second row describes the total number of times each performance metric is used.

	Total	F1	Precision/ PPV	Recall/ Sensitivity	Specificity	Accuracy	NPV	ROC AUC	PR-AUC	MAE
Used in n articles	28	14	13	17	8	9	3	10	2	2
Total uses	121	88	74	81	10	26	5	34	5	4

Table 5
Number of times each machine learning method was used in the selected studies. Unique uses are counted as the number of studies that used the method.

ML method	Unique uses	Total uses	ML method	Unique uses	Total uses
Tree	5	6	RNN	3	5
SVM	9	16	Gradient boosting	1	1
BERT	2	2	RF	4	5
word2vec	3	3	Elastic net	1	1
USE	1	1	LDA	1	1
KNN	1	1	SVC	1	1
LR	5	6	Rule-based	1	1
LSTM	4	5	DNN	1	2
BERT + LSTM	1	1	HAN	1	1
ML	4	8	CRF	2	8
LASSO	2	4	FOIL	1	2
XgBoost	1	1	LexRank	1	8
NN	1	1	NRT	1	8
SGD	1	1	DCA	1	8
ULMFiT	1	2	Naive Bayes	1	6
CNN	4	9			

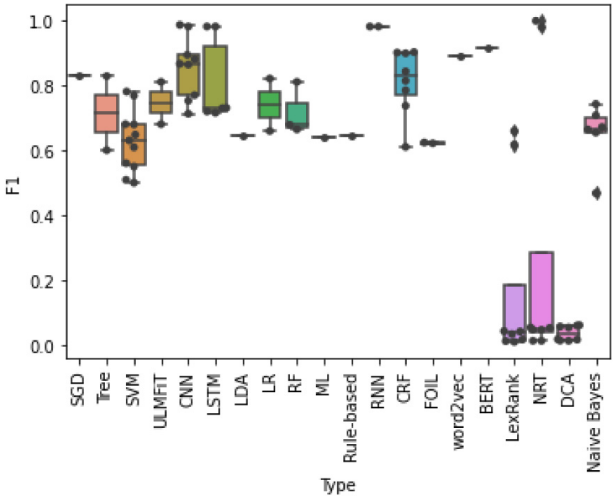

Fig. 5. F1-score of the types of methods in the selected studies.

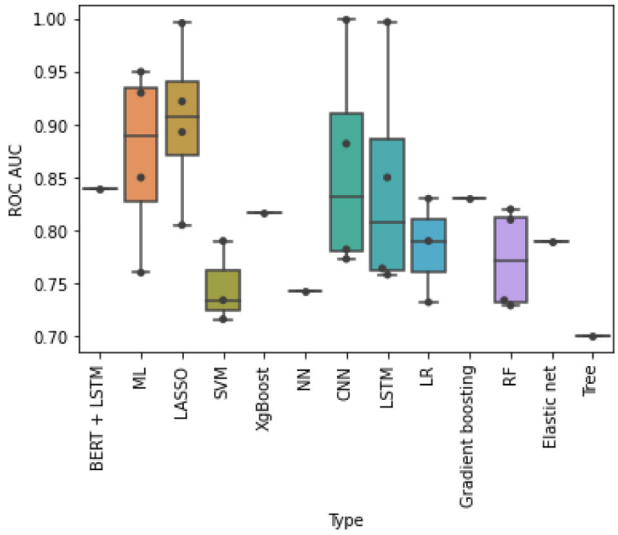

Fig. 6. Receiver operating characteristic area under the curve of the types of methods in the selected studies.

4. Discussion

In this discussion, an evaluation of the findings is presented, along with a reflection on the potential limitations of this study. At the time of writing this article, to the best of our knowledge, this is the first SLR comparing event prediction machine learning algorithms, used for different purposes, and using different data sources, that use text as the main data source.

By following the Kitchenham et al. [17] guidelines, 198 original articles were identified prior to the exclusion and quality assessment stages. As more than 80% has been published in the last 5 years (i.e. after 2017,) one can presume that this is a relatively recent area of machine learning research. Until 2020, every year increasing more articles were published about the topic of event prediction from text (see Fig. 3). However, this was not the trend for 2021–2022, which saw a decline. This decrease in publications since 2020 might suggest a decline in interest, but this is unclear until further investigation, yet the decrease could have also been caused by the Covid-19 pandemic, which caused a change of focus for many researchers within the medical data domain.

Fig. 6 clearly indicates that there is an increased percentage of inclusion for more recent articles. This might reinforce the idea that this is a young topic of research. There was, however, not enough data to verify if the selection criteria had a higher chance of excluding older studies.

The articles retrieved from PubMed were reviewed first, followed by papers from IEEE and then WebOfScience. Duplicate papers were marked as exclude as they appeared in the review. Because of this no duplicated were marked in PubMed and the most WebOfScience. This explains the higher percentage of inclusion for PubMed papers, as indicated in Table 3.

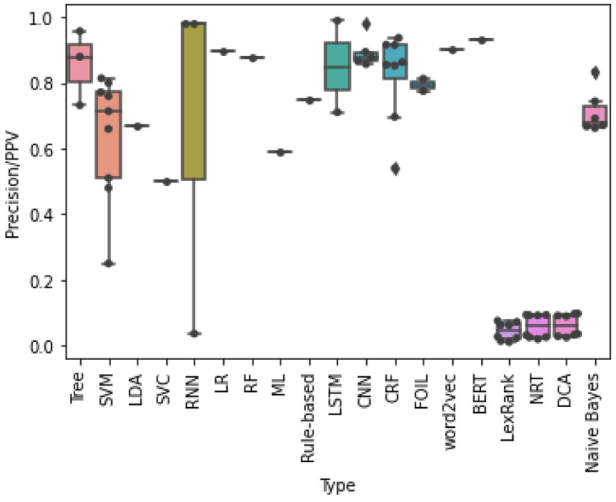

Fig. 7. Precision or PPV of the types of methods in the selected studies.

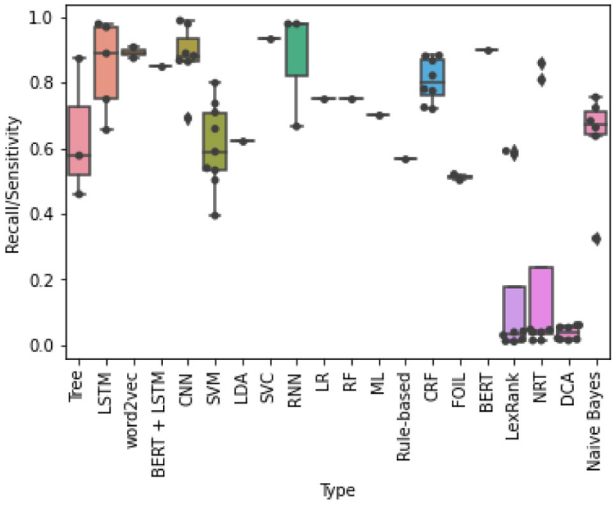

Fig. 8. Recall or Sensitivity of the types of methods in the selected studies.

4.1. Principal findings

Using natural language processing, computers are able to use the information available in text. With this information, the computer could predict events in the future. Sometimes this is done based on the text alone [50], but in many cases, it is combined with structured or numerical data to improve the prediction [49]. Some studies use complicated deep learning models that process temporal information to predict events [31]. Other studies use a more rudimentary approach, such as classification based on marker words in the text [43].

From assessment of all the selected studies it was clear that there is no set method that is considered the best performing at event prediction from text. Many different machine learning methods have been, and still are being used to predict events based on textual data. Although all these methods work in different ways, and not all of these types of algorithms can perform every task, they are valid methods of using machine learning to predict events from text.

As previously demonstrated, there is a large variation in the performance results of machine learning algorithms used for event prediction that is published in the selected articles. The most likely cause for the difference in performance, is that the algorithms used in the studies were used to perform widely different tasks. Additionally, even when tasks were similar between studies, the data used in each respective study could still significantly impact the performance. Simpler models, such as linear regression or decision tree models, sometimes perform well on some metrics such as precision [41], but they seem to show less consistent performance for the different metrics assessed in this review. It is notable that simple models could perform well on the task of event prediction from text; even when the problem is complex, as demonstrated in Hoogendoorn et al. (2017) [41]. The disadvantage is, however, that these simple models are also more specific and likely are less resistant to alterations in the data or experimental setup.

The more complex models that are based on neural networks are more versatile and seem to perform better across the different performance metrics assessed in this review. They are also more capable in a wider variety of tasks and are therefore more easily adapted for different tasks or datasets.

A selection was made to categorize the methods into 32 classes. This was necessary as otherwise each method could be considered slightly different, and all comparisons would be with singular results. Within the classes, differences between algorithms could be recognized and thus they could split into more or less than 32 classes. Examples of this are categories such as CRF and LSTM, which are based on RNNs. 32 classes was the result after assessing which methods had clear differences and which could be considered similar enough. Changes in the categorization could influence which performance results are associated with each other and, therefore, which methods seem the best performing.

It can be seen in Fig. 5 that, for most of the categorized methods, there are very few F1-scores. 50% of the machine learning algorithms had only 3 or less F1-scores reported in the selected articles. This number is even lower for all the other performance metrics. This means that for many of the methods there are not enough results to draw a conclusion about their performance. Only the most performance of the most commonly used algorithms can be compared in this study. More studies comparing event prediction from text need to be included to fairly assess the more specific methods, such as BERT [33].

For some of the metrics, such as the precision (see Fig. 7), the variation of scores within the algorithms is much larger than can be expected. As we explained above, this is likely caused by the method of reporting, making it not directly comparable to results from other studies. If a statistical analysis were preformed, outliers like the those shown in Fig. 7 would need to be corrected or removed first.

4.2. Strengths and weaknesses

Of the 198 studies, only 35 were selected for the study after the exclusion and quality assessment. The selected studies were of high quality according to the quality assessment, as only 3 studies did not achieve the cut-off score. Further, the included articles fitted the scope of this review well.

A limitation of this study is that there is a comparison of methods on different datasets. This is generally not recommended, as is shows an unfair comparison between the performance of algorithms. Yet, the goal of this review was to show the performance achievable by different machine learning methods on the tasks that they were deemed appropriate for by the authors of the primary studies. The published performance results are used in this study as an indicator of the result that can be achieved when using the different algorithms. It is not an exact comparison between performances of different methods, as for that purpose the same task and data should be used. For clarity, the findings of the SLR are used to compare the better results achieved between methods, and not the results for a single task on a specific dataset.

There were many difficulties with comparing reported performance metrics. Because there is not a standard for reporting machine learning method performance, all results had to be manually extracted from the

Table 6
Data extraction form to extract general information about the articles.

No.	Extraction element	Contents
General information		
1	DOI	
2	Title	
3	Authors	
4	Year of publication	
5	Repository	
8	SLR category	☐ Include ☐ Exclude
9	Notes about selection	
Description		
10	Keywords article	
11	Database terms	
12	Link	
13	Case study application	
14	Goal	
15	Approach	
16	Techniques	
17	Tools used	
Evaluation		
18	Quality assessment	Q1: Q2: Q3: Q4: Q5: Q6: Q7: Q8: tot:
19	QA notes	

Table 7
Machine Learning algorithm and performance statistics extracted from the selected studies.

Extraction element	Examples
Algorithm	SVM, HNN, Doc2Vec, ULMFiT, etc.
Class of Algorithm	SVM, RNN, word2vec,
Task	Classification, Prediction, Identification, Regression, Entity recognition
Performance Metric	Explanation
F1-score	Harmonic mean of precision and recall of a classifier: $\frac{TP}{TP+\frac{1}{2}(FP+FN)}$
Precision/PPV	Fraction of positive observed items among all observed items: $\frac{TP}{TP+FP}$
Recall/Sensitivity	True positive rate/ Fraction of observed positive items among all positive items: $\frac{TP}{TP+FN}$
Specificity	True negative rate/ Fraction of observed negative items among all negative items: $\frac{TN}{TN+FP}$
Accuracy	Proportion of correct predictions among total number of cases: $\frac{TP+TN}{TP+TN+FP+FN}$
NPV	Negative Predict Value: $\frac{TN}{TN+FN}$
ROC AUC	Area Under the Curve Receiver Operating Characteristics
PR-AUC	Precision–Recall Area Under the Curve (Average Precision)
MAE	Mean Absolute Error

articles. This creates many opportunities for making mistakes. The lack of a standard method to compare the performance between algorithms and between datasets, makes it difficult to compare published performance results. It is often unclear what exactly a performance metric is describing, and many studies show unexpected results. An example of this we encountered were reported F1-scores above 1, while it should be a percentage (0%–100%) or a fraction (0 to 1).

Another challenge encountered was the low number of results for some performance metrics. The metrics MAE and PR-AUC were only used in 2 different selected studies, as outlined in Table 4. This limits the applicability of the metric in comparing the performance of event prediction methods. The same problem existed for the algorithms. Because many of these algorithms were only used by a few studies, it was not possible to formulate an average performance for these methods, and this resulted in a reliance on the results of a low number of sources (>1 <2).

From the results it is clear that application of neural network-based algorithms, such as CNN [36], RNN [35] and LSTM [35], result in the most consistently high average F1, precision, recall and accuracy scores. This cannot be seen as proof that these methods will give the best performance for a given tasks, but it does indicate that these methods show the most potential and are often worth considering. Similarly do the algorithms that were specifically designed for natural language processing, such as BERT [33], show potential. These do however not have enough published performance results for the tasks of event prediction to be certain about their reliability.

5. Conclusions

This article reviewed 35 articles on event prediction based on textual data by machine learning algorithms published over 3 separate bibliographic databases by means of an SLR approach. The reported performance of these algorithms was compared on multiple different metrics such as the F1-score, precision, recall and accuracy. As demonstrated, numerous different tasks and datasets were used by the selected studies, giving a general overview of the performance of different machine learning methods. In the selected articles 89 different machine learning methods are used to predict events from textual data. Both complex methods based on neural networks, as well as simple methods focusing on single purpose tasks (e.g. classification), can be used for this purpose. Not enough performance results are published under comparable circumstances to give conclusive results about which methods perform the best. An indication about which machine learning methods perform well according to the authors of the selected studies are provided, which show that Neural network-based algorithms have the most consistently high performance over the different metrics. The

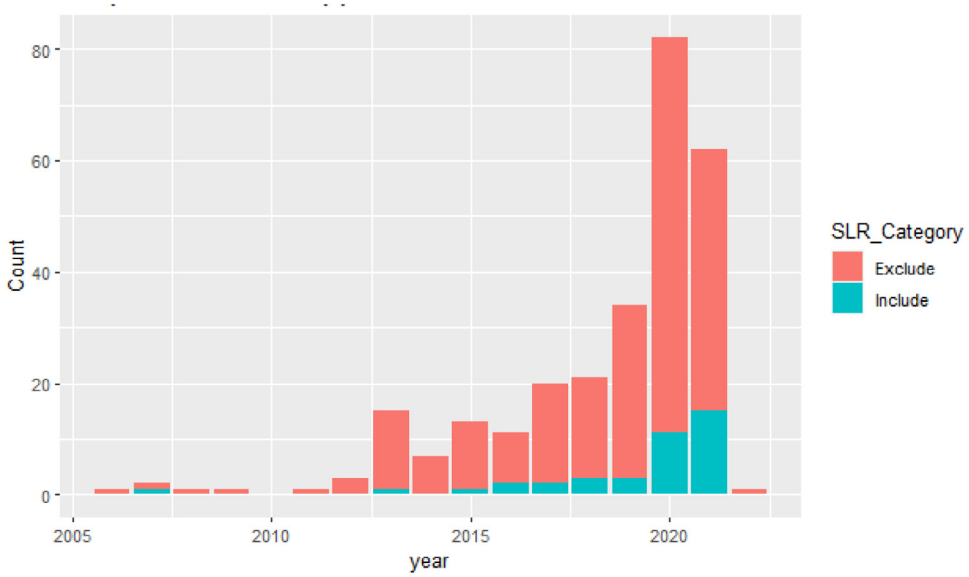

Fig. 9. Results of the study selection by year of publication.

Table 8
Explanation of the classes the machine learning methods were divided into.

Class	Explanation
Tree	Simple Tree based classification/regression model
SVM	Support Vector Machine
Classifier	General classification model
BERT	Bidirectional Encoder Representations from Transformers
word2vec	Neural network model to learn word associations from a large corpus of text
USE	Universal Sentence Encoder
KNN	K-Nearest Neighbours algorithm
LR	Logistic Regression
LSTM	Long Short-Term Memory artificial RNN
BERT + LSTM	Sequential use of BERT and LSTM
ML	General Machine Learning (not further specified in article)
LASSO	Least Absolute Shrinkage and Selection Operator
XgBoost	eXtreme Gradient Boosting
NN	(general) Neural Network
SGD	Stochastic Gradient Descent
ULMFiT	Universal Language Model Fine-tuning for Text
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
Gradient boosting	Prediction model in the form of an ensemble of weak prediction models, usually decision trees
RF	Random Forest
Elastic net	Extension of linear regression that combines penalties from LASSO and Ridge Regression
LDA	Latent Dirichlet allocation
SVC	Support Vector Machine
Rule-based	Simple rule based algorithm
DNN	Deep Neural Network
HAN	Hierarchical Attention Network
CRF	Conditional Random Field Logistic Regression
FOIL	First-Order Inductive Learner
LexRank	Unsupervised stochastic graph-based method for computing relative importance of textual units
NRT	Neural Rating Regression with Abstractive Tips Generation for Recommendationv
DCA	Deep Concept-aware Model
Naive Bayes	Classification technique based on Bayes' Theorem

state of the art in NLP models, such as BERT also perform well at event prediction in the 3 selected studies [34,48,51], we included in this review but more research would help to prove this. Under specific conditions, simple models also could perform well, but they are generally less versatile. To validate the results of this review and the performance metrics on which they were based, we would like to see the methods being tested on the same dataset and under similar conditions. Additionally, for future research we would recommend that an investigation into a standardized method for publishing machine learning method results is done. This would help massively in being able to compare performance results between different studies.

Appendix

A.1. Tables

See Tables 6–8.

A.2. Figures

See Figs. 9–14.

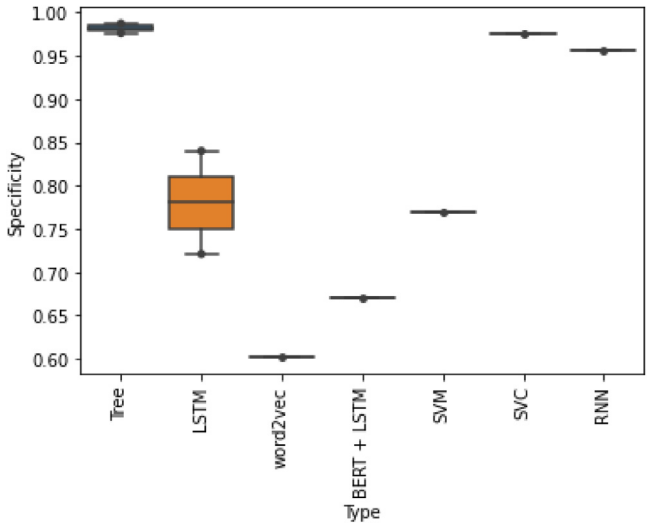

Fig. 10. Specificity of the types of methods in the selected studies.

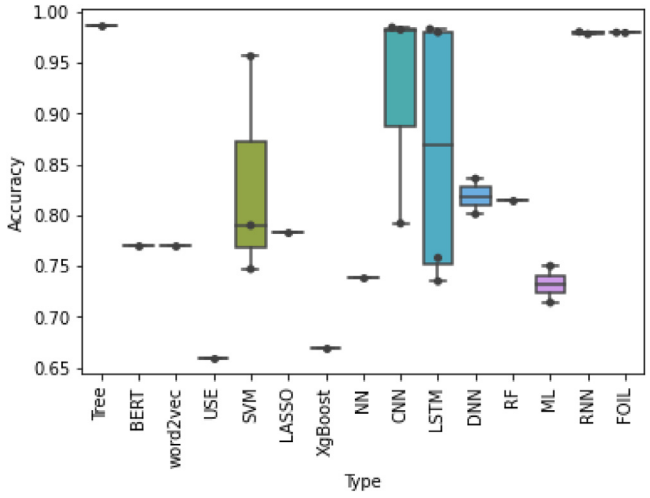

Fig. 11. Accuracy of the types of methods in the selected studies.

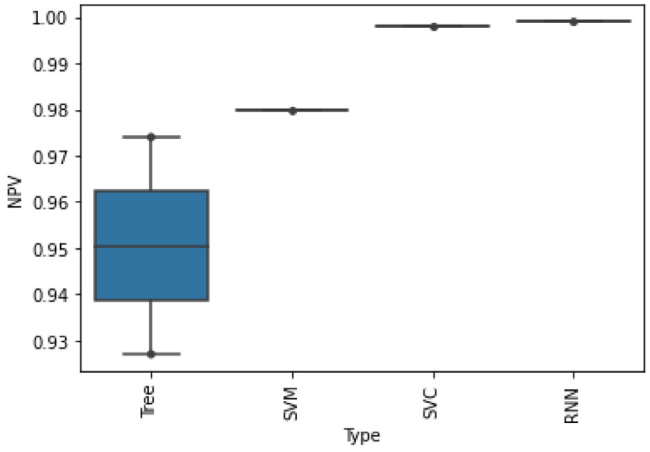

Fig. 12. Negative Predictive Value of the types of methods in the selected studies.

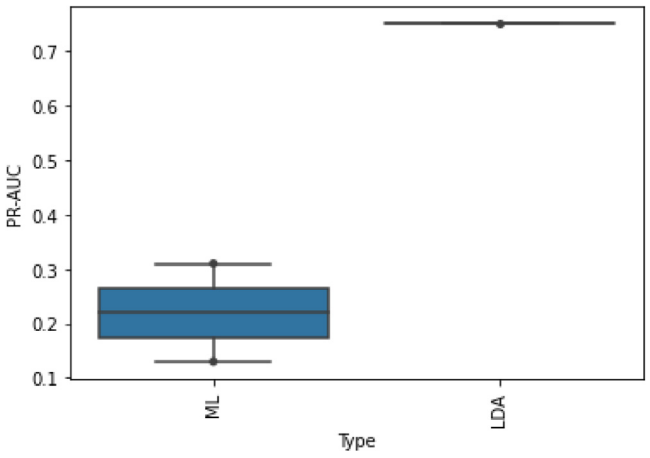

Fig. 13. Precision-Recall area under the curve of the types of methods in the selected studies.

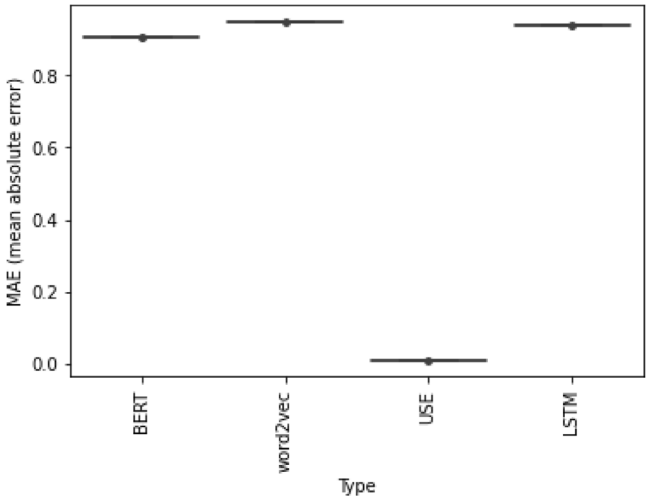

Fig. 14. Mean absolute error of the types of methods in the selected studies.

References

Studies included as primary studies in SLR: [21–31], [34], [39–50], [52–62]

[1] R.C. Schank, Computer understanding of natural language, Behav. Res. Methods Instrum. 10 (2) (1978) 132–138, http://dx.doi.org/10.3758/BF03205115.

[2] R.C. Schank, C.J. Rieger, Inference and the computer understanding of natural language, Artificial Intelligence 5 (4) (1974) 373–412, http://dx.doi.org/10.1016/0004-3702(74)90003-4.

[3] T. Winograd, Understanding natural language, Cogn. Psychol. 3 (1) (1972) 1–191, http://dx.doi.org/10.1016/0010-0285(72)90002-3.

[4] S. Joseph, K. Sedimo, F. Kaniwa, H. Hlomani, K. Letsholo, Natural language processing: A review, Nat. Lang. Process.: Rev. 6 (2016) 207–210.

[5] E. Ghazizadeh, P. Zhu, A systematic literature review of natural language processing: Current state, challenges and risks, Adv. Intell. Syst. Comput. 1288 (2021) 634–647, http://dx.doi.org/10.1007/978-3-030-63128-4_49/FIGURES/7.

[6] J. Hirschberg, C.D. Manning, Advances in natural language processing, Science (1979) 349 (6245) (2015) 261–266, http://dx.doi.org/10.1126/SCIENCE.AAA8685/ASSET/D33AB763-A443-444C-B766-A6B69883BFD7/ASSETS/GRAPHIC/349_261_F5.JPEG.

[7] C. Stephanidis, E eAccessibility.

[8] C.S.G. Khoo, S. Chan, Y. Niu, Extracting causal knowledge from a medical database using graphical patterns, 2000, pp. 336–343, http://dx.doi.org/10.3115/1075218.1075261.

[9] E. Blanco, N. Castell, D. Moldovan, Causal relation extraction, 2008.

[10] R.F. Cekinel, P. Karagoz, Event prediction from news text using subgraph embedding and graph sequence mining, World Wide Web (2022) 1–26, http://dx.doi.org/10.1007/S11280-021-01002-1/FIGURES/17.

[11] K. Liu, Y. Chen, J. Liu, X. Zuo, J. Zhao, Extracting events and their relations from texts: A survey on recent research progress and challenges, AI Open 1 (2020) 22–39, http://dx.doi.org/10.1016/J.AIOPEN.2021.02.004.
[12] B.G. Patra, et al., Extracting social determinants of health from electronic health records using natural language processing: A systematic review, J. Am. Med. Inform. Assoc. 28 (12) (2021) 2716–2727, http://dx.doi.org/10.1093/JAMIA/OCAB170.
[13] I.J.B. Young, S. Luz, N. Lone, A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis, Int. J. Med. Inform. 132 (2019) http://dx.doi.org/10.1016/J.IJMEDINF.2019.103971.
[14] G. Alfattni, N. Peek, G. Nenadic, Extraction of temporal relations from clinical free text: A systematic review of current approaches, J. Biomed. Inform. 108 (2020) 103488, http://dx.doi.org/10.1016/J.JBI.2020.103488.
[15] Y. Boneski Gumiel, et al., Temporal relation extraction in clinical texts, ACM Comput. Surv. 54 (7) (2021) http://dx.doi.org/10.1145/3462475.
[16] L. Zhao, Event prediction in the big data era, ACM Comput. Surv. 54 (5) (2021) http://dx.doi.org/10.1145/3450287.
[17] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering – A systematic literature review, Inf. Softw. Technol. 51 (1) (2009) 7–15, http://dx.doi.org/10.1016/J.INFSOF.2008.09.009.
[18] P. Alderson, S. Green, J.P.T. Higgins, Cochrane Reviewers' Handbook 4.2.2, 2004, http://www.cochrane.org/resources/handbook/hbook.htm. (Accessed 20 January 2022).
[19] A. Kovačević, A. Dehghan, M. Filannino, J.A. Keane, G. Nenadic, Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives, J. Am. Med. Inform. Assoc. 20 (5) (2013) 859–866, http://dx.doi.org/10.1136/AMIAJNL-2013-001625.
[20] A. Caccamisi, L. Jørgensen, H. Dalianis, M. Rosenlund, Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records, Ups J. Med. Sci. 125 (4) (2020) 316–324, http://dx.doi.org/10.1080/03009734.2020.1792010.
[21] J.T. Baillargeon, L. Lamontagne, E. Marceau, Mining actuarial risk predictors in accident descriptions using recurrent neural networks, Risks 2021 9 (1) (2020) 7, http://dx.doi.org/10.3390/RISKS9010007, 9, 7.
[22] B.S. Zanotto, et al., Stroke outcome measurements from electronic medical records: Cross-sectional study on the effectiveness of neural and nonneural classifiers, JMIR Med. Inform. 9 (11) (2021) e29120, http://dx.doi.org/10.2196/29120, 9 (11) (2021) e29120, https://medinform.jmir.org/2021/11/e29120.
[23] K. Kreimeyer, et al., Feature engineering and machine learning for causality assessment in pharmacovigilance: Lessons learned from application to the FDA adverse event reporting system, Comput. Biol. Med. 135 (2021) 104517, http://dx.doi.org/10.1016/J.COMPBIOMED.2021.104517.
[24] J.L. Izquierdo, J. Ancochea, J.B. Soriano, Clinical characteristics and prognostic factors for intensive care unit admission of patients with COVID-19: Retrospective study using machine learning and natural language processing, J. Med. Internet Res. 22 (10) (2020) http://dx.doi.org/10.2196/21801.
[25] Z.T. Korach, et al., Unsupervised machine learning of topics documented by nurses about hospitalized patients prior to a rapid-response event, Appl. Clin. Inform. 10 (5) (2019) 952–963, http://dx.doi.org/10.1055/S-0039-3401814.
[26] N. Rybak, M. Hassall, Deep learning unsupervised text-based detection of anomalies in U.S. chemical safety and hazard investigation board reports, in: International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2021, Oct, 2021, http://dx.doi.org/10.1109/ICECCME52200.2021.9590834.
[27] Y. Kim, S. Park, J. Lee, D. Jang, J. Kang, Integrated survival model for predicting patent litigation hazard, Sustainability 2021 13 (4) (2021) 1763, http://dx.doi.org/10.3390/SU13041763, 13 (2021) 1763.
[28] D.J. Feller, J. Zucker, M.T. Yin, P. Gordon, N. Elhadad, Using clinical notes and natural language processing for automated HIV risk assessment, J. Acquir. Immune Defic. Syndr. 77 (2) (2018) 160–166, http://dx.doi.org/10.1097/QAI.0000000000001580.
[29] X. Zhang, P. Srinivasan, S. Mahadevan, Sequential deep learning from NTSB reports for aviation safety prognosis, Saf. Sci. 142 (2021) 105390, http://dx.doi.org/10.1016/J.SSCI.2021.105390.
[30] Dr. G. Sakarkar, et al., Advance approach for detection of DNS tunneling attack from network packets using deep learning algorithms, ADCAIJ: Adv. Distrib. Comput. Artif. Intell. J. 10 (3) (2021) 241–266, http://dx.doi.org/10.14201/ADCAIJ2021103241266.
[31] H. Liao, X. Zhang, X. Li, M. Zhou, A. Vidmer, R. Mao, A deep concept-aware model for predicting and explaining restaurant future status, in: Proceedings - 2020 IEEE 13th International Conference on Web Services, ICWS 2020, 2020, pp. 559–567, http://dx.doi.org/10.1109/ICWS49710.2020.00081.
[32] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2022, [Online]. Available: http://ronan.collobert.com/senna/. (Accessed 27 May 2022).
[33] J. Devlin, M.-W. Chang, K. Lee, K.T. Google, A.I. Language, BERT: Pre-training of deep bidirectional transformers for language understanding, Naacl-Hlt 2019, no. Mlm, 2018, 2022, [Online]. Available: https://github.com/tensorflow/tensor2tensor. (Accessed 26 May 2022).
[34] D. Sun, Y. Peng, H. Li, Construction of knowledge graph of HIV-associated neurocognitive disorders syndrome based on deep learning, in: Proceedings - 2020 International Conference on Artificial Intelligence and Computer Engineering, ICAICE 2020, 2020, pp. 134–141, http://dx.doi.org/10.1109/ICAICE51518.2020.00032.
[35] A. Sherstinsky, Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, Physica D 404 (2020) 132306, http://dx.doi.org/10.1016/J.PHYSD.2019.132306.
[36] K. O'Shea, R. Nash, An Introduction to Convolutional Neural Networks, 2015, http://dx.doi.org/10.48550/arxiv.1511.08458.
[37] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/NECO.1997.9.8.1735.
[38] J. Lafferty, A. Mccallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
[39] J.S. Obeid, et al., Identifying and predicting intentional self-harm in electronic health record clinical notes: Deep learning approach, JMIR Med. Inform. 8 (7) (2020) http://dx.doi.org/10.2196/17784.
[40] G. Moharasan, T.B. Ho, Extraction of temporal information from clinical narratives, J. Healthcare Inform. Res. 3 (2) (2019) 220–244, http://dx.doi.org/10.1007/s41666-019-00049-0.
[41] M. Hoogendoorn, T. Berger, A. Schulz, T. Stolz, P. Szolovits, Predicting social anxiety treatment outcome based on therapeutic Email conversations, IEEE J. Biomed. Health Inform. 21 (5) (2017) 1449–1459, http://dx.doi.org/10.1109/JBHI.2016.2601123.
[42] D.A. Szlosek, J.M. Ferretti, Using machine learning and natural language processing algorithms to automate the evaluation of clinical decision support in electronic medical record systems, EGEMS (Wash DC) 4 (3) (2016) 5, http://dx.doi.org/10.13063/2327-9214.1222.
[43] J.J. Sivaraman, S.K. Proescholdbell, D. Ezzell, M.E. Shanahan, Characterizing opioid overdoses using emergency medical services data : A case definition algorithm enhanced by machine learning, Public Health Rep. 136 (1_suppl) (2021) 62S–71S, http://dx.doi.org/10.1177/00333549211026802.
[44] S.T. Parker, Estimating nonfatal gunshot injury locations with natural language processing and machine learning models, JAMA Netw Open 3 (10) (2020) http://dx.doi.org/10.1001/JAMANETWORKOPEN.2020.20664.
[45] K. de Silva, et al., Clinical notes as prognostic markers of mortality associated with diabetes mellitus following critical care: A retrospective cohort analysis using machine learning and unstructured big data, Comput. Biol. Med. 132 (2021) http://dx.doi.org/10.1016/J.COMPBIOMED.2021.104305.
[46] B.D. Wissel, et al., Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery, Epilepsia 61 (1) (2020) 39–48, http://dx.doi.org/10.1111/EPI.16398.
[47] J. Sanyal, A. Tariq, A.W. Kurian, D. Rubin, I. Banerjee, Weakly supervised temporal model for prediction of breast cancer distant recurrence, Sci. Rep. 11 (1) (2021) 1–11, http://dx.doi.org/10.1038/S41598-021-89033-6, 2021 11:1.
[48] Y. Zhang, et al., Applying artificial intelligence methods for the estimation of disease incidence: The utility of language models, Front. Digit. Health (2020) http://dx.doi.org/10.3389/FDGTH.2020.569261.
[49] B.D. Wissel, et al., Early identification of epilepsy surgery candidates: A multicenter, machine learning study, Acta Neurol. Scandinavica 144 (1) (2021) 41–50, http://dx.doi.org/10.1111/ANE.13418.
[50] Y. Nakajima, K. Takagi, M. Ptaszynski, H. Honma, F. Masui, A proposal of prediction method using word polarity information for future event prediction support system, in: Proceedings - 2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019, 2019, http://dx.doi.org/10.1109/ICAICTA.2019.8904426.
[51] F. Amrollahi, S.P. Shashikumar, F. Razmi, S. Nemati, Contextual embeddings from clinical notes improves prediction of sepsis, in: AMIA Annu Symp Proc, 2020, 2021, pp. 197–202, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33936391.
[52] D.M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, S.S. Ghosh, Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: Observational study, J. Med. Internet Res. 22 (10) (2020) http://dx.doi.org/10.2196/22635.
[53] S. Gupta, A. Belouali, N.J. Shah, M.B. Atkins, S. Madhavan, Automated identification of patients with immune-related adverse events from clinical notes using word embedding and machine learning, http://dx.doi.org/10.1101/2020.05.19.20106583.
[54] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 I2B2 challenge, J. Am. Med. Inform. Assoc. 20 (5) (2013) 806–813, http://dx.doi.org/10.1136/amiajnl-2013-001628.
[55] S. Sarkar, V. Lodhi, J. Maiti, Text-clustering based deep neural network for prediction of occupational accident risk: A case study, 2018, http://dx.doi.org/10.1109/iSAI-NLP.2018.8692881.
[56] C. Lin, et al., Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record, J. Am. Med. Inform. Assoc. 22 (e1) (2015) e151–e161, http://dx.doi.org/10.1136/amiajnl-2014-002642.

[57] G.E. Weissman, et al., Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay, Crit. Care Med. 46 (7) (2018) 1125–1132, http://dx.doi.org/10.1097/CCM.0000000000003148.

[58] T.R. Goodwin, D. Demner-Fushman, A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision, J. Am. Med. Inform. Assoc. 27 (4) (2020) 567–576, http://dx.doi.org/10.1093/jamia/ocaa004.

[59] K. Syed, W. Sleeman, M. Hagan, J. Palta, R. Kapoor, P. Ghosh, Automatic incident triage in radiation oncology incident learning system, Healthcare (Switzerland) 8 (3) (2020) http://dx.doi.org/10.3390/healthcare8030272.

[60] F. Amrollahi, S.P. Shashikumar, F. Razmi, S. Nemati, Contextual embeddings from clinical notes improves prediction of Sepsis.

[61] S. Komaki, F. Muranaga, Y. Uto, T. Iwaanakuchi, I. Kumamoto, Supporting the early detection of disease onset and change using document vector analysis of nursing observation records, Eval. Health Professions 44 (4) (2021) 436–442, http://dx.doi.org/10.1177/01632787211014270.

[62] J. Poveda, M. Surdeanu, J. Turmo, A comparison of statistical and rule-induction learners for automatic tagging of time expressions in English, 2007, [Online]. Available: http://chasen.org/.