Ontologies and Machine Learning Models to Enhance Health Informatics: A Survey, Challenges and Future Directions

Hafida Tiaiba, *Member*, *IAENG,* Lyazid Sabri, Okba Kazar

***Abstract*— Medical reports and narratives especially present considerable challenges due to their complicated medical terminology, frequent use of abbreviations, and diversity of language structures. As a result of these complexities, it can be hard to extract and interpret meaningful information straightforwardly and consistently. It is essential for a comprehensive understanding of healthcare data in the medical domain to identify patterns, recognize subtle cues, and distill critical information from different sources. Natural language processing, machine learning, and semantic analysis play fundamental roles in overcoming these obstacles. These tools streamline the information retrieval process and enable the discovery of hidden correlations and trends within medical texts, as a result, more informed decision-making is enabled and healthcare insights are enhanced. This study analyzes a range of academic publications, including books, doctoral theses, and articles, and it conducts a comparative study of recent survey papers to address limitations in prior research. While previous surveys have focused on specific areas like named entity recognition (NER), relationship extraction, text vectorization, and classification methods, this research adopts a broader perspective by exploring various aspects of information extraction and medical document classification. The study highlights the critical purpose of ontologies, especially medical ones, in knowledge representation. These ontologies ensure semantic interpretation, reduce ambiguity, and enhance information sharing among researchers and healthcare professionals. Additionally, the research emphasizes the benefits of integrating multiple datasets and combining machine learning methods with ontologies to improve the accuracy and efficiency of medical text analysis providing better decision-making opportunities.**

***Index Terms*—Health informatics, Medical unstructured texts, Embedding, Machine Learning, Natural language processing, Ontology**

I. INTRODUCTION

IN the medical field, the significance of classification, clustering, and prediction as fundamental data mining parameters cannot be overstated. These techniques present a key in unraveling complex patterns, identifying behavioral trends, and predicting potential outcomes, thereby contributing crucial insights for effective diagnosis, treatment, and intervention strategies in healthcare. Therefore, text mining and Natural Language Processing (NLP) are strongly involved in medicine, pharmacology, and the natural sciences. Nevertheless, analyzing medical documents using text mining is a complex domain that requires considerable time and effort. The storage, collection, and transmission of knowledge encoded in hospital reports, particularly chronic disease data, are controlled by textual knowledge management. Today's medical environment uses data mining to predict various illnesses, assist in diagnosis, and advise physicians in clinical decision-making. Likewise, data mining has far greater potential to provide question-based answers, anomaly-based detection, more informed decision-making, probabilistic measurements, predictive models, and decision support. Moreover, many data sources such as symptoms, exams, patient history, procedures, treatments, and medications allow rapid exploration of diseases. Therefore, the analysis of medical texts requires knowledge in a variety of fields, including clinical-specific areas, data mining, text mining, statistics, medical texts, and clinical and hospital procedures.

The definition of text-mining is the extraction of hidden and valuable information from unstructured texts. So only in 1999, the author of this study [1] consider Text Data Mining (TDM) as a variant of Data Mining (DM), since these methods (as in data mining) allow extracting knowledge from the web. A text mining process can be defined as a knowledge-intensive activity in which analytical tools are used to identify and explore hidden patterns among documents. However, clinical texts written by clinicians describe the patients' pathologies, the social and medical history, and the observations made during the interview or the care procedures. The term "clinical texts" covers the full range of narratives appearing in the medical patient record [2] or texts used to manage medical, financial, administrative, and legal aspects of a hospital [3]. In data mining, different methods are used to discover useful information. These include association, clustering, classification, prediction, and sequential patterns [4]. In fact, many text-mining techniques derive from data mining, including summarization and entity relations [5].

Developing classification models for clinical documents can be challenging as the medical texts' structure is heterogeneous. In most cases, these documents consist of narrative text. Additionally, doctors use particular jargon such as abbreviations and disease codes, which requires additional tools for interpreting the designated terms and extracting semantic information. The text classification

Manuscript received June 24, 2024; revised February 19, 2025.

Hafida Tiaiba is a PhD in Department of Computer Science, Mohamed El Bachir El Ibrahimi University, Bordj Bou Arreridj, Algeria (corresponding author e-mail: hafida.tiaiba@univ-bba.dz).

Lyazid Sabri is a supervisor of Intelligent Systems and Cognitive Computing (ISCC) Laboratory, Mohamed El Bachir El Ibrahimi University, Bordj Bou Arreridj, Algeria (e-mail: lyazid.sabri@univ-bba.dz).

Okba Kazar is a Professor in College of Arts, Sciences and Information Technology, Department of Computer Science, University of Kalba, Sharjah, United Arab Emirates (e-mail: Okba.Kazar@ukb.ac.ae).

consists of partitioning the data into pre-defined groups identified by their labels. For example, to predict the categorical label y_i for data instance x_i as that of learning a function f [6]. Where, for the documents

$X = (x_1, x_{2, \dots,} x_n) \,/\, i \in [1..n]$ and the classes

$Y = (y_1, y_{2, \dots,} y_m) \,/\, j \in [1..m]$: $f(x_i) = y_j$

The idea of the mind map method pioneered by Tony Buzan [7] is utilized in this study to enhance readers' comprehension of the content. To achieve this, each section is summarized with a figure.

The structure of the paper is as follows: First, the paper presents the main challenges and future directions in the field of medical text classification, followed by a comparison of recent surveys on medical texts. Next, the paper reviews related works on information extraction. Moreover, an updated view of text representation methods, natural language pre-processing and machine learning techniques for extracting medication information are provided. In addition, given the importance of ontological semantic technology in handling natural language meaning, the use of ontologies to classify medical documents is examined. Furthermore, the methods for evaluating the model and estimating it are explained. Finally, the paper concludes with discussion and recommendations for future researches.

II. CURRENT CHALLENGES AND FUTURE DIRECTIONS

The major challenge for medical text classification is the structure of the medical text itself. The vocabulary used by physicians to write conclusions differs from that in other fields. Hence, using NLP tools to extract the meaning of a text and deduce the context is only sometimes obvious. This text can hold intricate vocabulary with medical terms, abbreviations, acronyms, local dialectal terms, mistakes, and misspellings (Fig. 1). Moreover, clinical narrative texts containing personal health knowledge require data privacy and special processing to keep the information secure. Another challenge is that minor clinical corpus size influences the inferred embedding quality. Indeed, embedding induced from a large corpus encode more information than those generated from a small corpus. In the clinical field, available corpora are small compared to general corpora. For example, the Google News Corpus consists of about 100 billion tokens, MIMIC clinical notes consist of just about 0.53 billion tokens and the PubMed corpus consists of around 4.35 billion tokens [8]. Models such as Word2Vec, GloVe, and FastText, among others, allocate a singular representation that overlooks the multifaceted nature of meanings, consequently diminishing the quality of the inferred embedding. For instance, aspirin is utilized in the treatment of both fever and cardiovascular diseases. Nevertheless, it is important to note that the representation of text significantly influences the efficacy of a model in subsequent tasks. On the other hand, extracting information from clinical textual data is a complex and challenging task in NLP. Temporal information and extraction of temporal relations [9] between clinical events play essential roles in clinical assessment and decision-making. Therefore, extracting relationships from clinical textual data is challenging as it lies between medical NLP, temporal representation, and temporal reasoning. As highlighted by authors in the study [10], converting clinical text into causal knowledge (i.e., causal relationships) is the most difficult and complex challenge. Extracting clinical events as well as temporal knowledge is a challenging task. That is why existing systems frequently have numerous independent components based on a collection of norms or classical machine learning models [11].

The increasing medical vocabulary is a consequence of new diseases, symptoms, drugs, etc. Today, ontologies constitute an ineluctable design to reduce this ambiguity by providing a generic conceptualization of notions, especially in medicine. Therefore, using ontologies has become necessary to extract terms and their synonyms from the text to identify concepts and handle the nontrivial aspects of spatial-temporal reasoning based on semantic causal explanation and semantic analysis of relationships between events. Nevertheless, defining and using ontologies in the medical field presents a significant challenge. We highlight that combining machine learning technics and ontology should increase the efficiency of models for knowledge extraction in medical reports.

III. ONTOLOGY: SEMANTIC DISAMBIGUATION EXAMPLES

Transforming these two example text1 and text2 (excerpt from the PubMed 200k-RCT dataset [12]) into numerical representations show the ontologies' usefulness for word-sense disambiguation purposes in medical reports. For text1 and text2, we employed Word2Vec using the Skip-gram and Continuous Bag of Words (CBOW) methods. The objective is to retrieve words that are similar to the word pain. The word pain is a disease according to the European Federation, but it's also a sensory and emotional experience according to disease ontology. The main goal is to identify semantic relationships among various words and contexts, with vector dimension parameters adjusted across one, two, and three dimensions.

The comparison of vectorization methods across the obtained results highlights how CBOW and Skip-gram models represent clinical vocabulary, emphasizing the need to distinguish between terms with different meanings in medical texts. While models capture semantic relationships, words like "pain" have specific meanings that differ from common terms such as "years," "as," "was," and "to the." These terms may appear in medical contexts but serve different roles, such as indicating time, comparison, or tense, rather than describing a medical condition like "pain". Despite the model chosen for vectorization, which reflects semantic structures, ontologies are essential to accurately interpret these distinctions. For example, "pain" refers to a physical or psychological condition, while "years" indicates duration, and "as" is a comparative term. Their meanings shift depending on context, but they do not alter the core meaning of terms like "pain."

In TABLE I different word vectorization techniques (CBOW and Skip-gram with Bigrams and Trigrams) applied to text1 with specific vocabulary sizes (CBOW with v = 108), Skip-gram Bigrams with (v = 110), and Skip-gram Trigrams with (v = 112). Each method produces a distinct set of vocabulary, which reflects its ability to capture different aspects of the clinical text. CBOW includes terms like "therapy," "efficacy," and "hypertension," indicating a focus on general treatment and condition-related terms. In contrast, Skip-gram with Bigrams includes additional terms such as "randomly," "mg/day," and "receive," which suggest a deeper focus on treatment administration. Skip-gram with Trigrams extends this further, capturing terms like "taking," "diarrhea," and "monitoring," pointing toward patient symptoms and monitoring processes. These differences reveal that while CBOW provides a generalized semantic representation, the Skip-gram models, particularly with Trigrams, capture a richer and more detailed context by including terms related to dosing, patient demographics, and side effects. Thus, Skip-gram Trigrams may offer enhanced utility for applications requiring detailed clinical data extraction and precise interpretation of patient information in medical text analysis.

Similarly, TABLE II compares the performance of different word vectorization methods (CBOW and Skip-gram) applied to text2 from a specific vocabulary size, with CBOW at (v = 89) and Skip-gram at (v = 95). Each method yields distinct vocabularies, where CBOW does not include the word "pain," highlighting a potential limitation in capturing certain clinical terms relevant to pain management. The context words captured in each model further reveal the differences: CBOW generally includes words associated with clinical actions and contexts, such as "combinations," "postoperative," and "injected." In contrast, the Skip-gram approach, particularly with Bigrams and Trigrams, captures a wider range of clinical terms, including "analgesic" and "narcotics," which may provide a richer context for understanding patient treatments and outcomes. These Skip-gram representations suggest an improved ability to capture nuanced medical language, as they include more varied and specific terms related to pain and postoperative care. Overall, the table indicates that Skip-gram models, especially those using bigrams and trigrams, may offer better semantic representations of complex clinical vocabulary compared to CBOW, enhancing the model's potential for accurately interpreting medical narratives.

In summary, while Skip-gram models, especially with higher-order n-grams, perform better in capturing complex medical terms and context, careful attention must still be given to interpreting words that may have varied or ambiguous meanings. This highlights the importance of further development of vectorization methods, including their incorporation at domain-specific knowledge and ontologies level to enhance models accuracy while dealing with clinical texts.

TABLE III presents the numerical representations (embeddings) of the word "pain" (using text1) and a set of semantically similar words across three dimensions, illustrating the relationships between terms within a vector space. Each dimension captures unique values for "pain" and its related terms, such as "therapy," "efficacy," and "blood," which help define their semantic proximity. In Dimension 1, "pain" has a value of [-0.08035642], while similar terms like "therapy" and "hypertension" show varied values, such as [-0.4403279] for "1-2" and [-0.7665176] for "efficacy," aligning with treatment-related terms. Dimension 2, represented by a two-value vector, captures "pain" as [-0.04514251, 0.4190363], with nearby terms like "abdominal" and "therapeutic" also positioned closely, reflecting descriptive or contextual features in medical language. In Dimension 3, "pain" is expressed as [-0.248204, -0.2452565, -0.08156046], alongside terms like "blood" and "monitoring," which show similar vector patterns, potentially indicating associations with physiological or monitoring aspects. Altogether, these dimensions reveal how embedding techniques structure complex medical terminology, providing interpretable relationships that highlight semantic similarity and context in clinical narratives.

TABLE IV provides numerical embeddings for the word "pain" (using text2) and related terms across three distinct dimensions, illustrating how embedding models capture word meanings based on contextual similarity. In dimension 1, "pain" is represented by the vector [-0.37439027], with similar terms like "combinations," "postoperative," and "protocol" holding values such as [-0.02600246] [-0.44625682], and [-0.76460844], suggesting associations with treatment procedures and medical protocols. Dimension 2 shows "pain" as [-0.40871736, -0.46739444], alongside terms like "combination" and "receiving" with vectors like [-0.20421192, -0.3823596], [-0.24343628, -0.18285151], indicating connections to procedural efficacy and patient treatment contexts. In dimension 3, "pain" is represented by [0.18839891, -0.2540631, -0.12656955], with related terms such as "experienced," "preoperative," and "minutes" showing vectors like [0.10089256,-0.22837418,-0.04392448], [0.07130048,-0.26479003,-0.08826005], and [0.04674968,-0.0908818,-0.14382423], capturing temporal and procedural elements tied to patient experiences. Together, these dimensions reveal layered relationships between "pain" and similar terms, which aid in understanding clinical vocabulary in structured, context-driven spaces (a valuable approach for precise language interpretation in medical narratives).

Word vectorization depends on various factors: corpus, dimensions, and vectorization model to use (e.g. CBOW or Skip-gram). Vector representation's dimensionality affects how words are positioned within the semantic space. Furthermore, the choice of the corpus is crucial, where, a medical corpus, for example, will produce embeddings that are more specialized in healthcare terminology, while a general corpus might fail to capture domain-specific nuances such as those related to pain, treatments, and medical conditions. On top of these factors, the upcoming section covers the intrinsic limitations regarding word vectorization techniques. While they can capture semantic similarity between words, they may not fully account for the complex relationships and nuances present in clinical narratives, particularly when terms carry multiple meanings or depend heavily on the context. Hence, the addition of domain-specific knowledge (e.g., by linking the model representation with ontologies) could lead to a more accurate understanding and creation of medical language.

Based on the results obtained, it is recommended to integrate medical ontologies to further refine the semantic understanding of medical terminology and improve the model's performance. The code for these examples is available at the GitHub link https://github.com/T-

HAFIDA/Word2Vec.

Text1: "We studied 58 patients with grade 1-2 essential hypertension (25 men and 33 women) , 48.7 (11.9) years of age , randomly assigned to receive torasemide (5 mg/day) either upon awakening or at bedtime . Blood pressure was measured by ambulatory monitoring for 48 consecutive hours before and after 6 weeks of therapy . Efficacy of torasemide was significantly higher with bedtime dosing (11.2 and 8.0 mmHg reduction in the 24-hour mean of systolic and diastolic blood pressure , respectively) as compared to the administration of the drug on awakening (6.2 and 3.7 mmHg reduction in systolic and diastolic blood pressure) . The percentage of patients with controlled ambulatory blood pressure after treatment was also higher after bedtime treatment (54 % versus 27 %) . The time-response curves indicate a full 24-hour therapeutic duration only when torasemide was administered before bedtime . With regard to the safety profile , 2 patients presented secondary effects (abdominal pain , diarrhea) in morning dose , and 4 patients taking the drug at bedtime reported nicturia . "

Text2: "Fifty patients successfully completed the study protocol . Patients receiving combinations of morphine , bupivacaine , and epinephrine or bupivacaine and epinephrine yielded lower pain scores and narcotics consumption than patients receiving epinephrine alone , which was statistically significant irrespective of the timing of injection (P < .0001) . Patients receiving the study medication preoperatively had significantly lower pain scores at the first measurement (t = 0) than those receiving the study medication postoperatively (P = .0343) . There was no statistically significant effect of timing of the treatment medication administration at either 60 or 120 minutes postoperatively . Comparison of fentanyl consumption between groups receiving the treatment medication preoperatively versus postoperatively showed no significant difference .The combination of morphine , bupivacaine , and epinephrine , as well as the combination of bupivacaine and epinephrine provide excellent postoperative pain control when used either preoperatively or postoperatively in knee arthroscop . There was a trend that patients receiving preoperative analgesic injections experienced superior pain control than did those injected postoperatively . "

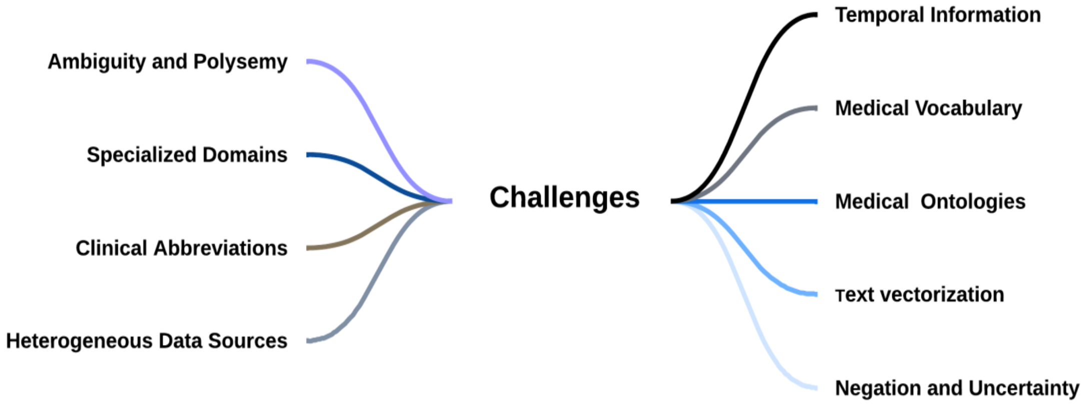

Fig. 1. The Principal challenges of clinical text.

TABLE I
THE TEN MOST SIMILAR WORDS TO THE WORD PAIN: TEXT1 WITH THE MODELS CBOW, SKIP-GRAM (BIGRAMS AND TRIGRAMS) FOR ONE, TWO, AND THREE DIMENSIONS (DIM). "V" DENOTES THE SIZE OF THE VOCABULARY

DIM	CBOW v= 108	Skip-gram (Bigrams) v = 110 (vocabulary)	Skip-gram (Trigrams) v= 112
1	[’1-2’, ’grade’, ’therapy’, ’efficacy’, ’25’, ’hypertension’, ’essential’ ’years’, ’)’, ’nicturia’]	[’randomly’, ’1-2’, ’nicturia’, ’upon’, ’mg/day’, ’5’, ’receive’, ’2’, ’grade’, ’33’]	[’taking’, ’58’, ’morning’, ’33’, ’5’,’men’,’4’, ’diarrhea’, ’therapy’,] ’monitoring’
2	[’morning’,’in’, ’for’, ’abdominal’, ’therapeutic’, ’3.7’, ’studied’,’at’, ’women’, ’presented’]	[’effects’,’secondary’,’48’, ’to the’, ’on’, ’2’, ’men’, ’years’, ’patients’, ’assigned’]	[’in’, ’treatment’, ’5’, ’ and’,’,’, ’versus’, ’3.7’, ’patients’, ’54’, ’age’]
3	[’men’, ’indicate’, ’monitoring’, ’hours’, ’to’,’blood’, ’curves’, ’mmhg’, ’as’, ’weeks’]	[’dosing’, ’randomly’, ’torasemide’, ’we’, ’only’, ’respectively’, ’.’, ’33’, was’,’as’,’receive’]	[’mean’, ’25’, ’as’, ’essential’,’54’, ’monitoring’,’torasemide ’hours’,’administration’,’.’]

TABLE II
THE TEN MOST SIMILAR WORDS TO THE WORD PAIN: TEXT2 WITH THE MODELS CBOW, SKIP-GRAM (BIGRAMS AND TRIGRAMS) FOR ONE, TWO, AND THREE DIMENSIONS (DIM). "V" DENOTES THE SIZE OF THE VOCABULARY

DIM	CBOW
	v=89
		Skip-gram(Bigrams)
		v=95
			Skip-gram (Trigrams)
			v=95

1	[’combinations’, ’which’,’postoperative’, ’arthroscop’, ’t’, ’protocol’, ’in’, ’alone’, ’at’, ’injected’]	The word ‘pain’ is not included in the vocabulary	[’0’, ’narcotics’, ’trend’,, ’arthroscop’, ’used’, ’postoperative’, ’significantly’, ’analgesic’,’of’, ’combinations’]
2	[’120’, ’)’, ’successfully’, ’,’, ’receiving’, ’combination’,’excellent’, ’between’,’or’, ’treatment’		[’,’, ’combinations’, ’receiving’, ’consumption’, ’did’,, ’medication’, ’significant’, ’study’,’minutes’, ’and epinephrine’]
3	[’the’, ’experienced’, ’preoperative’, ’t’, ’than’,’ and’, ’injected’, ’)’,’minutes’, ’provide’]		[’irrespective’, ’had’, ’arthroscop’, ’that’, ’significantly’, ’experienced’, ’) .’, ’patients’, ’injections’, ’study’]

TABLE III
NUMERICAL REPRESENTATION OF TEN MORE SIMILAR WORDS TO ’PAIN’ IN THE TEXT1, USING CBOW

Numerical representation of the word: pain	Similar words	Numerical representation of similar words
Dimension 1: [-0.08035642]	[’1-2’, ’grade’, ’therapy’, ’efficacy’, ’25’, ’hypertension’, ’essential’, ’years’,’)’,’nicturia’]	[[-0.4403279] [-0.8769923] [-0.48153663] [-0.3108462] [-0.7665176] [-0.02992293] [-0.00310332] [-0.41340765] [-0.05348039] [-0.5937165]]
Dimension 2: [-0.04514251, 0.4190363]	[’morning’,’in’, ’for’,’abdominal’, ’therapeutic’, ’3.7’, ’studied’, ’at’, ’women’, ’presented’]	[[-0.03414124, 0.38525942] [-0.04363778, 0.28883508] [-0.0430724, 0.27686176] [-0.00874919, 0.17324676] [-0.08426419, 0.3351643] [0.0070153, 0.15392697] [-0.04220278, 0.14129873] [-0.17484386, 0.47061527] [0.04652259, 0.3195849] [-0.15218966, 0.3950207]]
Dimension 3: [-0.248204, -0.2452565 -0.08156046]	[’men’, ’indicate’ ,’monitoring’, ’hours’, ’to’, ’blood’, ’curves’, ’mmhg’, ’as’, ’weeks’]	[[-0.23371187, -0.30359426, -0.01094574] [-0.17208306, -0.23322713, -0.1610568] [-0.2975674, -0.2330507, 0.03119168] [-0.095222, -0.20392987, -0.01255612] [-0.2942829, -0.14061719, 0.00221788] [-0.24695018, -0.07242171, -0.06219727] [-0.23660916, -0.25768462, -0.30295923] [-0.08901064, -0.28964493, -0.02691508] [-0.19178434, -0.15672255, -0.24436885] [-0.19845465, -0.05969124, -0.14281404]]

TABLE IV
NUMERICAL REPRESENTATION OF TEN MORE SIMILAR WORDS TO ’PAIN’ IN THE TEXT2, USING CBOW VECTORIZATION

Numerical representation of the word pain	Similar words	Numerical representation of similar words
Dimension 1: [-0.37439027]	[’combinations’, ’which’,’postoperative’, ’arthroscop’ ’t’, ’protocol’, ’in’, ’alone’,’at’, ’injected’]	[[-0.02600246] [-0.8715803] [-0.44625682] [-0.41450462] [-0.815236] [-0.76460844] [-0.35509628] [-0.43826872] [-0.19329406] [-0.47433007]]
Dimension 2: [-0.40871736, -0.46739444]	[’120’, ’)’, ’successfully’, ’,’, ’receiving’,’combination’, ’excellent’, ’between’, ’or’, ’treatment’]	[[-0.22952682, -0.25210172] [-0.11299069, -0.09376078] [-0.44509122, -0.351733] [-0.46091023, -0.35077196] [-0.24343628, -0.18285151] [-0.20421192, -0.3823596] [-0.14129485, -0.30785048] [-0.18319613, -0.43615544] [-0.3731453, -0.19463071] [-0.43532884, -0.21746512]]
Dimension 3: [0.18839891, -0.2540631, -0.12656955]	[’the’, ’experienced’, ’preoperative’, ’t’, ’than’,’and’, ’injected’, ’)’,’minutes’, ’provide’]	[[0.2950617, -0.31452298, -0.23044617] [0.10089256, -0.22837418, -0.04392448] [0.07130048, -0.26479003, -0.08826005] [0.2594061, -0.31887758, -0.00474538] [0.26621985, -0.20084439, 0.00535775] [0.20968314, -0.11466264, -0.02830276] [0.2506244, -0.12669171, -0.01810657] [0.14730744, -0.13950509, 0.0308135] [0.04674968, -0.0908818, -0.14382423] [0.25347137, -0.2860597, 0.11112635]]

IV. Comparison with earlier surveys

TABLE V summarizes articles in medical text processing type surveys, from 2015 to 2023. In [13] automated de-identification of medical free text by applying LSTM methods is discussed. More than half of the approaches to automatically de-identify free text data rely on either rule-based systems or hybrid models combining machine learning with rules In this study, we found that LSTM based model outperformed Conditional Random Fields (CRF) and Rule-based systems. The results for hybrids and ensembles of LSTMs were even poorer compared to the LSTM-only models.

To improve understanding of the literature regarding the use of NLP for classifying incident reports and analyzing adverse events, the researchers conducted a systematic review and synthesis in [14]. Their specific objectives were to understand, the techniques employed, and to highlight areas of future research in this field. Complementing this study, [15] described the different types of clinical reports that were classified automatically using text classification and NLP. In addition, they conducted a comprehensive review of the datasets that have been employed for the purpose of clinical report classification. Furthermore, they identified various pre-processing and data sampling techniques, analyzing the different feature sets, feature representation, and feature reduction. Likewise, [16] presents a comprehensive analysis of methodologies, obstacles, and tools utilized in text mining as it pertains to medical documentation, decision-making assistance, health administration, and classification frameworks.

A key aspect of medical text processing is the International Classification of Diseases (ICD) coding task, which [17] emphasizes as critical. The authors reviewed recent research in this area, translating the problem into a learnable study, and highlighted two essential tasks: document representation learning for diagnostic information and ICD code classification. Additionally, [18] discusses the application of patient safety ontologies for document classification, detailing methods for model evaluation. Furthermore, [19] reviewed the methodologies of clinical concept extraction, cataloging development processes, available methods and tools, and specific considerations when developing clinical concept extraction applications. This review discusses the critical steps of clinical concept extraction application development, the trends and associations of clinical concept extraction research over different approaches, and the main problems in this field. This study shows that the method adopted for a specific task can be impacted by five factors: data and resource availability, domain adaptation, model interpretability, system customizability, and practical implementation. Supporting this, [20] examines the current state of NER and Relational Extraction (RE) techniques, identifying the F1-score as the most common metric. The study confirms that other metrics, such as sensitivity, specificity, Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC), may also be useful for evaluating NER and RE in future research. Building on temporal aspects, [21] explores time modeling in the clinical domain, with a focus on ontology-based representations and temporal reasoning. This study highlights the management of temporal information within standardized clinical models. Likewise, [22] summarizes research on event and event relation extraction, outlining these tasks in medical text analysis.

Furthermore, [9] surveys existing temporal relation (TLINK) extraction methods in English clinical text noting, the main challenges in TLINK extraction. Furthering the conversation on evaluation, it is described in [23] how NLP algorithms can be used to map clinical text ontology concepts to the heterogeneity of methodologies. Moreover, [24] investigates global NLP applications in healthcare, focusing on clinical corpora development across languages and demonstrating the relevance of multilingual corpora for unstructured clinical text extraction.

Following this theme, [25] reviews how unstructured text from Electronic Health Record (EHR) data is used in developing and validating the models. In parallel, [26] examines EMR text-based case detection for specific clinical conditions, detailing information extraction methods and the added benefits of using text data over structured data alone. Additionally, [27] presents EMR data preprocessing, with applications in medical decision support, risk prediction, mobile health, network medical treatment, and drug reaction detection. Complementary to this, [28] defines techniques and applications of text mining in fields such as digital libraries, social media, and business intelligence.

Biomedical text summarization is the focus of [29], which reviews recent biomedical text summarization applications in literature and EHR documents, examining techniques, applications, and evaluation methods. Finally, the study [30] investigates the representation of textual semantics via the utilization of word embedding.

This comprehensive overview captures significant advancements and methodologies in medical text processing, as documented in recent survey articles. The field is marked by varied approaches, particularly in automating de-identification, classification, and clinical information extraction. Studies reveal that LSTM-based approaches for de-identification outperform traditional CRF and rule-based models. However, hybrid models are sometimes less effective due to potential overfitting issues, especially with datasets like MIMIC. Further findings from systematic reviews indicate that the scope of NLP in healthcare has expanded to tasks such as incident report classification, adverse event analysis, and clinical report classification. The reviews emphasize a variety of preprocessing and sampling techniques, feature representation, and reduction strategies. Notably, standardized metrics (such as F1-score, sensitivity, specificity, and AUC) are emerging as essential benchmarks in clinical entity and RE, though the reviews recommend future studies apply these metrics more consistently. The ICD coding task is identified as critical for clinical text processing, where accurate document representation learning and effective classification are essential. Ontologies, such as those related to patient safety, have shown utility in document classification, underscoring the importance of model interpretability and customizability. Temporal reasoning and event relation extraction are also noted as growing fields, aiming to map temporal information within clinical narratives effectively. Additionally, multilingual corpora development is highlighted as a crucial area for extending NLP's cross-linguistic applicability in healthcare.

Overall, while NLP in medical text processing has made significant strides, key challenges remain. Notably, hybrid model optimization is necessary to prevent overfitting in specialized datasets. Standardizing evaluation metrics across studies will further strengthen NLP applications by enabling consistent comparisons. Multilingual corpora with standardized labeling protocols could broaden NLP's clinical research utility, making it applicable across languages and regions. There is also a recognized need for continued refinement in document representation learning, especially for ICD coding and TLINK extraction, to enhance diagnostic tools, decision-making, and clinical timeline analysis. Although healthcare NLP has advanced considerably, optimizing hybrid models, advancing TLINK extraction, and promoting global corpus development remain essential to achieving scalable, effective NLP solutions for clinical applications.

TABLE V
DISTRIBUTION OF SURVEY'S TOPICS BY YEARS

Survey Topics	2015	2016	2018	2019	2020	2021	2022	2023	Total
Machine Learning Models	[16]	[21], [26]	[27]	[2], [14], [15]	[9], [13], [22], [23]	[17]-[19], [23], [29]			16
Classification Performance	[16]	[26]	[27]	[14], [15]	[9], [13], [23]	[20], [24]			10
Preprocessing Techniques				[15]	[9]				2
Feature Extraction				[15]					1
Feature Representation				[30]					1
Use of ontologies		[21]		[15]		[24]	[18]	[23]	5
Methods Development ontologies							[18]		1
ICD coding	[16]					[17]			2
Temporal Information		[21]			[9]	[23]			3
NER Extraction						[18]-20]			3
Relation Extraction						[20]	[25]		2
Event Extraction					[22], [23]	[20]			3
Causal Relation					[22]				1
Temporal Relation					[22], [23]	[31]			3
Evaluation Methods						[24]		[23]	2
Text Summarization						[29]			1

The Total in the last column reflects the total number of survey articles identified for each topic across the specified years. For instance, in the realm of machine learning models, 18 survey articles were identified, spanning the years 2015 to 2023. This table provides a comprehensive overview of the evolution of survey articles, offering insights into the trends and focus areas within the field of medical text processing over the examined period. Researchers and practitioners can utilize this information to navigate the wealth of knowledge generated in the literature.

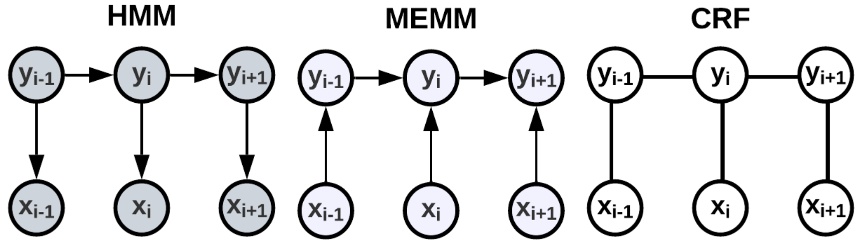

Fig. 2. Graphical Representation of Linear-Chain HMM, MEMM, and CRF (redrawing based on [32]).

V. INFORMATION EXTRACTION (IE)

The primary facets of information extraction include NER, events, and relation extraction. The rule-based approach, exemplified by the REgenstrief EXtraction (REX) tool [33], relies on "if-then" rules to match conditions in the text and extract entities. Studies such as [34], and [35] have successfully employed the REX model for NER. Beyond rule-based approaches, statistical methods, including Hidden Markov Models (HMMs), have found application in various domains, especially, part-of-speech tagger [36]. For identifying Protected Health Information [37] utilized a non-parametric Bayesian and Hidden Markov Model, showcasing its efficacy in capturing subtle variations in data

Additionally, Maximum Entropy Markov Models (MEMMs), proposed by [38], amalgamate the strengths of HMMs and maximum entropy models, exhibiting superior performance. In conclusion, CRF is highly successful for information retrieval and closely related to direction. Linear-chain conditional random fields, as used by [39], and [40], simplify predictions by considering only immediately neighboring tags, showcasing their utility in Named Entity Recognition, especially when combined with models like BiLSTM-CRF [41].

Fig. 2 delineates the variations among linear sequences of hidden Markov models, maximum entropy Markov maximum entropy Markov models. CRFs, being undirected graphical patterns, can handle dependencies in any models, and conditional random fields. This emphasizes the distinct attributes and utilizations of each methodology. Furthermore, alternative algorithms are utilized, as depicted in Fig. 5.

VI. TEXT VECTORIZATION METHODS

Manipulating textual data using data mining models requires transforming text into a digital format. Among the first methods is the One-hot–encoding representation (i.e.; representing the word by a single vector). In this instance, the vectors are expressed independently of the words' contexts. For vocabulary with V size, for each ith word w_i, the word w_i is represented by a vector where 1 is assigned to the ith element, and 0 is assigned to the rest of the elements of the vector [42].

The authors in [43] explored region embedding via one-hot representation. The concatenation of a one-hot vector gives us Bag of Words (BOW), where the major disadvantage is the length of its characteristic vector [44]. Term-Frequency Inverse Document Frequency (TF-IDF) is an extension of the Term Frequency model and the Inverse Document Frequency model [30]. Authors in [45] and [46] highlighted that word embedding outperform topic modeling and TF-IDF.

$$\mathrm{TF}(w_i) = \frac{\text{Number of times } w_i \text{ appear}}{\text{Total number of words}} \tag{1}$$

$$\mathrm{IDF}(\ w_i\ ) = \log \frac{N}{\mathrm{DF}(\ w_i\ )} \tag{2}$$

N represents the total number of documents, while DF(w_i) represents the number of documents containing the term w_i. TF-IDF presented in (3) is calculated from (1) and (2), as follows:

$$\mathrm{TF}-\mathrm{IDF}(w_i) = \mathrm{TF}(\ w_i\ ) \times \mathrm{IDF}(\ w_i\ ) \tag{3}$$

Moreover, word embedding has become an integral component of many NLP tasks due to its widespread application to functions such as machine translation, chatbots, image legend generation, and language modeling. This technique gives a semantic representation of characteristics unlike other existing techniques. Models for learning word vectors fall into two categories: models based on global matrix factoring or models based on the local context window. Latent Semantic Analysis (LSA) [47], skip-gram and CBOW are other prediction algorithms based on global matrix factoring. LSA is an analysis technique that maps words in documents to concept, or a method of representing words in a document. For example, the Word2Vec method combines words which appear in similar ”contexts” and distances terms in different contexts by adjusting numbers in the vector. Authors in [27] used Word2Vec and the Chinese word segmentation with java implementation (Ansj) word segmentation tool. However, this method achieved an accuracy rate of 25%, presenting a meagre result. As for the thesis, [48] applied Word2Vec encoding to multiple datasets; in his study, the author obtained 0.9342 as the best result for COVID-19 data sets and 0.3164 as the worst result for Drugs.com data sets. Unlike the One-hot, BOW or word co-occurrence matrix, the Word2Vec representation vector's size is not dependent on vocabulary size [42]. Skip-gram, where exploiting the context, is the goal of this algorithm to predict a word’s neighboring. The objective of the Skip-gram model is to minimize the log probability error function [48], as follows:

$$-\sum_{t=1}^{T} \sum_{j=-m,j\neq 0}^{m} logP(w_{(t+j)}|w_t) \tag{4}$$

Where, T represents the number of words in the sequence (w_1,w_2,w_3,...,w_T) and m the size of the context.

$$-\sum_{t=1}^{T} \log P(w_t|w_{t-m},\ldots,w_{t-1},w_{t+1},\ldots,w_{t+m}) \tag{5}$$

Likewise, the CBOW approach permits a word’s prediction from neighboring words. CBOW approach has been used in many studies; for example, authors in [49] used CBOW and Skip-gram in gene and protein synonym recognition tasks. While authors in [50] rely on CBOW, Word2Vec methods, and the full MIMIC-III dataset to pre-train word embedding. To address the issue of polysemy in word vectors by generating distinct word vectors for different meanings and providing pre-trained models on extensive corpora, a continuous bag-of-words approach is used to pre-train word embeddings on the entire training dataset [6]. While in [51] the authors used Adaptive Skip-gram (AdaGram) model introduced by [52]. Conversely, Paragraph2vector, commonly referred to as Doc2Vec [8], constitutes an advancement of Word2Vec as proposed by [53]. Doc2vec functions as an unsupervised model adept at depicting textual data of varying lengths, including sentences, paragraphs, and full documents. This research [54] proposed the Global Vectors (GloVe) model. This model based on global matrix factoring effectively exploits the global corpus' statistics (e.g., the co-occurrence of words in global scope). To automate the assignment of higher-level codes from the ICD version 9 (ICD-9) using clinical records in human and veterinary databases, the GloVe model is employed as a vector representation in the research presented in [39]. Another model of text vectorization used by [43] is region embedding or the tv-embedding model (two-view embedding). Unlike word embedding, which embeds text words into vectors with fixed dimensions, an embedding is referred to as a two-view embedding if it retains the information necessary to predict one view (word or region) from another perspective (context word or region). Among the recent methodologies for vectorization BERT [55] employs transformer architecture, characterized by an attention mechanism that discerns contextual interrelations among lexical units (or sub-lexical units) within a text. The transformer is composed of two discrete elements: an encoder, which processes the text input, and a decoder, which generates predictions for the task. BERT generates vector representations for a token depending on its context. The biLSTM layers capture various types of semantic information about words in context, and utilizing all layers enhances overall task performance. In contrast, BERT's representations are jointly conditioned on the left and the right context across all layers. Moreover, SciBERT [56] is a pre-trained BERT-based language model for performing scientific tasks, BlueBERT [57] is a pre-trained model for the medical field, and BioBERT [58] is a pre-trained language model for the biological domain. On the other hand, Embeddings from Language Models (ELMo) [59] uses the left-to-the-right and the right-to-the-left LSTM concatenation. ELMo considers the entire phrase when assigning an embedding to each word, capturing both the context before and after the word to generate more dynamic and context-aware representations.

Meta-embedding learning is an embedding-based model that combines several existing embedding sets [60]. This method comes with two benefits. The first one enhances performance as it leverages multiple word embedding sets, the second enhances vocabulary coverage results from using various word embedding sets (CW [61], Huang [62], Glove, HLBL [63], and Word2Vec). In [45], authors used topic embedding to allow the representation of any document by a set of topics. Fig. 4 summarizes the different text vectorization approaches.

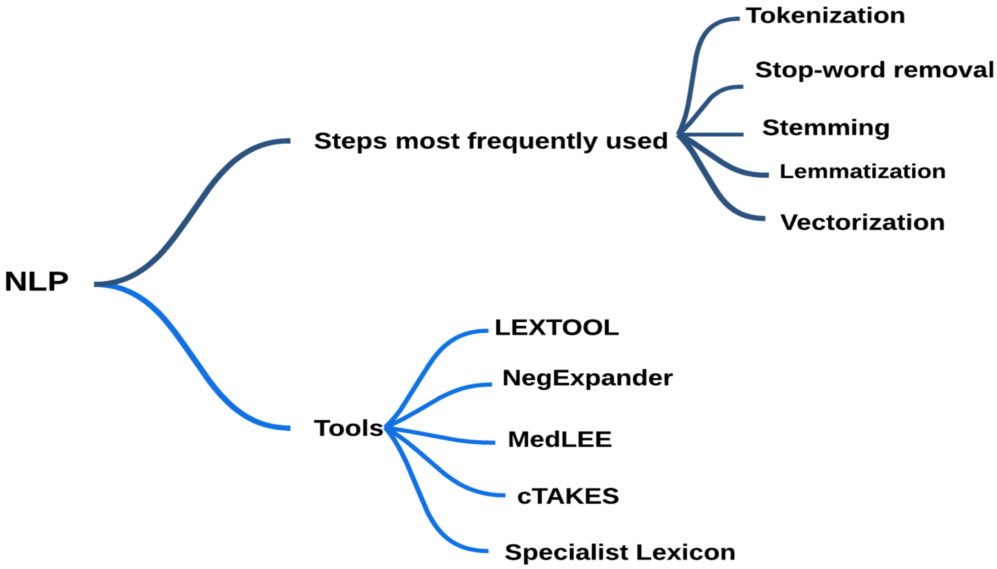

Fig. 3. NLP and used Tools in Medical Field.

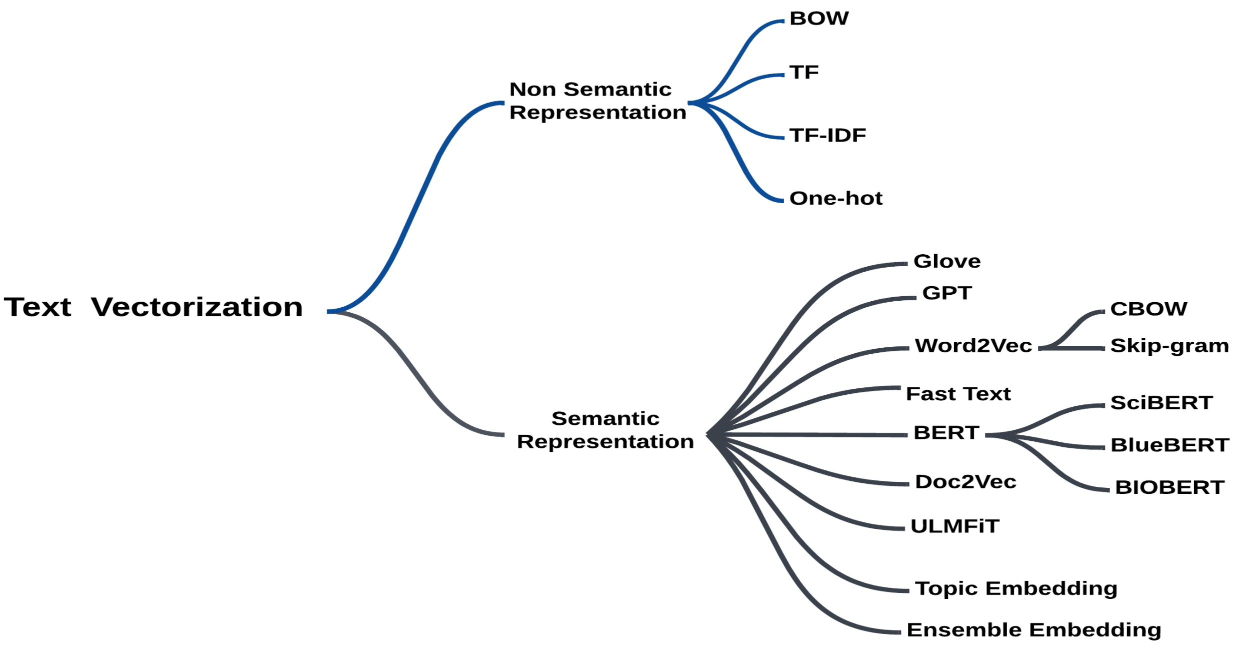

Fig. 4. The Key Types of Text Vectorization.

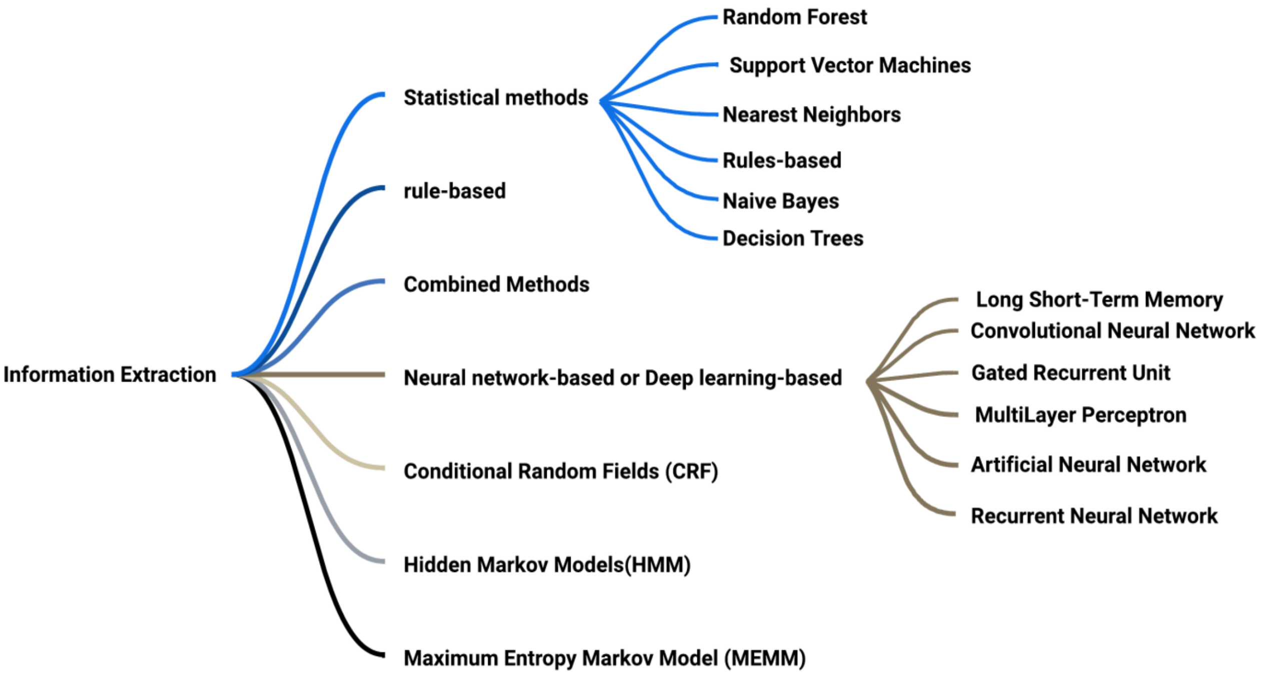

Fig. 5. The Main Algorithms used for Knowledge Extraction and NER extraction from Medical Text.

VII. NATURAL LANGUAGE PREPROCESSING

The text underscores the pivotal role of data preprocessing in enhancing data quality and mining results, with distinct approaches based on data types. For structured data, preprocessing involves cleaning, integration, transformation, and reduction. In contrast, for semi-structured or unstructured data, tokenization, stop-word removal, stemming, and vectorization methods are essential (Fig. 3). The sensitivity and complexity of preprocessing are magnified in the medical field, as elucidated by [2], due to factors like abbreviations, negations, spelling errors, and medical codes. Notably, medical record spelling mistakes are reported at a substantial rate of approximately 10% [2]. According to [64], medical text preprocessing includes unifying terms and acronyms using regular expressions and procedural methods. Vectorization methods are then applied to EHR. The information extraction tasks in medical texts, particularly NER and relationship extraction are crucial for feature detection. Works such as Specialist [65], LEXTOOL, and UMLS-based spelling correction tools [66] introduced linguistic analysis, morphological analysis, and spelling correction. Negation detection, highlighted by NegExpander [67], gains significance in medical texts, and temporal analysis, as proposed by [68], involves complex temporal models. Other studies, such as [69] and [70], propose modular architectures integrating NLP techniques and knowledge bases for processing temporal information in clinical narrative documents. The text also emphasizes the effectiveness of data augmentation for small datasets and advocates dimension reduction, employing machine learning approaches and methods like PCA, LSA, and Latent Dirichlet Allocation (LDA) [6].

VIII. MACHINE LEARNING ALGORITHMS

This section offers an overview of various approaches used in medical document classification, covering four primary categories of machine learning algorithms: supervised, semi-supervised, unsupervised, and reinforcement learning. Within each category, diverse models are utilized based on the nature of the data and the learning objectives. Text classification methods incorporate decision trees (DT), rules-based approaches, Support Vector Machines (SVM), k-nearest neighbors (K-NN), neural networks, and Bayesian models. These methods are sometimes optimized using meta-boosting algorithms like AdaBoost or BoosTexter. Notable machine learning algorithms in this context include Linear Regression (LR) for both regression and classification problems [71], SVM specifically for classification problems [72], DT [73], and Random Forest (RF) [74]. These algorithms represent a diverse set of tools that cater to different characteristics of medical data and learning goals, providing a comprehensive toolkit for document classification in the medical domain.

A. Support Vector Machines (SVM)

A groundbreaking machine learning approach specifically designed to tackle classification problems. SVM operates on two fundamental principles: firstly, the maximization of the margin, allowing for errors in the training sets by determining the distance between the decision boundary and the nearest observations, commonly referred to as support vectors. Secondly, SVM facilitates an expansion into a new space, potentially of infinite dimensions, where linear separation becomes feasible. The primary objective of SVM's algorithm is the selection of planes defining decision boundaries between different classes of samples. Researchers have conducted extensive testing and comparisons, as noted in studies such as [44], [46], [75], and [76]. The authors in [45] have pitted SVM against other algorithms like Multilayer Perceptron Neural Networks (MLPNN), Random Forest, and Convolutional Neural Networks (CNN) using i2b2 2006 datasets. According to the collective findings of these authors, SVM consistently emerges as one of the models delivering superior performance and achieving commendable results compared to its algorithmic counterparts.

B. K-Nearest Neighbors (K-NN)

The K-NN algorithm operates by memorizing all samples in the training set and subsequently comparing them to a test sample, often referred to as memory-based learning or instance-based learning. While constructing the K-NN model is computationally inexpensive, as it involves storing the training data, classifying unknown samples can be relatively costly this is because it necessitates computing the K-nearest neighbors of the testing sample to assign a label, which involves calculating distances between the new sample and all objects in the training set. This computational demand can become significant, especially with large training sets, highlighting the need for careful consideration when determining the number of neighbors and selecting distance methods.

In the context of heart disease prediction, the authors in [77] employed various classification algorithms, including Naïve Bayes, Decision Trees, Support Vector Machines, and K-NN. Their experimental findings revealed that combining the K-means algorithm and decision tree improves the accuracy and underscores the potential efficacy of hybrid approaches in enhancing predictive models for heart disease.

C. Naïve Bayes (NB)

It is the first algorithm designed for text classification. It works based on the Bayes probability theorem and is used to solve problems associated with text and web classification [78]. For training dataset S = {S_1, S_2, …, S_m} (m samples) where every sample Si, is represented as an digital vector {x_1, x_2, …, x_n}, and k classes {C_1, C_2, …, C_k}, every sample belongs to one of these classes. Given that each sample belongs to one of these classes, for a data sample X (with an unknown class), it is possible to predict the class of X by using the conditional probability $P(C_i|X)$ where $i\in[1..k]$. This is the fundamental concept behind a naïve Bayesian classifier, where probabilities are calculated using Bayes' theorem, presented in (6).

$$P\left(C_i|X\right)=\frac{P\left(X|C_i\right)\times P\left(C_i\right)}{P\left(X\right)} \tag{6}$$

P($X|C_i$) presented in (6) is very complex to calculate. Based on the (naïve) assumption of total independence of the variables, proposed in a 1997 article by [79] the formula is as follows:

$$P\left(X|C_i\right)=\prod_{t=1}^{n}P\left(x_t|C_i\right) \tag{7}$$

Where, x_t in (7) values for attributes in the sample X. Although NB model is used sparingly in the medical field [80], however, similar to SVM, the NB model achieved significant results.

D. Decision Trees (DT)

Automatic Interaction Detection (AID) methods were abandoned by statisticians. Nevertheless, they were revived by the work of [73]. DT model allows the prediction of quantitative (regression trees) or qualitative (decision, classification, and segmentation trees). The general purpose of a decision tree is to explain a value from a series of discrete or continuous variables. Among the methodologies designed for the analysis of continuous variables, the chi-squared Automatic Interaction Detector (CHAID) algorithm [81] employs the evaluation of the χ^2 statistic to assess deviations from independence, alongside Tschuprow's measure [82]. The Classification and Regression Tree (CART) [78] methodology and the C4.5 [83] algorithm, which represents an advancement over the Iterative Dichotomiser 3 (ID3) algorithms, fundamentally rely on the principles of the Gini index and entropy, respectively. Regarding the adjustment of the tree size, there is post-pruning to CART and C4.5: that makes the tree pruned with all the segmentation; then, the model uses a criterion for comparing trees of different sizes. But CHAID proceeds by pre-pruning and setting a stopping rule to stop the construction. According to Shannon, for two possible outcomes with probabilities, p and $(1-p)$, the entropy H, is defined as follows [84]:

$$H\left(X\right)=-\sum_{i=1}^{n}p\left(x_i\right)log_2 p\left(x_i\right) \tag{8}$$

For X discrete random variable, the entropy $H(X)$ of X, that can assume values from a set of features$\{x_1, x_2, \ldots, x_n\}$ is presented in (8). The idea of generating a decision tree is not only to divide the original heterogeneous set into more homogeneous subsets but also to keep their size as small as possible in terms of the number of nodes. In addition, the Gini index function indicates the purity of the leaf nodes (a mixture of training data associated with each node) [85].

$$G=\sum_{k=1}^{n}p_k\times\left(1-p_k\right) \tag{9}$$

Equation (9) indicates that G is the Gini index for all classes, and p_k is the percentage of training instances. All classes of the same type (perfect class purity) have G=0.

Researchers referenced in [86] explored a range of artificial intelligence techniques, including DT, NB, K-NN, Logistic Regression, SVM, and neural networks like GRU and LSTM. These methods allowed them to evaluate model performance using various classification metrics. The results showed that the Symbiotic Gated Recurrent Unit (SGRU) outperformed the other models, achieving an F1-score of 0.69, compared to the DT's lower score of 0.52. To automatically code accident descriptions, specifically the US OSHA accident database, the investigation carried out by [87] assessed six distinct machine learning algorithms, including decision tree, NB, RF, linear regression, k-NN, and SVM. Those studies highlight that the SVM algorithm produced the best performance.

E. *Random Forest (RF)*

RF is used for regression or classification; it consists of constructing an ensemble of DT, generally trained via the bagging method or pasting. Moreover, it consists of making several independent DT to avoid ending up with equal trees; it gives each tree a piecemeal view of the problem, both on the input observations and the variables used. This double sampling is randomly pulled.

Thus, RF = tree bagging + feature sampling. Where:

- Tree bagging: random pull with a replacement on the lines (the observations).
- Feature sampling: random pull on the columns (the variables).

The prediction for new data is an average (in regression) or a vote (in classification). In [80], the authors used five classifiers: RF, SVM, decision tree, Naïve Bayes, and K-nearest neighbor for sentiment analysis of medical drug reviews. The SVM classifier outperforms the other classifiers. For example, to predict acute appendicitis in patients with undifferentiated abdominal pain in the emergency department, the authors in [88] utilized RF and binary logistic regression models, drawing on datasets from the United States National Hospital Ambulatory Medical Care Survey (NHAMCS). They concluded that the RF model demonstrated superior accuracy compared to the logistic regression model.

F. *Rule-based Approach*

Rule-based classifiers are widely used for text classification, especially medical text classification. The experiences of [89] demonstrate the performance of the Recurrent Neural Network (RNN) model and the combination of the rules-based engine generated by Pool-based Simulated Annealing (PSA). The authors in [45] developed a rules-based NLP algorithm and applied it to unlabeled clinical text to automatically generate weak labels. Authors in [90] proposed an NLP algorithm based on automated rules to extract cancer stage statements from narrative EHR data. Recently, the authors in [91] used BioBERT to classify entity to one of these labels (violence presence, perpetrator, victim, domestic, physical, and sexual) using the rules-based model.

G. *Artificial Neural Networks (ANN)*

An artificial neuron, known as a perceptron, is a mathematical model designed to mimic the functioning of the human brain. It receives input from neighboring neurons, each scaled by the corresponding connection weights. Then the inputs are summed, and an activation function transforms this sum into an output, which is passed to the neurons in the next layer for further processing. Among the neural networks, the most used are RNN, LSTM, CNN, GRU, etc. Where, neural network with multiple hidden layers is referred to as a multilayer perceptron (MLP), which is classified as deep learning [92].

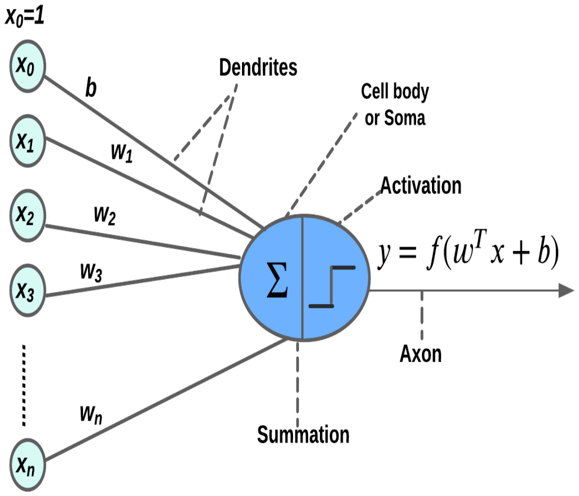

Fig. 6. Artificial Neuron (redrawing based on [93]).

Fig. 6 describes the structure of an artificial neuron. The input X, $X \in R^n$, passes through successive layers of neural units. The outputs of each layer are sent to the neurons of the following layers. The weight $w_{ij}^{(l)}$ corresponds to the weight connection between the i^{th} neuron in layer l and the j^{th} neuron in layer (l+1). Furthermore, each neuron unit i in a specific layer l is connected by a bias b^l, the predicted output y for the input vector x. When the label of the data is y, and y takes continuous values, the neural network learns the weights and biases by minimizing the prediction error $(y - \hat{y})^2$. The estimated W is calculated as follows: $\hat{W} = W \sum_i (y_l - \hat{y}_l)^2$

Fig. 5 illustrates the machine learning algorithms that are thoroughly examined in our paper. Recent studies in text classification within healthcare have explored diverse models and methods to automate processes and enhance accuracy. [41] Addressed the assignment of ICD-9 codes to clinical records using RNNs networks with LSTM. Their study involved human (MIMIC-III) and veterinary (Veterinary medical hospital at Colorado State University) datasets, employing RF and DT for model evaluation. Notably, LSTM outperformed DT and RF models, emphasizing the impact of using MetaMap (a tool for recognizing medical concepts in text) on accuracy. The proposed model FastTag was implemented for clinical text categorization, covering various medical domains. Another innovative model, CLUB-DRF [94], leveraged RF and clustering, demonstrating improved precision through the grouping of similar trees. [35] Introduced a bidirectional LSTM-CRF model with multitask attention for clinical NER. This model exhibited superior performance compared to traditional rules-based systems like MedLEE and MetaMap. This study [95] proposed the BART model (Bidirectional and Auto-Regressive Transformers) combining BERT and GPT in deep learning for text reconstruction. In the domain of biomedical event extraction, hybrid deep neural networks using CNN and RNN were employed by authors [130], emphasizing end-to-end learning. Efforts to automate the ICD coding process were made by [50], introducing an automatic model with an embedding layer, bidirectional LSTM layer, label attention layer, and output layer. For automated diagnostic coding [44], SVM models (flat and hierarchical) were utilized, with a fusion strategy to enhance performance. Additionally, a new model for extracting medical relationships, treating forests as latent variables, was proposed by [96].

Exploring ensemble techniques, [48] employed Majority Classifier Committee (MCC) aggregation, combining classifiers like KNN, Complement Naive Bayes (CNB), DT, RF, AdaBoost, XGBoost, and SVM through "Hard voting". Evolutionary neural network models for time-evolving text classification were introduced by [97], extending baseline models TextCNN [98], RCNN [99] and HAN [100] with temporal frameworks to predict cancer stage statements from HER. In the realm of probabilistic modeling, this study [101] presented a CRFs framework, overcoming the limitations of MEMMs and HMMs. [76] for detecting illicit drugs and recent studies have delved into Graph Neural Networks (GNN) for text classification, exemplified by the novel framework InducT-GCN [102], tested on medical-related datasets like Ohsumed. These studies collectively showcase the diverse approaches and models employed to advance text classification within the healthcare domain. [103] Introduces a NER method for Chinese Electronic Medical Records (EMRs) utilizing LSTM networks combined with CRF and Word2Vec embedding. This approach achieves high precision, recall, and F1-scores for body parts and treatment entities with scores exceeding 90%. However, it encounters challenges with symptoms, signs, and disease entities, showing lower performance with an accuracy of 66.9% and recall of 61.7% for diseases and diagnoses, and an F1-score of 80.0% for symptoms and signs. These limitations suggest that while the model is effective in certain areas, it struggles with the variability and complexity of other medical entities. In contrast [104] presents a novel NER approach using BERT-BiGRU-Att-CRF, which integrates the BERT model with Multi-Head Attention and Bi-directional Gated Recurrent Units (BiGRU). This model addresses challenges such as unannotated data and complex syntax in EMR texts, achieving an improved F1-score of 86.97% on the CCKS2019 dataset. The success of this model highlights how advanced techniques, such as attention mechanisms and pre-trained models, can enhance the accuracy of NER. Similarly, the study [105] focuses on improving healthcare documentation through text data mining using the KH Coder software. By analyzing nursing care records and identifying frequent terms like "toilet" and "wheelchair," this study emphasizes the potential for enhancing nursing care efficiency and documentation through semi-automated systems. This approach parallels the improvements in data handling seen in paper [104], demonstrating the broader applicability of advanced data processing techniques in healthcare settings.

Paper [106] introduces the PSI (Patient Similarity Identification) framework, which utilizes a medical knowledge graph to improve patient similarity assessments. This framework employs graph representation learning to generate embedding for medical entities and assesses patient similarity using a Siamese CNN with Spatial Pyramid Pooling (SPP). The framework's success in handling sparse data and representing complex relationships reflects the potential of ontologies to enhance data interpretation and integration, similar to the benefits observed in [104]. Building on the integration of data types, [107] presents the DiseaseNet model, which combines structured and unstructured EMR data using the BART model for summarization and BERT for data integration. The framework’s use of Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks improves diagnostic accuracy, demonstrating how comprehensive data integration can lead to better outcomes. This approach resonates with the use of ontologies in [106], highlighting the importance of integrating diverse data sources to enhance diagnostic capabilities. This study [108] utilizes a lexicon-based sentiment analysis approach with VADER to evaluate patient feedback on healthcare services. By classifying sentiments and verifying accuracy, this study underscores the importance of understanding patient opinions for improving healthcare quality. This focus on patient feedback complements the data-driven improvements demonstrating how sentiment analysis can enhance decision-making and service quality. Finally, [109] introduces a template-based Natural Language Generation (NLG) approach for summarizing hematological examination results. The system generates clear and coherent reports by mapping non-linguistic data to predefined templates, improving accessibility and usability. This method of structured report generation links to the broader theme of improving data interpretation and communication seen throughout the studies, including the uses of ontologies for managing and relating complex data.

IX. Ontologies and Machine learning

The use of machine learning technologies for digital health applications holds great promise, and the need for extensive, hand-labeled datasets for training poses a significant challenge to their broad deployment [110]. Automated disambiguation processing of clinical documents demands meticulous annotation of abbreviations, contributing to the imbalance of labeled training data and hindering machine learning deployment across clinical and research workflows. Knowledge management in digital health, crucial for human well-being, faces challenges related to semantic misinterpretation, which may lead to medical errors adversely affecting patient health. As highlighted by authors in [111]–[115], ontology emerges as a necessity to recognize entities, detect implicit knowledge, and extract semantics from clinical narratives, addressing these challenges effectively ontologies, surpassing the limitations of thesauri and taxonomies, offer a formal representation of conceptual meanings, facilitating reasoning and knowledge generation. Through explicit specifications of shared conceptualizations, ontologies enable advanced mapping between domains like anatomy, disease, phenotype, and laboratory investigation [116]. Ambiguities inherent in word meanings, such as the term "mole," highlight the challenges of context-dependent meanings, reinforcing the importance of ontology mappings [117], and [118]. Medical ontologies, including SNOMED-CT, HPO [119], Open Biomedical Ontologies (OBO), Unified Medical Language System (UMLS), and MeSH, contribute significantly to structured medical terminologies, improving vector space representations and supporting relation extraction models [117]. Recent studies showcase the synergies between machine learning and ontological approaches, notably in solving abbreviation disambiguation problems. Combining machine learning with ontological approaches, as demonstrated in the Finley database [120] and CASI dataset, signifies a promising advancement, yielding a 3% improvement in results [117]. The integration of ontologies in medical terminologies and knowledge representation, exemplified by the GALEN project, contributes to structured archiving and consolidation of simulation databases [111].

Ontologies continue to find applications in various domains (Fig. 7), such as dependency extraction using active learning [121], building pedagogical models in intelligent tutoring systems [122], and discovering gene/protein synonyms for applications like Twitter analysis [39]. Semantically rich biomedical ontologies, like the Cardiovascular Disease Ontology (CVDO), enhance word embedding and enable the development of NER architectures in healthcare domains [39] and [49]. Additionally, utilizing ontologies, relying on UMLS, facilitates semantic annotation of biomedical documents [77] or for semantic relations as seen in the study [114], focusing on the treatment of breast cancer. However, ontologies, which provide structured frameworks for representing knowledge, play a crucial role in these advancements. By creating well-defined models of medical concepts and relationships, ontologies facilitate more accurate data interpretation and integration. The PSI framework's [106] use of a medical knowledge graph and DiseaseNet's integration of structured and unstructured data rely on ontological principles to represent and relate medical entities. Ontologies enable the development of more sophisticated models that can handle diverse data types and complex relationships, leading to improved diagnostic accuracy, better patient care, and more efficient healthcare documentation and analysis. Furthermore, in [123], the authors generated Word2Vec embedding using a publicly available de-identified Electronic Health Records dataset. Then this embedding was augmented using three different algorithms, each employing a unique approach to integrating ontology information. The performance of these enhanced vectors was evaluated based on their correlation with human-annotated lists using Spearman's correlation coefficient and their effectiveness in NER tasks. Both quantitative metrics and empirical evaluations were used to evaluate the strengths and weaknesses of each approach. Where Word2Vec vectors enhanced with UMLS ontology information demonstrated the highest correlation with human-annotated evaluation lists, achieving a Spearman's correlation of 0.733 with the mini-Mayo clinical annotation. On the other hand, Bio + Clinical BERT outperformed the Word2Vec vectors in the NER task, achieving an F1-score of 0.87 on the i2b2 2010 dataset and 0.811 on the i2b2 2012 dataset, highlighting its superiority in this specific application. Clinically adapted Word2Vec vectors effectively capture lexical and clinical relationships such as synonymy, antonymy, and to a lesser extent hierarchical relationships like hyponymy and hypernymy. However, Bio + Clinical BERT proves to be more effective in NER tasks and in handling out-of-vocabulary words, demonstrating its robustness in clinical NLP applications. In addition, UmlsBERT proposed by [124] is a contextual embedding model aimed at improving biomedical natural language processing by incorporating structured domain knowledge during pre-training. Unlike models such as BioBERT and Bio_ClinicalBERT which focus exclusively on domain-specific corpora, UmlsBERT integrates expert knowledge from the UMLS. It achieves this by connecting words with the same underlying UMLS concept and utilizing UMLS semantic type information to create more meaningful input embedding. This methodology enables UmlsBERT to surpass existing models in tasks like NER and clinical natural language inference. Where, UmlsBERT achieved the highest F1-scores on the i2b2 2006, 2010, and 2012 tasks (93.6%, 88.6%, and 79.4%, respectively) and the best accuracy on the MedNLI task (83.0%). Although it did not surpass BERT on the i2b2 2014 task, this is likely due to differences in how Protected Health Information (PHI) is masked in the training data, affecting sentence structure. Despite this, UmlsBERT still outperformed other biomedical BERT models on i2b2 2014, confirming the benefit of integrating domain-specific biomedical knowledge into contextual embedding. An ablation test further demonstrated that UmlsBERT's performance improved when semantic type embedding was included, reinforcing the positive impact of this embedding on the model's effectiveness across different datasets. As demonstrated in recent studies the integration of ontologies and semantic information into medical text processing is essential for improving the effectiveness of NLP models in the biomedical field. Ontologies, such as those provided by the UMLS, offer a structured framework that captures intricate relationships between medical concepts, including synonymy, antonymy, and hierarchical associations. By embedding this domain-specific knowledge into NLP models, these systems can better understand and process complex medical language, leading to more accurate and meaningful interpretations. Models that incorporate UMLS concepts and semantic type information are able to create richer and more contextually relevant word embedding, enhancing their ability to handle tasks like NER and clinical natural language inference. This approach highlights the critical role of structured domain knowledge in advancing the capabilities of clinical NLP applications.

In conclusion, the integration of ontologies in healthcare and biomedical domains plays a pivotal role in addressing challenges related to data semantics, knowledge representation, and interoperability. These ontologies, coupled with machine learning approaches, contribute significantly to advancements in automated processing, decision support, and information extraction from clinical narratives. As demonstrated by the referenced studies, the synergy between machine learning and ontological methodologies continues to propel innovation in the digital health landscape.

X. Evaluation Methods and Model Estimation

In this section the most used evaluation methods and model estimation to evaluate models are presented, as shown in Fig. 8. In the resubstitution methodology, both the train and the test dataset are the same. This approach is infrequently employed in practical data mining applications in the real world. Authors in [94] used the Holdout method, where they used half or two-thirds of the datasets for training the model and the remaining data for testing. The training and test sets are independent. It appears helpful to do a different partitioning and then repeat the process to improve model estimation with other randomly selected training and testing sets. Authors in [125] used the Leave-one-out method, which consists of using (n-1) samples for training and testing the remaining sample and repeating n times with different training sets (n-1). The authors in [91], [94]–[97], [102], [111]–[115], [121], [122], and [125]–[130], employed the rotation method (n-fold cross-validation), which represents an amalgamation of the holdout and leave-one-out methodologies. It partitions the dataset into P mutually exclusive subsets, which is widely regarded as the most prevalent approach in empirical applications. The Bootstrap method is defined by [4] as creating new samples from the initial dataset to generate several fake data. Empirical findings indicated that bootstrap estimations have the potential to surpass the performance of cross-validation estimations. This method is beneficial in cases of small datasets such as this study [51]. Evaluating the performance of the models, accuracy and recall measures used by the majority of authors as: [48], [51], [94], [121], [130], and [131]–[136].

$$Accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{10}$$

Formula (10) delineates the metric of accuracy, defined as the proportion of accurate forecasts. Where, TP=True Positive; FP=False Positive; TN=True Negative; FN=False Negative.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{11}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{12}$$

$$F1-score = 2\times\frac{Accuracy \times Recall}{Accuracy + Recall} \tag{13}$$

Equation (11) represents the positive prediction accuracy. Equation (12) presents the proportion of instances that are accurately identified by the classifier. Furthermore, (13) articulates the weighted parameters of precision and recall. Supplementary evaluation metrics utilized by [94], including Mean Absolute Error (MAE) delineated in (14), Mean Squared Error (MSE) articulated in (15), and Root Mean Squared Error (RMSE) referenced in (16), are crucial for the evaluation of model performance.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{14}$$

Where, e_i = actual output- predicted value = $\left(y_i - \hat{y}_i\right)$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}e_i{}^2 \tag{15}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n}} \tag{16}$$

The authors in this study [122] consider the Laplace Accuracy (17) metric as an important measure to evaluate the performance of algorithms for the exploration of predictive association rules.

$$\text{Laplace Accuracy} = \frac{m_c+1}{m_{tot}+k} \tag{17}$$

Knowing, k represent the number of classes, m_{tot} is the total number of examples satisfying the body of the rule, where m_c examples belong to the class c. Various metrics such as macro and micro-averaged F1-score and AUC are used by authors in [50]. The ROC curve plots the false positive rate as a function of the true positive rate. [49], [51], [135]–[137], and [138]–[140].

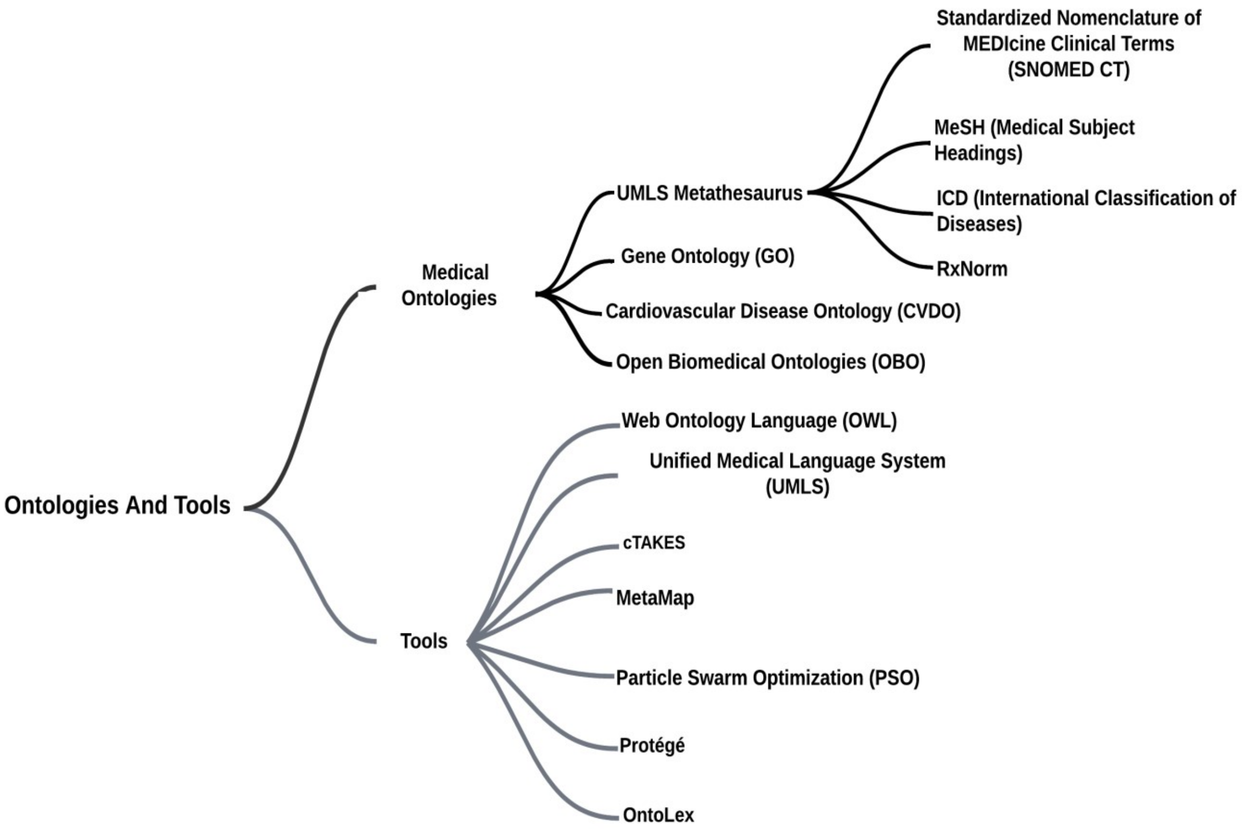
Ontologies And Tools
Medical Ontologies
UMLS Metathesaurus
Standardized Nomenclature of MEDIcine Clinical Terms (SNOMED CT)
MeSH (Medical Subject Headings)
ICD (International Classification of Diseases)
RxNorm
Gene Ontology (GO)
Cardiovascular Disease Ontology (CVDO)
Open Biomedical Ontologies (OBO)
Tools
Web Ontology Language (OWL)
Unified Medical Language System (UMLS)
cTAKES
MetaMap
Particle Swarm Optimization (PSO)
Protégé
OntoLex

Fig. 7. The main Medical Ontologies and Tools, which significantly improve the organization, accessibility, and analysis of medical information, and contributing to better patient care and more efficient healthcare delivery.

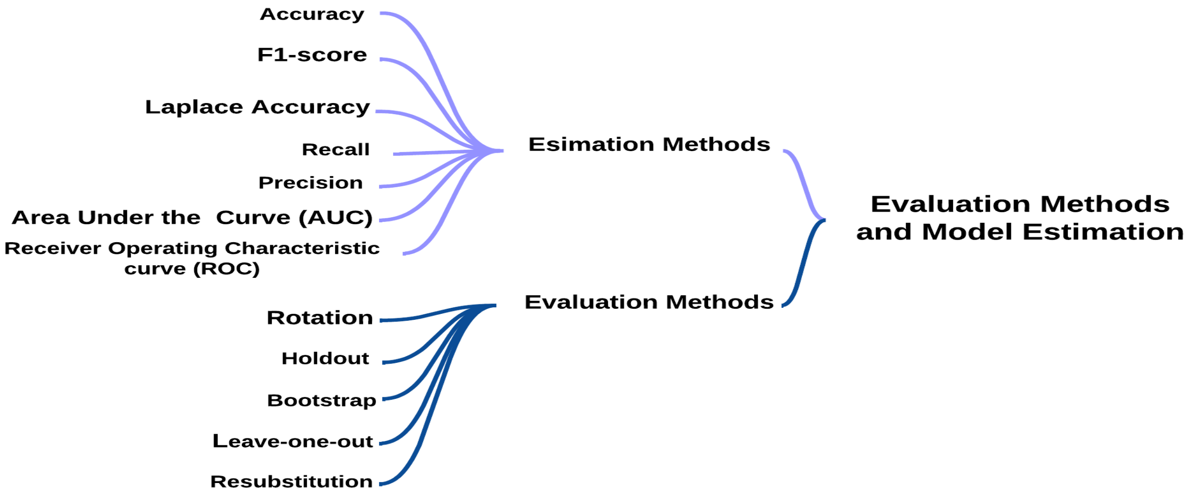

Fig. 8. The main Evaluation Methods and Model Estimation Techniques, for validating the effectiveness of machine learning models in medical narrative analysis, ensuring that they provide accurate, reliable, and clinically useful insights.

In summary, these metrics collectively address various aspects of model effectiveness, from classification accuracy to error magnitude, offering a holistic evaluation model that ensures both predictive reliability and balanced performance across different dimensions of the task.

XI. Discussion and Analysis

TABLE VI provides an overview of studies that have employed a variety of text representation methods and tools for medical text processing and classification tasks. Traditional methods like TF-IDF, n-grams, and Word2Vec are commonly used across many studies, demonstrating their foundational role in text analytics. More recent studies have used advanced techniques such as BERT, ELMo, and FastText, which provide a deeper contextual understanding of the text, especially in the domain of medical narratives. Several studies, such as those using Word2Vec (e.g., [36], [44], [50]), highlight the importance of vector-based models that capture semantic relationships between words. Techniques like CBOW and Skip-gram are specifically noted for their performance in medical tasks, including gene and protein synonym recognition. TF-IDF remains widely used in text classification tasks, as seen in studies like [45], [46], and [102], but these methods are increasingly being supplemented or replaced by deep learning models like BERT and BioBERT, which are capable of learning from complex linguistic patterns. Python emerges as the dominant programming language across studies, reflecting its versatility and the availability of libraries like scikit-learn, Gensim, and PyTorch. TensorFlow is also commonly used for deep learning tasks, and specific tools like MetaMap are employed for domain-specific tasks in the medical field, such as semantic and negation detection. FastText and GloVe are used in several studies for their efficiency in producing word embeddings, particularly when speed and scalability are crucial. As the field progresses, newer studies lean toward advanced contextualized embeddings like BERT and ELMo, which allow for a more nuanced understanding of the medical text. BERT models, including BioBERT and UmlsBERT, are particularly prominent in more recent research due to their effectiveness in capturing domain-specific terminology and complex relationships in clinical narratives. While the table demonstrates a diverse range of methodologies and tools used in medical text analysis, the evolution from traditional approaches (such as TF-IDF and n-grams) to deep learning-based models (such as BERT and ELMo) marks a significant shift in the field. However, there remains an opportunity to develop more unified frameworks that combine these traditional and modern approaches. For instance, hybrid models that integrate TF-IDF or n-grams with deep contextual embeddings like BERT could offer a balance between computational efficiency and nuanced text understanding.

TABLE VII offers a comprehensive examination of various studies dedicated to medical text processing, detailing their objectives, methodologies, utilized ontologies, and designated models. This compilation spans multiple years, reflecting the continuous evolution of methodologies and technologies within the field. Noteworthy studies, such as those by [36] and [39], focus on NER extraction, employing sophisticated approaches like BiLSTM-CRF and LSTM-CRF, frequently in conjunction with the UMLS ontology. The pervasive use of deep learning architectures marks a transformative shift towards more intricate interpretations of medical narratives.

For instance, research conducted by [37] illustrates the effectiveness of non-parametric Bayesian methods and HMM in NER, showcasing the methodological flexibility available for extracting critical medical information. In the classification domain, studies such as [44] and [76] utilize SVM alongside advanced deep learning architectures, including CNNs and BiLSTM-CRF. The integration of both hierarchical and flat SVM techniques in [44] highlights the essential requirement for precise coding in alignment with the ICD. The diversity of classifiers explored -ranging from RF to NB and DT in [80] illustrates the multifaceted strategies that researchers are deploying to address the complexities of classification tasks effectively. Feature extraction also emerges as a critical component, with studies like [97] leveraging neural network architectures to enhance classification performance. The adoption of GNN in [102] signifies a novel trajectory in modeling relationships among data points, facilitating deeper insights into medical texts. Furthermore, the rise of models that incorporate attention mechanisms, as evidenced in [104], indicates a growing trend towards employing advanced architectures to refine NER extraction processes. This aligns with the broader movement toward sophisticated models like BERT, as highlighted in studies such as [124], which excel in capturing the intricate semantic relationships inherent in clinical narratives. MetaMap integrates semantic and negation detection in clinical narrative text, the model proposed by [41] handles 17 categories of diseases. On the other hand, the approach suggested by [45] for clinical text classification aims to reduce the need for human-labeled training data and extensive feature engineering by leveraging weak supervision and deep representation. However, their model still required domain experts to formulate specific rules. The research additionally revealed that the nature of classification multi-class versus binary-has a significant impact on the efficacy of CNN models. For instance, precision results include 0.93 for Mayo Clinic Smoking Status classification, 0.97 for Proximal Femur (Hip) Fracture classification, and 0.76 for the i2b2 2006 smoking status classification, indicating that dataset choice influences model precision.

In the model proposed by [43], two approaches were combined: one using tv-embeddings, LSTM, and max pooling, and another using tv-embeddings, CNN, and max pooling. This hybrid model performed better than either the model (CNN or LSTM) alone. In contrast, the study by [94] did not compare their model with others to validate its performance but focused on examining how the pruned number of DT impacted outcomes. In another study, [90] demonstrated that cancer stages could be effectively extracted from narrative text in most electronic health record samples, achieving high accuracy. Additionally, the research by [40] revealed that text vectorization using ELMo (Clinical) outperformed ELMo (General), Word2Vec, and GloVe, though it was limited to extracting medical tests, problems, and treatment concepts. The ensemble model exhibited values of F1-score, precision, and recall metrics of 88.78, 89.11, and 88.46, respectively. In comparison to the previously established optimal solution, "ELMo (Clinical) and BiLSTM–CRF," as delineated by [141], the ensemble model demonstrated marginal enhancements in F1-score (+0.18%) and recall (+0.59%), although it experienced a slight decrement in precision (-0.23%). We emphasize that despite the accuracy parameter allowing the evaluation of any model, most authors use only a few vital parameters, such as model time learning and time prediction. We concluded that TextCNN outperforms SVM in all metrics (Precision, Recall, and F1-score), but the SVM-based method takes less than 1 hour, and the TextCNN process takes 11 hours for training. Both methods take less than 0.05 seconds for prediction, which is the most important. They used 10-fold cross-validation procedures to evaluate the accuracy of both the SVM and CNN-based methods.

This study concludes that the accuracy of the models depends on several parameters. The most crucial factor is the choice of a dataset in terms of content and size. A dataset directly affects the model's accuracy, where models constructed (trained and tested) from a large corpus are better than models built from a small corpus. Thus, researchers should combine several datasets. In addition, a training text indexing model, and using several methods allows better classification. For the Assignment of medical imaging procedure protocols, in this article [46] authors made a comparative study between the models by changing the classification methods without changing the text vectorization for a better and more accurate comparison. The best system proposed is SVM, where the results obtained using the Computed Tomography (CT) dataset got 92% accuracy and 76% F1-score, for 87% accuracy and 76% F1-score for the Magnetic Resonance Imaging (MRI) dataset. In the majority of studies, the incorporation of ontologies improved the results obtained by the models proposed. The methodological diversity presented in the table underscores the rapid advancements in medical text processing. While traditional techniques, such as rule-based systems and SVMs, continue to hold significance, there is a clear shift towards more advanced methodologies, including deep learning and attention-based models. This evolution reflects the increasing complexity of medical data and underscores the pressing need for more accurate, context-aware processing methods. As the field of medical text analytics continues to evolve, several key propositions emerge that could significantly enhance research and application in this domain, the propositions are as follows:

- There is a valuable opportunity to develop hybrid models that integrate traditional machine learning techniques, such as SVM, with modern deep learning architectures. As a result of this combination, classification and extraction tasks may perform better.
- Future research should focus on creating standardized ontologies and frameworks for consistent application across studies. This approach would improve the comparability of results and encourage collaboration among researchers, advancing the field of medical text analytics.
- Investigating transfer learning techniques, particularly with models like BERT, could open new avenues for improving NER tasks. Fine-tuning pre-trained models on specific medical datasets could enhance performance while minimizing the need for extensive training data.
- Social Media and Public Health Data: Using social media platforms (e.g., Twitter, Google+) indicates an increasing interest in public health monitoring through user-generated content. These data sources provide insights into health trends, patient opinions, and behaviors in real-time, offering complementary information to traditional clinical datasets.

TABLE VI
TEXT VECTORIZATION: OVERVIEW OF THE PREDOMINANT TECHNIQUES EMPLOYED IN THE PROCESS OF TEXT DIGITIZATION

Related Studies	Text Vectorisation Methods	Tools
[36]	FastText, Word2Vec, Skip-Gram	Gensim Library, Python
[37]	Not used	Python 2.7, Numpy, Scipy
[39]	Word2Vec, GloVe, pyysalo, Chiu, ChenPM, Aueb	/
[40]	ELMo	Pytorch library, Python 3.7
[41]	GloVe	Python(2.7),Tensor-Flow(1.9),scikit-learn library(0.19.2) https://github.com/rivas-lab/FasTag, MetaMap
[44]	Word2Vec (CBOW), BOW	/
[45]	TF-IDF, Word2Vec, topic modeling [142]	http://creativecommons.org/publicdomain/zero/1.0/
[46]	TF-IDF	/
[50]	CBOW, Word2Vec	Python, PyTorch
[64]	Document Term Matrix	scikit-learn
[75]	n-grams	Weka toolkit
[74]	TF-IDF, Word2Vec	Python, SVM Based Method TextCNN
[80]	FastText based on TF-IDF	Python, Keras.
[89]	Word2Vec	/
[88]	BioBERT	/
[90]	n-grams	R software, Microsoft Access 2013, RODBC [143] iGraph
[97]	Word2Vec	github.com/RingBDStack/Time-evolving-Classification
[102]	TF-IDF	https://github.com/usydnlp/InductTGCN
[131]	TF-IDF, unigrams, bigrams, trigrams	NLTK library
[144]	TF-IDF	/
[145]	skip-gram	Python, Tensorflow
[146]	TF-IDF	MetaMap, PSO
[103]	Word2Vec	Jieba word segmentation
[104]	BERT	Python 3.7, Pytorch 1.7.0
[106]	HKGE (Heterogeneous Knowledge Graph Embedding **)**	TensorFlow
[107]	BERT	BART
[124]	UmlsBERT	/

TABLE VII
STUDIES DESCRIPTION, THIS TABLE PROVIDES A COMPREHENSIVE OVERVIEW OF VARIOUS STUDIES BY DETAILING KEY ASPECTS OF THEIR RESEARCH. IT INCLUDES THE YEAR OF PUBLICATION, THE PRIMARY OBJECTIVES EACH STUDY AIMED TO ACHIEVE, AND THE METHODOLOGIES UTILIZED TO ADDRESS THEIR RESEARCH QUESTIONS. THE TABLE ALSO HIGHLIGHTS THE ONTOLOGIES EMPLOYED FOR SEMANTIC ANALYSIS INDICATING HOW DOMAIN-SPECIFIC KNOWLEDGE WAS INTEGRATED INTO THE STUDIES. ADDITIONALLY, IT LISTS THE MODELS OR ALGORITHMS USED, RANGING FROM ADVANCED NEURAL NETWORK ARCHITECTURES LIKE LSTM, CNN, AND RNN TO TRADITIONAL APPROACHES SUCH AS NAÏVE BAYES, DECISION TREES, AND KNN. THIS TABLE ILLUSTRATES THE DIVERSE APPROACHES APPLIED IN THE ANALYSIS OF MEDICAL REPORTS, UNDERSCORING THE ONGOING RELEVANCE OF BOTH MODERN AND CLASSICAL TECHNIQUES IN HANDLING UNSTRUCTURED DATA

Studies	Year	Objective	Methods	Ontology	Model Designation
[36]	2019	Classification	BiLSTM-CRF, BiGRU-CRF, BiLSTM-S, BiGRU-S	/	/
[37]	2015	NER extraction	Non-parametric Bayesian HMM and Dirichlet Process (DP),HMM-DP with CRF	/	HMM-DP HMM-DP+ CRF
[39]	2019	NER extraction	LSTM–CRF, BiLSTM-CRF	UMLS	/
[40]	2019	NER extraction	BiLSTM–CRF	UMLS	/
[41]	2020	NER extraction	LSTM, RF, DT	UMLS	FastTag
[44]	2017	Automatic coding, Classification	Flat SVM, hierarchical SVM	ICD9CM [147] ICD9	
[45]	2019	Classification	Rules-based, SVM, RF, MLPNN and CNN	/	/
[46]	2021	Classification	SVM, RF, CNN, BILSTM, BETO [148].	/	/
[50]	2018	Classification (multi-label)	Bidirectional LSTM	/	LAAT JointLAAT
[64]	2018	Clusrting	t-SNE, PCA	/	
[75]	2015	Feature extraction Classification	SVM	SNOMED CT	/
[76]	2022	Classification,	SVM, CNN BiLSTM-CRF	/	TextCNN
[80]	2020	Classification	SVM, Decision Tree, RF, Naive Bayes, and K-NN	/	/
[89]	2020	NER extraction	Rules-based, Pool-based (Simulated Annealing)		
[88]	2022	Classification Feature selection	Rules-based	/	/
[90]	2016	Classification	Rules-based	/	/
[94]	2020	Regression	RF and K-means	/	CLUB-DRF
[97]	2018	Feature extraction Classification	Neural networks (TextCNN, RCNN and HAN)	/	/
[102]	2022	Classification	GNN	/	InducT-GCN
[131]	2020	Classification	RF, SVMlinear, SVMrbf	/	/
[144]	2020	Classification	SVM, NB, RNN, and ANN	/	/
[146]	2019	Feature Selection Classification	KNN, DT, SVM, NB, LR	UMLS	/
[103]	2023	NER extraction	LSTM+CRF	/	/
[104]	2024	NER extraction	BiGRU+ Multi-Head Attention +CRF		BERT-BiGRU-Att-CRF
[106]	2021	Patient Similarity Identification	Siamese CNN [149], with Spatial Pyramid Pooling	ICD-9	PSI
[107]	2022	Classification	BiLSTM+CNN	/	DiseaseNet
[124]	2021	Text vectorization	BERT	UMLS	UmlsBERT
[150]	2018	NER extraction, Classification	CRF, BiLSTM+CRF, BiGRU, BiGRU+CRF	/	/
[151]	2019	NER extraction, Classification	SVM, Dimensionality Reduction, SVM+ SESARF	UMLS	SESARF
[152]	2015	NER extraction, Classification	SVM, MCS [153], SVM+MCS, (WUP [154])		/

TABLE VIII
MAIN DATASETS USED, THIS TABLE SERVES AS A COMPREHENSIVE RESOURCE TO UNDERSTAND THE INTERCONNECTION BETWEEN DATASETS, STUDIES, AND THE TEXTUAL OR DATA COMPONENTS CRUCIAL TO THE RESEARCH PROCESS

Studies	Corpus	Designation
[36]	BioScop	Consists of 3 textual sources different: reports of radiological abstracts and Bioinformatics articles
[36]	ESSAI	French biomedical texts of clinical trial protocols
[36]	SemClinBr	Consisting of clinical texts that have been provided by three Brazilian hospitals linked to several specialties
[36]	CAS	French biomedical texts, 200 clinical cases.(2018)
[37],[124],[150]	i2b2/2014	UTHealth de-identification challenge
[39]	Twitter	This data was collected by Twitter between 12 July 2018 and 12 July 2019 using the search terms "healthcare"
[40], [146], [124]	i2b2/VA 2010	Datasets utilized to extract medical concepts
[39]	CSU	Veterinary medical hospital at Colorado State University
[39]	MIMIC-III + CSU	
[44],[50],[106] [41],[107]	MIMIC-III	Medical Information Mart for Intensive Care III
[45], [124]	i2b2 2006	Fracture classification in the Mayo Clinic: Proximal Femur (Hip)
[46]	MRI	Magnetic Resonance Imaging
[46]	CT	Computed Tomography
[50]	MIMIC-III-50 MIMIC-II-full	MIMIC-III Medical Information Mart for Intensive Care II subset of the 50 most frequent codes
[64]	Russian EHR	EHR of cardiovascular patients at the Alamazov Center
[69]	Discharge summaries	Columbia University Medical Center (CUMC)
[75]	Death certificate	Data from Cancer Institute New South Wales
[76]	Google+,Flickr Tumblr	Dataset collected from Google+, Flickr and Tumblr
[89]	China’s dataset Online medical	Consultation platform
[88]	CRIS	Clinical Record Interactive Search mental HR
[90]	VCR	(Vanderbilt Cancer Registry) Vanderbilt University
[97]	NYTimes	Articles published in the New York Times Manualy labelled newswire collection of Reuters
[97]	RCV1	Produced from RCV1, with 12 subtrees, for each
[97]	RCV1-org	category contain half of the concepts
[97]	RCV1-noise	Present RCV1-org by adding pseudo-random noise
[97]	RCV1-drift	From RCV1-org with half of concepts as new concepts
[102]	Ohsumed	Produced by the MEDLINE database,23 diseases abstracts
[80], [144]	Drugs.com	https://archive.ics.uci.edu/datasets
[131]	SU-ADE	Swedish Health Data Research Bank (Stockholm University)
[145]	NMLEC	Medical Licensing Examination in China
[103]	Chinese EMRs	Entity recognition task
[104]	CCKS2019	Chinese Electronic Medical Records
[106]	DrugBank	https://go.drugbank.com/
[124]	MedNLi	Medical history dataset annotated by doctors performing a natural language inference task
[150]	CRTT-MED	French medical corpus, https://quaerofrenchmed.limsi.fr/
[150]	QUAERO	http://perso.univ-lyon2.fr/~maniezf/Corpus/Corpus_medical_FR_CRTT.htm
[151]	TREC CDS	Text Retrieval Conference Clinical Decision Support
[152]	SemEval-2014 task 7	http://alt.qcri.org/semeval2014/task7/

- Global Representation: Datasets from non-English-speaking countries, such as SemClinBr (Brazil) [36], Chinese EMRs in [103], the Swedish Health Data Research Bank in [131], the French corpora CRTT-MED and QUAERO [149], highlight the global efforts in medical NLP. They underscore the need for multilingual and culturally diverse data, which are crucial for creating models applicable in international healthcare contexts.

In order to improve the performance of the models, we suggest the following strategies:

- Integration of Diverse Datasets: Using datasets from varied sources (e.g., clinical records, medical publications, and social media) the models to generalize across contexts. For example, hybrid models that incorporate the structured nature of clinical texts with the flexibility of social media data could improve performance in both medical concept extraction and public health surveillance.
- Standardization across Corpora: There is a need for standardized frameworks to make comparisons between studies more consistent. Creating benchmark datasets that span multiple languages and medical systems would enhance the reproducibility of research and facilitate collaboration. By aligning datasets from different sources under a unified ontology, such as UMLS, researchers could streamline the integration of knowledge and improve model interpretability.
- Leveraging Transfer Learning for Niche Domains: Transfer learning techniques could be further explored to apply models trained on large datasets (e.g., MIMIC-III or i2b2) to niche domains like veterinary medicine (e.g., CSU in [41]) or mental health records (e.g., CRIS in [88]). Fine-tuning pre-trained models on these specialized datasets can improve performance while reducing the need for extensive, domain-specific training data.
- Enhancing Multilingual Capabilities: Given the increasing use of datasets in languages other than English (e.g., SemClinBr, Chinese EMRs), there is an opportunity to develop more multilingual NLP models. These models should not only translate medical concepts accurately but also capture the cultural nuances and medical terminologies unique to each language.

Finally, the range of corpora highlighted in this study offers underscores the importance of dataset diversity in advancing medical NLP. By integrating diverse data sources, standardizing frameworks, and leveraging transfer learning, the field can continue to evolve and address complex healthcare challenges more effectively.

In conclusion of this section, in the field of biomedical natural language processing, the performance of predictive models is heavily influenced by the intricate interplay between vectorization methods, machine learning models, datasets, and the tools employed particularly ontologies. Vectorization methods, such as word embeddings, serve as the foundational layer by transforming textual data into numerical representations that models can process. The choice of vectorization method can significantly impact the model's ability to capture semantic nuances, especially in complex and specialized domains like healthcare. Machine learning models, whether they are traditional algorithms like NB and DT or advanced neural networks like CNNs and LSTMs, rely on these vectors to learn patterns and make predictions. The effectiveness of these models is further augmented by the quality and diversity of datasets used for training and evaluation. Datasets that are representative of the domain and cover a wide range of scenarios ensure that the models generalize well to real-world applications. Model semantics are enhanced by ontologies, in fact those offered by the UMLS. Integrating structured domain knowledge (ontologies) helps refine vectorization processes, ensuring that the generated embeddings are not only syntactically, but also semantically meaningful.

XII. Conclusion and future works

This survey provides an in-depth review of studies on text mining techniques applied to healthcare, focusing on the extraction of knowledge from unstructured data in medical reports. We explore foundational concepts and key techniques, including various NER methods, widely-used classification algorithms, diverse preprocessing processes, and the role of ontologies in enhancing medical text analysis. We emphasize that text mining can significantly streamline patient diagnosis and treatment recommendations, potentially leading to more efficient healthcare delivery.

While evaluating and comparing machine learning models typically involves using a single dataset, our survey advocates for the use of multiple datasets to better validate and confirm model performance. We highlight the need for standardized evaluation criteria in text mining to ensure consistent and reliable results. The challenges of varying medical vocabulary comprising complex terms, abbreviations, acronyms, and errors present significant difficulties for clinical text classification.

Medical texts can be more efficiently analyzed with ontologies and machine learning. Additionally, we address the importance of preprocessing in optimizing model performance. Our study includes various applications of text mining in medical documents, such as drug classification, disease-based patient categorization, and document classification according to different medical conditions.

In future works, we propose:

1) Enhanced Ontology Integration: Future research should aim to further integrate ontologies with advanced machine learning models. This involves creating more detailed and domain-specific ontologies that capture the subtleties of medical language more effectively, thereby enhancing the accuracy of entity recognition and relationship extraction.
2) Multi-Dataset Validation: To verify the effectiveness of text mining techniques, future studies ought to employ multiple datasets for validation. This approach will ensure that the models are robust and generalizable across various medical contexts and data sources.
3) Addressing Vocabulary Challenges: Future work had better focus on innovative strategies to overcome challenges related to medical vocabulary. This includes developing techniques to manage domain-specific terminology, abbreviations, acronyms, and errors in

medical texts, which will improve the performance of text-mining models.

4) Advanced Preprocessing Techniques: Further research is needed into advanced preprocessing methods to better handle and normalize medical text data. This should include refining techniques for error correction, addressing dialectal variations, and standardizing medical terminology.
5) Application of Multimodal Data: Integrating text mining with multimodal data sources, such as imaging and genetic information, could offer a more holistic understanding of patient conditions and enhance diagnostic accuracy.
6) Real-World Implementation: Upcoming research should focus on implementing text mining techniques in practical healthcare settings. This includes developing tools and systems that leverage advanced text mining and ontology-based methods to support clinical decision-making and improve patient care.

By addressing these future directions, researchers can further enhance the capabilities of text-mining techniques and ontologies in healthcare, leading to more effective knowledge extraction and ultimately better healthcare outcomes.

REFERENCES

[1] M. A. Hearst, "Untangling text data mining," in *Proc. 37th Annual. Meeting of the Association for Computational Linguistics, College Park, MD*, June 1999, pp. 3-10.

[2] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: a review of recent research," *Yearbook of Medical Informatics*, vol. 17, no. 01, pp. 128-144, 2008.

[3] S. Tsumoto, S. Hirano, and Y. Tsumoto, "Clustering-based analysis in hospital information systems," in *2011 IEEE International Conference on Granular Computing*, IEEE, 2011, pp. 669-674.

[4] M. Kantardzic, *DATA MINING Concepts, Models, Methods, and Algorithms*. Wiley-IEEE Press, 2020.

[5] Y. H. Tseng, C. J. Lin, and Y. I. Lin, "Text mining techniques for patent analysis," *Information Processing & Management*, vol. 43, pp. 1216–1247, 2007.

[6] C. C. Aggarwal, *Machine Learning for Text*. Springer International Publishing, 2018.

[7] B. Tony, *Developpez votre intelligence avec le Mind Mapping*. Alisio, 2018.

[8] K. S. Kalyan and S. Sangeetha, "SECNLP: A survey of embeddings in clinical natural language processing," *Journal of Biomedical Informatics,* vol. 101, ID. 103323, 2020.

[9] G. Alfattni, N. Peek, and G. Nenadic, "Extraction of temporal relations from clinical free text: A systematic review of current approaches," *Journal of Biomedical Informatics*, vol. 108, ID. 103488, 2020.

[10] M. Hussain, F. A. Satti, J. Hussain, T. Ali, S. I. Ali, H. S. M. Bilal, and T. Chung, "A practical approach towards causality mining in clinical text using active transfer learning," *Journal of Biomedical Informatics*, vol. 123, pp. 103932, 2021.

[11] Z. Li, C. Li, Y. Long, and X. Wang, "A system for automatically extracting clinical events with temporal information," *BMC Medical Informatics and Decision Making,* vol. 20, no. 1, pp. 1-13, 2020.

[12] F. Dernoncourt, PubMed 200k RCT dataset. (Online). Available: https://github.com/Franck-Dernoncourt/pubmed-rct. Accessed June 10, 2021.

[13] J. L. Leevy and T. M. Khoshgoftaar, "A Short Survey of LSTM Models for De-identification of Medical Free Text," in *2020 IEEE 6th International Conference on Collaboration and Internet Computing (CIC)*, IEEE, pp. 117-124, 2020.

[14] I. J. B. Young, S. Luz, and N. Lone, "A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis," *International Journal of Medical Informatics*, vol. 132, ID. 103971, 2019.

[15] G. Mujtaba, L. Shuib, N. Idris, W. L. Hoo, R. G. Raj, K. Khowaja, K. Shaikh, and H. F. Nweke, "Clinical text classification research trends: systematic literature review and open issues," *Expert Systems with Applications*, vol. 116, pp. 494-520, 2019.

[16] L. Pereira, R. Rijo, C. Silva, and R. Martinho, "Text mining applied to electronic medical records: a literature review," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 6, no. 3, pp. 1-18, 2015.

[17] C. Yan, X. Fu, X. Liu, Y. Zhang, Y. Gao, J. Wu, and Q. Li, "A survey of automated International Classification of Diseases coding: development, challenges, and applications," *Intelligent Medicine*, vol. 2, no. 3, pp. 161-173, 2022.

[18] S. Taheri Moghadam, N. Hooman, and A. Sheikhtaheri, "Patient safety classifications, taxonomies and ontologies: A systematic review on development and evaluation methodologies," *Journal of Biomedical Informatics*, vol. 133, ID. 104150, 2022.

[19] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, Y. Zhao, S. Sohn, and H. Liu, "Clinical concept extraction: a methodology review," *Journal of Biomedical Informatics*, vol. 109, pp. 103526, 2020.

[20] P. Bose, S. Srinivasan, W. C. S. Iv, J. Palta, R. Kapoor, and P. Ghosh, "A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts," *Applied Sciences*, vol. 11, no. 18, p. 8319, 2021.

[21] M. Madkour, D. Benhaddou, and C. Tao, "Temporal Data Representation, Normalization, Extraction, and Reasoning: A Review from Clinical Domain," *Computer Methods and Programs in Biomedicine*, vol. 128, pp. 52–68, 2016.

[22] K. Liu, Y. Chen, J. Liu, X. Zuo, and J. Zhao, "Extracting Events and Their Relations from Texts: A Survey on Recent Research Progress and Challenges," *AI Open*, vol. 1, pp. 22–39, 2021.

[23] M. G. Kersloot, F. J. van Putten, A. Abu-Hanna, R. Cornet, and D. L. Arts, "Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies," *Journal of Biomedic,* vol. 11, pp. 1–21, 2020.

[24] M. AlShuweihi, S. A. Salloum, and K. Shaalan, "Biomedical corpora and natural language processing on clinical text in languages other than English: a systematic review," *Recent Advances in Intelligent Systems and Smart Applications*, pp. 491-509, 2021.

[25] T. M. Seinen, E. A. Fridgeirsson, S. Ioannou, D. Jeannetot, L. H. John, J. A. Kors, A. F. Markus, V. Pera, A. Rekkas, R. D. Williams, C. Yang, E. M. van Mulligen, and P. R. Rijnbeek, "Use of unstructured text in prognostic clinical prediction models: a systematic review," *Journal of the American Medical Informatics Association*, vol. 29, no. 7, pp. 1292-1302, 2022.

[26] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, "Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: A Systematic Review," *Journal of the American Medical Informatics Association*, vol. 23, no. 5,pp. 1007-1015, 2016.

[27] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review," *Journal of Healthcare Engineering*, 2018.

[28] S. S. Tandel, A. Jamadar, and S. Dudugu, "A Survey on Text Mining Techniques," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, IEEE, 2019, pp. 1022-1026.

[29] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, and J. Mostafa, "*A Systematic Review of Automatic Text Summarization for Biomedical Literature and EHRs*," Journal of the American Medical Informatics Association, vol. 28, no. 10, pp. 2287-2297, 2021.

[30] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A Survey of Word Embeddings for Clinical Text," *Journal of Biomedical Informatics*, vol. 100, ID. 100057, 2019.

[31] Y. B. Gumiel, L. E. Silva e Oliveira, V. Claveau, N. Grabar, E. C. Paraiso, C. Moro, and D. R. Carvalho, "Temporal relation extraction in clinical texts: a systematic review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1-36, 2021.

[32] Aggarwal, C.C., and Zhai, C.X., *Mining Text Data*, Springer New York, 2012.

[33] J. Friedlin and C. J. McDonald, "A Natural Language Processing System to Extract and Code Concepts Relating to Congestive Heart Failure from Chest Radiology Reports," in *AMIA Annual Symposium Proceedings, American Medical Informatics Association*, 2006, pp.269-273.

[34] J. Friedlin, S. Grannis, and J. M. Overhage, "Using Natural Language Processing to Improve Accuracy of Automated Notifiable Disease Reporting," in *AMIA Annual Symposium Proceedings, American Medical Informatics Association*, 2008, pp. 207.

[35] M. A. Al-Haddad, J. Friedlin, J. Kesterson, J. A. Waters, J. R. Aguilar-Saavedra, and C. M. Schmidt, “Natural Language Processing for the Development of a Clinical Registry: A Validation Study in Intraductal Papillary Mucinous Neoplasms,” *HPB*, vol. 12, no. 10, pp. 688-695, 2010.

[36] C. Dalloux, “Fouille de texte et extraction d’informations dans les données clinique ,“ Ph.D. dissertation, Université de rennes1, France, 2020.

[37] T. Chen, R. M. Cullen, and M. Godwin, “Hidden Markov Model Using Dirichlet Process for De-identification,” *Journal of Biomedical Informatics*, vol. 58, pp. S60-S66, 2015.

[38] A. McCallum, D. Freitag, and F. C. Pereira, “Maximum Entropy Markov Models for Information Extraction and Segmentation,” in *ICML*, pp. 591-598, 2000.

[39] E. Batbaatar and K. H. Ryu, “Ontology-Based Healthcare Named Entity Recognition from Twitter Messages Using a Recurrent Neural Network Approach,” *International Journal of Environmental Research and Public Health*, vol. 16, no. 19, pp. 3628, 2019.

[40] J. Yang, Y. Liu, M. Qian, C. Guan, and X. Yuan, “Information Extraction from Electronic Medical Records Using Multitask Recurrent Neural Network with Contextual Word Embedding,” *Applied Sciences*, vol. 9, no. 18, pp. 3658, 2019.

[41] G. R. Venkataraman, A. L. Pineda, W. IV, A. M. Zehnder, S. Ayyar, R. L. Page, C. D. Bustamante, and M. A. Rivas, “FasTag: Automatic text classification of unstructured medical narratives,” *PLOS ONE*, vol. 15, no. 6, pp. e0234647, 2020.

[42] T. Ganegedara, *Natural Language Processing with TensorFlow: Teach Language to Machines Using Python’s Deep Learning Library*. Packt Publishing Ltd., 2018.

[43] R. Johnson and T. Zhang, “Supervised and Semi-Supervised Text Categorization Using LSTM for Region Embeddings,” in *International Conference on Machine Learning, PMLR*, 2016, pp. 526-534.

[44] S. Berndorfer and A. Henriksson, “Automated Diagnosis Coding with Combined Text Representations,” *Studies in Health Technology and Informatics*, vol. 235, pp. 201-205, 2017.

[45] Y. Wang, S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, S. Amin, and H. Liu, “A clinical text classification paradigm using weak supervision and deep representation,” *BMC Medical Informatics and Decision Making*, vol. 19, pp. 1-13, 2019.

[46] P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. Ureña-López, and M. T. Martin-Valdivia, “Automatic Medical Protocol Classification Using Machine Learning Approaches,” *Computer Methods and Programs in Biomedicine*, vol. 200, pp. 105939, 2021.

[47] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990.

[48] Y. Mercadier, “Classification automatique de textes par réseaux de neurones profonds: Application au domaine de la santé,“ Ph.D. dissertation, Université Montpellier, Montpellier, France, 2020.

[49] M. Arguello Casteleiro, G. Demetriou, W. Read, M. J. Fernandez Prieto, N. Maroto, D. Maseda Fernandez, G. Nenadic, J. Klein, J. Keane, and R. Stevens, “Deep Learning Meets Ontologies: Experiments to Anchor the Cardiovascular Disease Ontology in the Biomedical Literature,” *Journal of Biomedical Semantics*, vol. 9, no. 1, pp. 1-24, 2018.

[50] T. Vu, D. Q. Nguyen, and A. Nguyen, “A label attention model for ICD coding from clinical text,” in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 3335-3341.

[51] C. Coulombe, “Techniques d’amplification des données textuelles pour l’apprentissage profond,” Doctoral dissertation, Télé-université, Université de Québec, Québec, Canada, 2020.

[52] S. Bartunov, D. Kondrashkin, A. Osokin, and D. Vetrov, “Breaking Sticks and Ambiguities with Adaptive Skip-gram,” in *Artificial Intelligence and Statistics, PMLR*, 2016, pp. 130-138.

[53] Q. Le and T. Mikolov, “Distributed Representations of Sentences and Documents,” in *International Conference on Machine Learning, PMLR*, 2014, pp. 1188-1196..

[54] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global Vectors for Word Representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532-1543.

[55] J. D. M. W. Kenton and L. K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of NAACL-HLT*, pp. 1-2, 2019.

[56] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. (EMNLP-IJCNLP), Association for Computational Linguistics*, 2019, pp. 3613.

[57] Y. Peng, S. Yan, and Z. Lu, “Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets,” in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 58-65.

[58] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020.

[59] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep Contextualized Word Representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, 2018, pp. 2227–2237.

[60] W. Yin and H. Schütze, "Learning word meta-embeddings," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Aug. 2016, pp. 1351-1360.

[61] R. ColloBERT and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of ICML*, 2008, pp. 160-167.

[62] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, “Improving word representations via global context and multiple word prototypes,” in *Proceedings of ACL*, 2012, pp. 873-882.

[63] A. Mnih and G. E. Hinton, “A scalable hierarchical distributed language model,” *Advances in Neural Information Processing Systems*, vol. 21, pp. 1081-1088, 2008.

[64] S. Sikorskiy, O. Metsker, A. Yakovlev, and S. Kovalchuk, “Machine learning-based text mining in electronic health records: cardiovascular patient cases,” in *Computational Science–ICCS 2018: 18th International Conference, Wuxi, China, June 11–13,* 2018 Proc.

[65] A. T. McCray, J. L. Sponsler, B. Brylawski, and A. C. Browne, “The role of lexical knowledge in biomedical text understanding,” in *Proceedings of the Annual Symposium on Computer Application in Medical Care, Medical Informatics Association*, 1987, p. 103.

[66] H. D. Tolentino, M. D. Matters, W. Walop, B. Law, W. Tong, F. Liu, and D. C. Payne, "A UMLS-based spell checker for natural language processing in vaccine safety," *BMC Medical Informatics and Decision Making*, vol. 7, pp. 1-13, 2007.

[67] D. B. Aronow, F. Fangfang, and W. B. Croft, “Ad hoc classification of radiology reports,” *Journal of the American Medical Informatics Association*, vol. 6, no. 5, pp. 393-411, 1999.

[68] L. Zhou and G. Hripcsak, “Temporal reasoning with medical data—a review with emphasis on medical natural language processing,” *Journal of Biomedical Informatics*, vol. 40, no. 2, pp. 183-202, 2007.

[69] L. Zhou, C. Friedman, S. Parsons, and G. Hripcsak, “System architecture for temporal information extraction, representation and reasoning in clinical narrative reports,” in *AMIA Annual Symposium Proceedings, American Medical Informatics Association*, 2005, pp.869.

[70] H. Harkema, A. Setzer, R. Gaizauskas, M. Hepple, R. Power, and J. Rogers, “Mining and modelling temporal clinical data,” in *Proceedings of the UK e-Science All Hands Meeting*, vol. 2005, 2005, pp. 507-514.

[71] C. V. Ananth and D. G. Kleinbaum, "Regression models for ordinal responses: a review of methods and applications," *International Journal of Epidemiology*, vol. 26, no. 6, pp. 1323-1333, 1997.

[72] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, pp. 273-297, 1995.

[73] L. Breiman, Classification and Regression Trees (1st ed.), Routledge, 1984.

[74] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[75] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, and N. Grayson, “Automatic ICD-10 classification of cancers from free-text death certificates,” *International Journal of Medical Informatics*, vol. 84, no. 11, pp. 956-965, 2015.

[76] F. Zhao, P. Skums, A. Zelikovsky, E. L. Sevigny, M. H. Swahn, S. M. Strasser, Y. Yan, and Y. Wu, "Computational approaches to detect illicit drug ads and find vendor communities within social media platforms," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 180-191, 2020.

[77] J. Jovanović and E. Bagheri, “Semantic annotation in biomedicine: the current landscape,” *Journal of Biomedical Semantics*, vol. 8, no. 1, pp. 1-18, 2017.

[78] A. Saxena, *Artificial Intelligence and Machine Learning in Healthcare*, Springer, Singapore, 2021.

[79] P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss,” *Machine Learning,* vol. 29, pp. 103-130, 1997.

[80] A. Yadav and D. K. Vishwakarma, “A weighted text representation framework for sentiment analysis of medical drug reviews,” in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, IEEE, 2020, pp. 326-332.

[81] M. van Diepen and P. H. Franses, “Evaluating chi-squared automatic interaction detection,” *Information Systems*, vol. 31, no. 8, pp. 814-831, 2006.

[82] Wikipedia. “Tschuprow’s T.” [Online]. Available: https://en.wikipedia.org/wiki/Tschuprow%27s_T. Accessed August 20, 2021.

[83] S. L. Salzberg, “Book Review: C4.5: Programs for machine learning,” M*achine Learning*, vol. 16, no. 3, p. 235, 1994.

[84] J. Žižka, F. Dařena, and A. Svoboda, *Text Mining with Machine Learning: Principles and Techniques*, CRC Press, 2019.

[85] J. Brownlee. (2016). *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch* (v1.1) (Online). Available: https://datageneralist.wordpress.com/wp-content/uploads/2018/03/master_machine_learning_algo_from_scratch.pdf.

[86] M. Y. Cheng, D. Kusoemo, and R. A. Gosno, “Text mining-based construction site accident classification using hybrid supervised machine learning,” *Automation in Construction*, vol. 118, pp. 103265, 2020.

[87] Y. M. Goh and C. U. Ubeynarayana, “Construction accident narrative classification: An evaluation of text mining techniques,” *Accident Analysis & Prevention*, vol. 108, pp. 122-130, 2017.

[88] D. Su, Q. Li, T. Zhang, P. Veliz, Y. Chen, K. He, P. Mahajan, and X. Zhang, "Prediction of acute appendicitis among patients with undifferentiated abdominal pain at emergency department," *BMC Medical Research Methodology*, vol. 22, pp. 1-14, 2022.

[89] C. Tu and M. Cui, “Learning regular expressions for interpretable medical text classification using a pool-based simulated annealing approach,” *in 2020 IEEE Congress on Evolutionary Computation (CEC)*, IEEE, 2020, pp. 1-7.

[90] J. L. Warner, M. A. Levy, and M. N. Neuss, “ReCAP: Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data,” *Journal of Oncology Practice*, vol. 12, no. 2, pp. 157-158, 2016.

[91] R. Botelle, V. Bhavsar, G. Kadra-Scalzo, A. Mascio, M. V. Williams, A. RoBERTs, S. Velupillai, and R. Stewart, "Can natural language processing models extract and classify instances of interpersonal violence in mental healthcare electronic records: an applied evaluative study," *BMJ Open*, vol. 12, no. 2, p. e052911, 2022.

[92] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[93] S. Pattanayak, *Intelligent Projects Using Python: 9 Real-world AI Projects Leveraging Machine Learning and Deep Learning with TensorFlow and Keras*, Packt Publishing Ltd., 2019.

[94] K. Fawagreh and M. M. Gaber, “Resource-efficient fast prediction in healthcare data analytics: A pruned Random Forest regression approach,” *Computing*, vol. 102, no. 5, pp. 1187-1198, 2020.

[95] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL 2020)*, Online, July 5-10, 2020, pp. 7871–7880

[96] Z. Guo, G. Nan, W. Lu, and S. B. Cohen, “Learning latent forests for medical relation extraction,” in P*roceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 3651-3657.

[97] Y. He, J. Li, Y. Song, M. He, and H. Peng, “Time-evolving Text Classification with Deep Neural Networks,” in *(IJCAI)*, 2018, pp. 2241-2247.

[98] Y. Kim, “Convolutional neural networks for sentence classification,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP’14),* 2014.

[99] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[100] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the NAACL Conference*, 2016, pp. 1480–1489.

[101] J. D. Lafferty, A. McCallum, and F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proceedings of the International Conference on Machine Learning*, 2001, pp. 3.

[102] K. Wang, S. C. Han, and J. Poon, “Induct-GCN: Inductive Graph Convolutional Networks for Text Classification,” in Proceedings *of the 26th International Conference on Pattern Recognition (ICPR),* IEEE, 2022, pp. 1243-1249.

[103] Gang Ding, "Research on Record Named Entity Recognition of Chinese Electronic Medical based on LSTM-CRF," *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2023*, 5-7 July, 2023, Hong Kong, pp88-93.

[104] H. Yang, L. Wang, and Y. Yang, “Named Entity Recognition in Electronic Medical Records Incorporating Pre-trained and Multi-Head Attention,” *AENG International Journal of Computer Science*, vol. 51, no. 4, pp. 125–134, 2024.

[105] Muneo Kushima, Tomoyoshi Yamazaki, and Kenji Araki, "Text Data Mining of the Nursing Care Life Log from Electronic Medical Record," *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2019*, 13-15 March, 2019, Hong Kong, pp257-261.

[106] Zhihuang Lin, Dan Yang, Hua Jiang, and Hang Yin, "Learning Patient Similarity via Heterogeneous Medical Knowledge Graph Embedding," IAENG International Journal of Computer Science, vol. 48, no.4, pp868-877, 2021

[107] Rushan Long, Dan Yang, and Yang Liu, "DiseaseNet: A Novel Disease Diagnosis Deep Framework via Fusing Medical Record Summarization," IAENG International Journal of Computer Science, vol. 49, no.3, pp808-817, 2022

[108] V. I. S. RamyaSri, C. Niharika, K. Maneesh, and M. Ismail, “Sentiment Analysis of Patients' Opinions in Healthcare Using Lexicon-based Method,” *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 6977–6981, 2019.

[109] Opim Salim Sitompul, Erna Budhiarti Nababan, Dedy Arisandi, Indra Aulia, and Hengky Wijaya, "Template-Based Natural Language Generation in Interpreting Laboratory Blood Test," IAENG International Journal of Computer Science, vol. 48, no.1, pp57-65, 2021

[110] A. Rajkomar, E. Loreaux, Y. Liu, J. Kemp, B. Li, M.-J. Chen, Y. Zhang, A. Mohiuddin, and J. Gottweis, ”Deciphering clinical abbreviations with a privacy protecting machine learning system,” *Nature Communications*, vol. 13, no. 1, pp. 7456, 2022.

[111] J. E. Rogers and A. L. Rector, “Terminological systems: bri dging the generation gap,” in *Proc. AMIA Annu. Fall Symp., American Medical Informatics Association*, 1997, pp. 610-614.

[112] G. Héja, G. Surján, and P. Varga, “Ontological analysis of SNOMED CT,” *BMC Med. Inform. Decis. Mak.*, vol. 8, pp. 1-5, 2008.

[113] P. Kestel, P. Kügler, C. Zirngibl, B. Schleich, and S. Wartzack, “Ontology-based approach for the provision of simulation knowledge acquired by Data and Text Mining processes,” *Adv. Eng. Inform.*, vol. 39, pp. 292-305, 2019.

[114] B. M. Hsu, “Comparison of supervised classification models on textual data,” *Mathematics*, vol. 8, no. 5, pp. 851, 2020. Available: https://doi.org/10.3390/math8050851.

[115] O. Bodenreider, “The unified medical language system (UMLS): integrating biomedical terminology,” *Nucleic Acids Res.*, vol. 32, no. suppl_1, pp. D267-D270, 2004.

[116] P. Bernus, J. Blazewicz, G. Schmidt, and M. Shaw, *International Handbooks on Information Systems*, 2nd ed. Springer Berlin Heidelberg New York, 2008. Available: https://www.springer.com/series/3795.

[117] M. Skreta, A. Arbabi, J. Wang, E. Drysdale, J. Kelly, D. Singh, and M. Brudno, ”Automatically disambiguating medical acronyms with ontology-aware deep learning,” *Nature Communications*, vol. 12, no. 1, pp. 5319, 2021.

[118] I. Harrow, R. Balakrishnan, E. Jimenez-Ruiz, S. Jupp, J. Lomax, J. Reed, M. Romacker, C. Senger, A. Splendiani, J. Wilson, and P. Woollard, ”Ontology mapping for semantically enabled applications, ” *Drug Discovery Today*, vol. 24, no. 10, pp. 2068–2075, 2019.

[119] The Human Phenotype Ontology.” Available: https://hpo.jax.org/app/about. Accessed June 23, 2022.

[120] G. Finley, “towards_comprehensive.” Available: https://github.com/gpfinley/towards_comprehensive. Accessed July 20, 2022.

[121] G. Deshpande, Q. Motger, C. Palomares, I. Kamra, K. Biesialska, X. Franch, G. Ruhe, and J. Ho, “Requirements dependency extraction by integrating active learning with ontology-based retrieval,” in *Proc. 2020 IEEE 28th Int. Requirements Engineering Conf. (RE)*, 2020, pp. 78–89.

[122] M. Chang, G. D’Aniello, M. Gaeta, F. Orciuoli, D. Sampson, and C. Simonelli, “Building ontology-driven tutoring models for intelligent

tutoring systems using data mining," *IEEE Access*, vol. 8, pp. 48151–48162, 2020.

[123] N. Nath, S.-H. Lee, M. D. McDonnell, and I. Lee, "The Quest for Better Clinical Word Vectors: Ontology-Based and Lexical Vector Augmentation Versus Clinical Contextual Embeddings," *Computers in Biology and Medicine*, vol. 134, pp. 104433, 2021.

[124] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, and A. Wong, "UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus," in *Proc. NAACL-HLT 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1744–1753.

[125] J. Meier, A. Dietz, A. Boehm, and T. Neumuth, "Predicting treatment process steps from events," *J. Biomed. Inform*., vol. 53, pp. 308–319, 2015.

[126] Y. Kim, C. Denton, L. Hoang, & A. M. Rush, "Structured Attention Networks," in *International Conference on Learning Representations*, November 2016.

[127] Y. Liu and M. Lapata, "Learning structured text representations," Trans. Assoc. *Comput. Linguistics*, vol. 6, pp. 63-75, 2018.

[128] A. Hannun, (2019). The Label Bias Problem. (Online). Available: https://awni.github.io/label-bias.

[129] A. Ritter, S. Clark, and O. Etzioni, "Named entity recognition in tweets: an experimental study," in *Proc. 2011 Conf. Empirical Methods in Natural Language Processing*, 2011, pp. 1524-1534.

[130] L. Zhu and H. Zheng, "Biomedical event extraction with a novel combination strategy based on hybrid deep neural networks," *BMC Bioinformatics*, vol. 21, no. 1, pp. 47, 2020.

[131] M. Bampa and H. Dalianis, "Detecting adverse drug events from Swedish electronic health records using text mining," in *Proc. LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, 2020, pp. 1-8.

[132] S. Bringay, "Fouille de données de santé," Ph.D. dissertation, Université de Montpellier, Montpellier, France, 2015.

[133] R. Navarro-Almanza, R. Juárez-Ramírez, G. Licea, and J. R. Castro, "Automated ontology extraction from unstructured texts using deep learning," in *Intuitionistic and Type-2 fuzzy logic enhancements in neural and optimization algorithms: Theory and applications*, 2020, pp. 727-755.

[134] H. S. Nguyen, M. H. Le, C. Q. L. Lam, and T. H. Duong, "Smart interactive search for Vietnamese disease by using data mining-based ontology," *J. Inf. Telecommun*., vol. 1, no. 2, pp. 176-191, 2017.

[135] C. Liu, F. Wang, J. Hu, and H. Xiong, "Temporal phenotyping from longitudinal electronic health records: A graph-based framework," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discov. Data Min*., 2015, pp. 705-714.

[136] D. Li, P. G. Lyons, C. Lu, and M. Kollef, "DeepAlerts: Deep learning-based multi-horizon alerts for clinical deterioration on oncology hospital wards, in " *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence,* 2020, pp. 743-750.

[137] M. L. Kolling, L. B. Furstenau, M. K. Sott, B. Rabaioli, P. H. Ulmi, N. L. Bragazzi, and L. P. C. Tedesco, "Data mining in healthcare: Applying strategic intelligence techniques to depict 25 years of research development," *International journal of environmental research and public health*, vol. 18, no. 6, pp. 3099, 2021.

[138] R. Qin, L. Duan, H. Zheng, J. Li-Ling, K. Song, and Y. Zhang, "An ontology-independent representation learning for similar disease detection based on multi-layer similarity network," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 183-193, 2019.

[139] I. Ioniţă and L. Ioniţă, "Applying data mining techniques in healthcare," Studies in Informatics and Control, vol. 25, no. 3, pp. 385-394, 2016.

[140] M. S. Islam, M. M. Hasan, X. Wang, H. D. Germack, and M. Noor-E-Alam, "A systematic review on healthcare analytics: application and theoretical perspective of data mining," *Healthcare*, vol. 6, no. 2, pp. 54, May 2018.

[141] H. Zhu, I. C. Paschalidis, and A. M. Tahmasebi, "Clinical Concept Extraction with Contextual Word Embedding," *NIPS Machine Learning for Health Workshop*, 2018.

[142] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of machine Learning research*, vol. 3, pp. 993-1022, 2003.

[143] ODBC Database Access. (Online). Available: https://cran.r-project.org/web/packages/RODBC/index.html. Accessed October 20, 2022.

[144] S. Vijayaraghavan and D. Basu, "Sentiment analysis in drug reviews using supervised machine learning algorithms," *arXiv preprint arXiv*:2003.11643, 2020.

[145] Y. Hao, X. Liu, J. Wu, and P. Lv, "Exploiting sentence embedding for medical question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 938-945, 2019.

[146] M. Abdollahi, X. Gao, Y. Mei, S. Ghosh, and J. Li, "An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimization," in *2019 IEEE Congress on Evolutionary Computation (CEC), IEEE*, 2019, pp. 119-12.

[147] International Classification of Diseases, Version 9 - Clinical Modification. (Online). Available: http://bioportal.bioontology.org/ontologies/ICD9CM. Accessed July 12, 2022.

[148] BETO: Spanish BERT. (Online). Available: https://github.com/dccuchile/beto. Accessed June 20, 2022.

[149] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Sackinger, and R. Shah, "Signature Verification Using a Siamese Time Delay Neural Network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 669–688, 1993.

[150] C. Dalloux, N. Grabar, and V. Claveau, "Détection de la négation: corpus français et apprentissage supervisé," *Revue des Sciences et Technologies de l'Information-Série TSI: Technique et Science Informatiques*, vol. 1, pp. 1-21, 2019.

[151] S. Sabra, K. M. Malik, and M. Alobaidi, "Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives," *Computers in Biology and Medicine*, vol. 94, pp. 1-10, 2018.

[152] S. Perera, P. Mendes, A. Sheth, K. Thirunarayan, A. Alex, C. Heid, and G. Mott., "Implicit entity recognition in clinical documents," in *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, 2015, pp. 228-238.

[153] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," In *AAAI*, 2006, vol. 6, pp. 775-780.

[154] Z. Wu, and M. Palmer, 1994. "Verbs semantics and lexical selection, ," *In Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, ACL '94*, Stroudsburg 1994, pp. 133–138.