IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 December 2022, accepted 26 January 2023, date of publication 30 January 2023, date of current version 3 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3240898

SURVEY

Systematic Literature Review of Information Extraction From Textual Data: Recent Methods, Applications, Trends, and Challenges

MOHD HAFIZUL AFIFI ABDULLAH, (Graduate Student Member, IEEE), NORSHAKIRAH AZIZ, SAID JADID ABDULKADIR, (Senior Member, IEEE), HITHAM SEDDIG ALHASSAN ALHUSSIAN, AND NOUREEN TALPUR

Centre for Research in Data Science (CeRDaS), Computer Information Science Department, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia

Corresponding author: Mohd Hafizul Afifi Abdullah (mohd_20002084@utp.edu.my)

This research was supported by Universiti Teknologi PETRONAS through Yayasan Universiti Teknologi PETRONAS (YUTP) with Grant Number: 015LC0-277.

ABSTRACT Information extraction (IE) is a challenging task, particularly when dealing with highly heterogeneous data. State-of-the-art data mining technologies struggle to process information from textual data. Therefore, various IE techniques have been developed to enable the use of IE for textual data. However, each technique differs from one another because it is designed for different data types and has different target information to be extracted. This study investigated and described the most contemporary methods for extracting information from textual data, emphasizing their benefits and shortcomings. To provide a holistic view of the domain, this comprehensive systematic literature review employed a systematic mapping process to summarize studies published in the last six years (from 2017 to 2022). It covers fundamental concepts, recent approaches, applications, and trends, in addition to challenges and future research prospects in this domain area. Based on an analysis of 161 selected studies, we found that the state-of-the-art models employ deep learning to extract information from textual data. Finally, this study aimed to guide novice and experienced researchers in future research and serve as a foundation for this research area.

INDEX TERMS Information extraction, text extraction, named entity, named entity recognition, relation extraction, event extraction, deep learning.

I. INTRODUCTION

For decades, businesses and organizations have created vast amounts of data in various structures (e.g., structured, semi-structured, and unstructured) and formats. Consequently, a considerable amount of this information is either deposited locally as physical or digital files, or stored in a network in data warehouses, data lakes, and network storage. The files stored in physical or digital forms are kept for reference or later use. Extracting and utilizing this stored information can provide many untapped opportunities for businesses and organizations. For example, curating insights from this vast amount of data enables organizational leaders to make informed decisions and empower their organizations through a data-driven approach [1], [2]. It extends to extracting valuable data from job advertisement texts, which businesses or advertisers may use to obtain useful information, build job profiles, and limit relevant information to end users [3].

Other researchers have demonstrated several information extraction (IE) applications from textual data, including business and investment applications. For instance, .Selimi and Besimi [4] demonstrated a model for stock market forecasting based solely on mining or extracting financial news and sentiment analysis. Krstić et al. [2] presented a method to visualize big data texts (tweet data), which helped a firm (bank operator) better understand its clients and make better business decisions. Textual data extraction or mining is

The associate editor coordinating the review of this manuscript and approving it for publication was Giacomo Fiumara.

also important in text classification tasks for languages with limited resources [5].

Nevertheless, our technological capabilities to process such data, which come in various formats and structures, have hindered us from utilizing the information within [6], [7], [8]. It is generally understood that unstructured data cannot be processed directly by computing equipment since the information is not predefined in a particular structure [9], [10]. For example, techniques used to extract information from text data are incompatible with other data types, such as audio, video, and image data. Thus, different encoding mechanisms are required to extract information from each type of unstructured data [6], [10]. As unstructured data (including textual data) can be arranged or represented in various ways, the extraction method must specify mechanisms to extract the correct information from each data arrangement or representation. This makes IE using traditional methods difficult and impractical.

To address this issue, researchers have explored other approaches for extracting useful information from textual data using various IE techniques. These approaches aim to extract entities, relations, facts, events, objects, and other information from unstructured data and transform them into a structured form [6], [11]. The primary objective of IE in textual data is to label semantic information from the text and transform it into structured data. The extracted information can be used in a data analysis pipeline for data preparation before data preprocessing and analysis [6], [11]. An effective and efficient IE technique can accurately transform unstructured data into a structured form, leading to valid data preprocessing and improved performance during data analysis [8], [11], [12]. IE techniques can be divided into three categories: (i) traditional, (ii) machine learning (ML), and (iii) deep learning (DL)-based methods.

Baars and Kemper [9] documented the traditional methods for extracting information from textual data, which include: (i) integrated presentation, (ii) analysis of content collections, and (iii) distribution of analysis results and analysis templates. In the first method (integrated presentation), both structured and unstructured content are accessed simultaneously via an integrated user interface, navigation, and search functionalities. The advantage of this method is that it enables users to search and navigate structured and unstructured data simultaneously. By contrast, the automatic juxtaposition of search results can uncover hidden interrelationships between these data [9]. The disadvantage of the integrated presentation method is that the retrieved data are only used for viewing purposes and not extracted for processing. Alternatively, the second method (analysis of content collections) allows the extraction of metadata from unstructured data collections and stores them as structured data. Abdullah and Ahmad [13] implemented this method to ''transform'' unstructured data into a structured form by classifying and mapping metadata from unstructured data into structured database tables. This process comprises four main steps: (i) extraction, (ii) classification, (iii) repository development, and (iv) data mapping. The mapping process provides a more structured and comprehensive collection of thematic information for business decision-making [12]. The advantage of the second method is that the mapped metadata can be treated as structured information and used for analytical purposes. However, the disadvantage of this method is that an analytical system will not be able to access the actual content of the unstructured data. The third method (the distribution of analysis results and templates) offers the advantage of being more efficient and providing more accurate IE results. The disadvantage of this method is that it is more complicated and requires more steps. Thus, advanced skills are required for users to perform extensive recalibration and parameter optimization to achieve the desired output.

Several studies employed ML to extract useful information from textual data to address these problems. For instance, Chai et al. [14] presented a model for extracting sentiment information from textual data sources (e.g., news event data and investor comments) using natural language processing (NLP). The extracted sentiment data were then trained using stock trading data and an extended Hidden Markov Model (HMM). The experimental results show that the model effectively extracts positive and negative sentiments from textual data and predicts future stock price movements. In another study,Jenhani et al. [15] presented an ML model hybridized with a domain-based dictionary and linguistic rules for extracting drug abuse information from social media data (tweets). The learning model was trained and tested using 1,000,000 tweets from five different experimental settings. Based on their findings, ML alone was insufficient to produce satisfactory results when performing IE tasks using social media data. Combining ML with a domain-based dictionary and linguistic rules helps the model extract tweet data accurately into useful, structured data.

The studies mentioned above show that ML can be utilized for training models to extract information from textual data. However, ML methods struggle to handle massive amounts of highly complex data (e.g., complex compound texts, nested entities, and variety in data representation) [8], [11], [16]. For instance, Jenhani et al. [15] noted that the volume, velocity, and variety issues of big data cannot be effectively managed by a simple implementation of the hybrid ML model.

Several recent studies have investigated DL methods for extracting useful information from textual data. For instance, Zhang and El-Gohary [17] presented a DL-based model for extracting semantic and syntactic information from regulatory documents. The model significantly outperformed the other two baseline models in terms of precision, recall, and F1 measures for the closed-domain IE tasks. In another study, Liu et al. [18] presented a DL-based model called BERT-BiLSTM-MHATT-CRF (BBMC) [18] to identify and extract complex biochemical entities from scientific documents. The developed BBMC model comprises four different algorithms: (i) Bidirectional Encoder Representations from

Transformers (BERT) model to extract features in the text, (ii) bidirectional long short-term memory (BiLSTM) to learn the context represented in the text, (iii) multi-head attention (MHATT) to extract chapter-level features, and (iv) conditional random field (CRF) to label the sequence tag. The experimental results of this DL-based model showed significant improvement in identifying and extracting complex biochemical names from the datasets used for testing. According to [18], the DL-based method employing the BBMC model has significantly improved the extraction accuracy compared to the conventional BiLSTM-CRF algorithm. The study also concluded that DL-based methods can effectively identify and extract complex information from textual data. However, studies on IE using DL-based methods are still in their infancy, and both novice and experienced researchers have much ground to cover [8], [19], [20].

We firmly believe that the most recent advancements in this field provide a strong impetus for examining and discussing the current methods, applications, trends, research challenges, and future research directions. Thus, we structured our research around a compelling central point by defining specific research questions (RQs) that form a strong connection between the concepts surrounding the main objective. The following are the formulated RQs and their motivations for this study:

RQ1: What are the concepts related to IE from textual data?

Motivation: To discuss the fundamental concepts and knowledge related to IE from textual data.

RQ2: What are the current techniques for extracting textual data from unstructured data?

Motivation: To discuss the current techniques available for IE from textual data.

RQ3: What are the real-world applications of IE from textual data?

Motivation: To investigate the current application of IE from textual data in various domain areas and suggest potential future directions.

RQ4: What is the intensity of publications related to IE from textual data?

Motivation: To explore the intensity of publications and current research trends related to IE from textual data.

To the best of our knowledge, no comprehensive systematic literature review (SLR) study has been conducted with the sole focus on highlighting the current methods, applications, trends, and challenges in the domain of IE from textual data with detailed facts and figures. This SLR covers the literature published within the last six years (2017-2022) with the following main contributions:

1) This study presents a comprehensive methodology with a four-step study mapping process to systematically search studies related to IE from textual data (Figure 1).
2) Next, this study explains the fundamental concepts and knowledge of IE from textual data (Section IV-A).
3) This study describes the recent techniques used to extract information from textual data, including their purposes, mechanisms, and implementations (Section IV-B).
4) Furthermore, it highlights the practical applications of each IE technique for textual data categorized by domain area (Section IV-C).
5) This study conveys the status quo of research in this domain by visualizing and highlighting key insights from data synthesized through a systematic mapping process (Section IV-D).
6) Finally, this study discusses the current challenges in this research domain (Section V-A) and suggests future research directions (Section V-B).

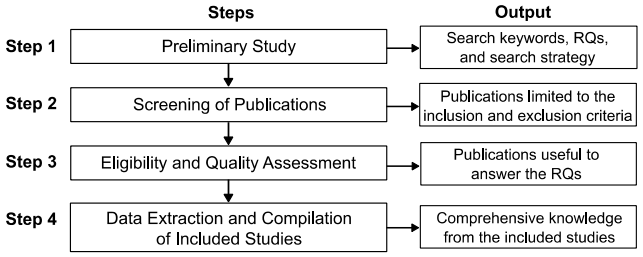

FIGURE 1. The literature study mapping process.

The findings of this study will aid researchers in capturing the current state of research by providing a thorough understanding of the most recent methods for extracting information from textual data, their applications, and current research trends. Consequently, researchers may find it beneficial to view the entire landscape of this domain from a general to a more specific perspective. Additionally, insights based on data from dissected studies will help create a paradigm for future development and advancement by describing recent challenges, research opportunities, recommendations, and future research directions. We strongly believe this study will serve as a central point of reference for novice and experienced researchers in understanding the current state of research, related fundamental concepts, and future research directions in this domain area.

The rest of this paper is organized as follows. Section II summarizes the available SLR/survey studies from the literature, and Section III defines the methodology used to conduct this systematic study. Section IV answers all the RQs presented earlier based on synthesized data from the included studies. Section V discusses current challenges and future recommendations for each RQ. Section VI presents the strengths and limitations of this SLR study. Finally, Section VII concludes the study.

II. LITERATURE REVIEW

The extensive growth of the Internet, mobile devices, social networks, and sensors has attracted researchers to develop and use various IE techniques to deliver valuable insights from data. Several studies have been published to date focusing on providing the most recent advancements in the domain of IE from textual data. However, few studies have provided comprehensive insights into the studies published in this domain. Since the IE task for unstructured data is still

unresolved, additional research activities have been attempted to design and develop robust methodologies for handling the extraction process. Therefore, the primary aim of this section is to highlight existing SLR and survey studies that focus on the domain of IE from textual data and summarize the motivations, contributions, advantages, and shortcomings of each related work. These studies were carefully selected based on their relevance to deliver an overview and status quo for research on IE from textual data.

First, we dissected the systematic studies performed in [6] and [16] that focused primarily on contemporary IE techniques while examining the limitations of each technique for various data types. The aforementioned studies described recent challenges, recommended potential solutions, and suggested future directions for research related to IE from big data. They also contributed significantly by presenting an overview of IE from heterogeneous unstructured data and compiling recent methods for extracting information from various data types (text, images, audio, and video) in a single SLR study. They provide a holistic view of IE, which helps novice researchers grasp the fundamental concepts of big data IE.

In contrast to previous studies, several SLR and survey studies have focused on Named Entity Recognition (NER) techniques. For example, a survey study performed by Etaiwi et al. [21] mainly focused on statistical approaches to NER for Arabic texts, which explained the working mechanisms of each statistical approach for the NER task. According to their findings, the implementation of NER in Arabic texts can be classified into six approaches: HMM, CRF, Naïve Bayes (NB), neural networks, entropy, and Support Vector Machines (SVM). In contrast, Goyal et al. [22] published a detailed SLR study of NER and its classification techniques. This study has provided essential knowledge, interesting developments, and progress related to NER from a technical perspective. Additionally, it offers future directions for this domain to encourage continuous improvement in the research community.

Next, it is crucial to emphasize studies inclined toward contemporary IE techniques from textual data. For example, Xiang and Wang [23] compiled the most recent techniques for extracting event-related information from textual data. Their study provides in-depth details on the fundamentals of IE, while comparing the strengths and weaknesses of each technique. Akkasi and Moens [24] published a survey of state-of-the-art models for extracting causal relationships from textual data using DL and deep neural network (DNN) approaches. Their study covers not only state-of-the-art methods, but also the applications, datasets used, and remaining challenges in the area. In addition, Nasar et al. [25] concluded that DL-based and joint models are state-of-the-art methods for extracting information from text.

Other SLR and survey studies have provided significant information on IE from textual data. However, these studies focused only on specific or more IE techniques for all data types rather than providing a comprehensive review of the current IE techniques for textual data. Table 1 summarizes the relevant SLR and survey studies, including their motivations, contributions, advantages, and limitations.

From our exploration, we found 12 relevant SLR and survey studies published within the last six years. It is important to note that these studies focused on various aspects of IE techniques for textual data. For example, studies [6] and [11] have concentrated on summarizing studies related to IE techniques for diverse data types (i.e., text, images, audio, and video). Studies [21], [22], and [26] focused on NER techniques only, whereas Akkasi and Moens [24] focused on relation extraction (RE) techniques, and Xiang and Wang [23] focused on event extraction (EE) techniques. Based on the literature, only a few survey studies [23], [24], [25] have implemented DL-based methods for extracting useful information from text. The primary goal of these studies was to develop general concepts of IE from textual data related to a specific technique or comparison from a broader perspective.

The current literature does not provide a comprehensive and updated review that focuses on the existing and contemporary methods, applications, trends, and challenges in this study area. Although these studies help us grasp the fundamental concepts of IE, few do not highlight recent trends or advancements in this research domain. Furthermore, no study currently offers a thorough analysis of different IE techniques, covering well-known and modern techniques, while focusing only on textual data.

This study not only intends to provide researchers with full knowledge of IE from textual data, but also aims to help guide novice and experienced researchers in shaping future research directions for IE from textual data. Therefore, our systematic survey aims to present the current IE methods for textual data, their applications, and current research trends, while describing the challenges and future research prospects in this domain area. This study includes the most recent articles and covers literature published between 2017 and 2022 to provide the most recent information on advances in this field. The following section (Section III) thoroughly explains the methodology used to construct this SLR study to achieve earlier goals.

III. METHODOLOGY

The methodology used in this study was adopted based on the SLR guidelines [31], [32]. Additionally, this systematic study followed the reporting style and flow of comprehensive studies [33], [34]. The modified mapping process employed in this study is shown in Figure 1 and further explained in Section III-A. Further details of the actual mapping process can be found in previous studies [31], [32].

A. STEP 1: PRELIMINARY STUDY

The literature review began with a preliminary study, wherein an initial search was performed to understand better the main topic being addressed. This step also serves as a ''kick-off'' motivation for authors to find relevant questions, keywords, and scope for this systematic study. From this initial stage,

TABLE 1. Summary of related works from SLR and survey studies.

Authors and Year	Motivations	Contributions	Advantages	Limitations
Etaiwi et al. [21], 2017	To conduct a comprehensive survey about statistical approaches to Arabic NER.	Provides a comprehensive study of NER on Arabis texts, which is less common in this domain area.	Explains the challenges and suitable extraction approaches for NER on Arabic texts.	The scope is limited to only NER techniques (specifically on Arabic text). Inadequate discussion on the application areas and trends of research in the domain area.
Goyal et al. [22], 2018	To explore and highlight the status of NER classification techniques, issues, challenges, and factors affecting their performances.	Analyzed and presented the factors affecting NER performance, with existing challenges and issues.	Provides in-depth analysis and explanation of the most recent NER and its classification techniques.	Less focus on DL-based approaches. The study was published in 2018. Thus, there is a more recent advancement in NER methods and strategies.
Adnan and Akbar [6], 2019	To review state-of-the-art methods for extracting information from various types of big data.	Presents the task-dependent and task-independent limitations of IE covering all data types in one study.	Covers IE strategies from a broad spectrum of data types (text, image, audio, video).	Does not provide an in-depth analysis of textual data, as the study covers literature related to various data types. Less discussion on IE approaches utilizing DL.
Xiang and Wang [23], 2019	This article aims to provide a comprehensive and up-to-date survey of EE from the text.	Provides a thorough analysis of EE, including the task or challenge, data sources, performance assessment, and future research direction.	Provides the latest in-depth analysis, including the fundamentals, strengths, and weaknesses of each EE approach up to the year 2019.	Good for understanding the recent EE approaches but lacks mathematical explanation to cater to advanced researchers.
Albared et al. [26], 2019	To assess the improved performance of NER for the most popular datasets and determine the best system recently proposed.	Tracks progress improvement in NER accuracy over the 4 most common datasets: CoNLL2003, OntoNotes v5, SciERC, and BC5CDR from 2016-2019. Several challenges are still open for further research.	Provides knowledge of accuracy improvement over popular datasets not presented in other studies. This paper also provides information regarding the methods used and their corresponding result.	Limited explanation of the contemporary IE methods, such as the DL-based approaches. The progress tracking is until 2019 only. Therefore, there may be further progress that is not captured until now.
Adnan et al. [16], 2019	To review the IE process and method based on the various data types (text, image, audio, video), and the desired output (structured information).	This study identifies the problems of the IE process when dealing with unstructured data for performance improvement. It emphasizes the need for a multistep IE pipeline to deal with the massive volume of unstructured data.	A single paper comprehensively reviews the IE pipeline process for various data types (text, image, audio, video).	Does not cover the most contemporary methods, such as using DL, DNN, Convolutional Neural Networks (CNN), and others. Limited discussion on the application areas and trends of research in the domain area.
Pejić Bach et al. [27], 2019	To provide a comprehensive review of the existing techniques used in text mining for corpora related to the financial sector.	This study makes a theoretical contribution through citation and co-citation analysis of studies and research trends in the area. It compiled a.	Provides a comprehensive list of text mining techniques based on data sources and typical applications in the financial sector.	Good for understanding the text mining techniques suitable for financial corpora, but the study is limited until 2019 and does not cover DL-based methods.
Liu et al. [28], 2020	To survey and summarize the most recent research progress focusing on EE and event RE.	Provides knowledge of the task description required for the EE and event RE. This study also presents the main challenges, the most common datasets used, and future research direction.	Provides a brief and straightforward explanation regarding recent research progress in the domain area. Utilizes suitable visualization and graphics to aid readers' comprehension of the topics discussed.	Less discussion on DL-based methods to support IE. Has limited explanation for extracting events on a document level.
Akkasi and Moens [24], 2021	To conduct a comprehensive survey on state-of-the-art models for cause-effect relationships and investigate the class imbalance problems.	Provides a comprehensive survey on cause-effect relationship (relation) extraction using DNN models.	In-depth review covering the state-of-the-art methods, applications, datasets available, and remaining challenges.	The review scope is limited to causal-relationship (RE) extraction. Does not cover application areas other than medical text.
Nasar et al. [25], 2021	To provide a summarized description of methods to extract "structured" insights from textual data, focusing on two subdomains of IE, which are NER and RE.	Based on their analysis, this study concludes that the DL-based hybrid technique and joint models are currently state-of-the-art models.	It covers the early approaches until the state-of-the-art models while focusing on advancement via the DL-based approach.	The study is mainly a survey rather than an SLR focusing on the trend and progress of IE from textual data. No discussion on the application

TABLE 1. *(Continued.)* **Summary of related works from SLR and survey studies.**

Authors and Year	Motivations	Contributions	Advantages	Limitations
				areas and trends of research in the domain area.
Frisoni et al. [29], 2021	To survey the recent methods for extracting complex events-related information (relations, participants, nested events, nested roles) from biomedical literature.	Provides a flexible definition of an event, EE tasks, challenges, and state-of-the-art methods for extracting events from biomedical literature.	Provides an in-depth survey on event extraction and NL in the biomedical domain. Each method of extracting events is analyzed thoroughly and discussed in this study.	The survey paper is limited to only studying EE literature from texts from the biomedical domain. However, it also provides the gist of EE-related ideas from other domains.
Li et al. [30], 2022	To provide a comprehensive review of the state-of-the-art methods implementing DL techniques of NER.	This study compiles the resources needed for NER research (i.e., annotated corpora and tools), categorizes the NER-related works into three groups, and presents the current challenges and future directions for NER research.	Compares various NER architectures, including DL-based methods for NER. It also identifies and explains the factors affecting NER's performance.	This study presents various implementations of the most recent NER models. This creates a challenge for novice researchers to choose the simplest or the right model to get started.

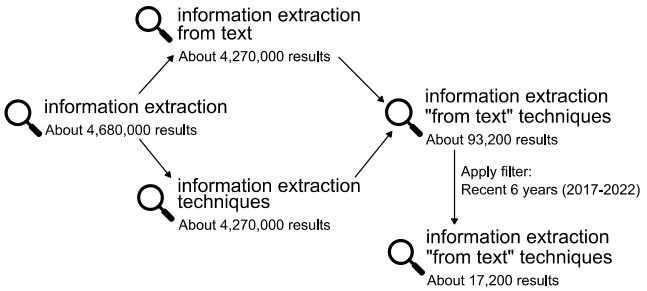

FIGURE 2. **The results of the keyword combination survey from Google Scholar.**

we conducted three tasks: identifying keywords, formulating RQs, and identifying search criteria and strategies.

1) KEYWORDS IDENTIFICATION

Our initial search was performed on Google Scholar using the keyword ''information extraction'' to survey the number of studies available on this topic before selecting keyword variations and other relevant keywords. From this step, we found four relevant studies [11], [16], [27], [35] that we believe will help to answer the RQs. The retrieved studies were used to identify a few more relevant keywords and obtain general ideas about the search venue. Prior to finalizing the search string, keywords were analyzed to determine the most appropriate keyword combinations that would return the most relevant articles related to ''information extraction strategies for textual data''. Figure 2 illustrates the results of the keyword combination survey, including the number of studies retrieved during the literature search using Google Scholar.

2) IDENTIFY SEARCH CRITERIA AND STRATEGIES

Next, it is essential to pivot the search into specific areas using search criteria. We selected five established databases for the literature search: IEEE Xplore, ScienceDirect, SpringerLink, Scopus, and ACM Digital Library. The search scope for retrieving related studies focused on the titles, abstracts, and author keywords. Meanwhile, the publication venue is limited to journals, conference proceedings, and book chapters published within the recent six years (2017–2022).

TABLE 2. **Domain focus, keywords, and the generalized search string.**

Domain Focus	Keywords	Generalized Search String
IE from textual data	Information extraction, text extraction, Named Entity Recognition, Named entity extraction, Event Extraction, Relation extraction, Rule-based IE	((“Information extraction” OR “data extraction”) AND (“from text”)) OR ((“Named Entity Recognition”) OR (“Named entity extraction”) OR (“Event Extraction”) OR (“Relation extraction”)) OR (extract* text data)

The search results were monitored to ensure that they contained information that could answer the RQs mentioned in Section I. The selected search keywords were combined with suitable advanced search operators and wildcards following the manual for each database (Appendix A). The domain focus, keywords, and generalized search strings used for the search are provided in Table 2.

From the search, 845 studies were found using the search criteria and strings (Table 2) executed on five scholarly databases: IEEE Xplore, ScienceDirect, SpringerLink, Scopus, and ACM Digital Library. Four additional studies [11], [16], [27], [35] already in hand were included in the total number of identified studies and labeled as others, making 849 identified studies in total.

The latest database search was performed on July 31, 2022. Relevant studies were identified and selected based on PRISMA standards [32]. The flow of information throughout the various phases of this SLR study is depicted in a PRISMA flowchart (Figure 3). Appendix B presents the PRISMA 2020 Checklist prepared for this study.

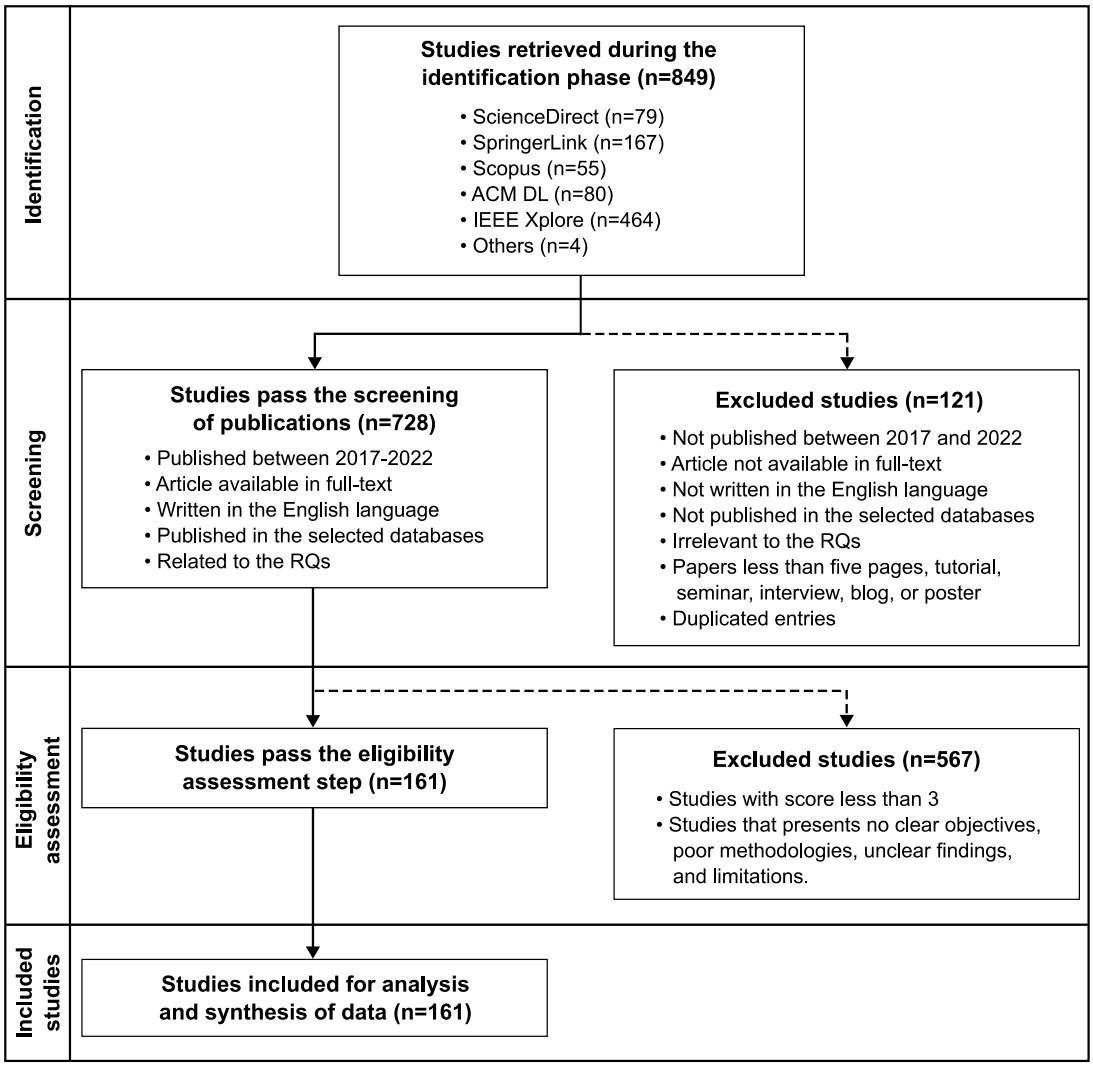

FIGURE 3. PRISMA flowchart of the systematic literature process.

TABLE 3. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
• Published between 2017–2022.	• Published before 2017 or after 2022.
• Available in full text.	• The study is unavailable in full text.
• Written in the English language.	• Written in other than English.
• Relevant to the RQs.	• The study is not relevant to the RQs.
	• Short papers (less than five pages).
	• Duplicates or redundant studies.

B. STEP 2: SCREENING OF PUBLICATIONS

The screening of publications involved a two-step process. First, the studies were screened based on inclusion and exclusion criteria (Table 3). All 849 identified studies retrieved in the previous step were screened to exclude those not meeting the selection criteria.

Subsequently, a formal analysis and data curation team assessed the literature to determine the relevance of the studies based on their titles, abstracts, and conclusions. A thorough examination of each literature content was performed rigorously based on their relevance to the topic to ensure that we did not miss any important studies. The studies were selected based on their relevance to the RQs and themes of the study. As a result, 728 studies passed the screening, and 121 were excluded from the 849 identified studies.

C. STEP 3: ELIGIBILITY AND QUALITY ASSESSMENT

After the screening process, the remaining 728 studies were assessed for eligibility and quality prior to inclusion in the SLR study. The eligibility and quality assessment were conducted based on the criteria with scoring values of 1 (agree), 0.5 (partly agree), and 0 (disagree), as shown in Table 4. This was performed to ensure that only studies with clear objectives and goals, methodologies, limitations, and findings were included in the final selection.

The scoring criteria described in the table above were used for each study during the eligibility and quality assessment steps. Only studies with a minimum score of 3.0 were selected

TABLE 4. Scoring criteria for the eligibility and quality assessment.

Criteria	Score	Description
Does the study provide clear objectives and goals?	1	Yes, the study presents clear objectives and goals.
	0.5	The study presents its objectives, but the goals of the study are not clearly defined.
	0	No, the study does not provide clear objectives and goals.
Does the study present a clear methodology?	1	Yes, the study presents clear, systematic, and well-documented methodology.
	0.5	The methodology documentation is present but incomplete/not systematic.
	0	No, the study presents poor/missing methodology documentation.
Does the study present limitations of the work?	1	Yes, the study provides a well-acknowledged limitation of the study.
	0.5	The study states its limitations, but not in detail.
	0	No, the study does not provide a declaration of limitations of the study.
Does the study present clear research findings?	1	Yes, the study presents clear, comprehensive, and well-presented research findings. Suitable visualizations are used to present the findings/results.
	0.5	The study findings were presented, but more explanation could be provided. The data presented relate to the results.
	0	No, the study does not present clear research findings, and/or elaboration is not provided. Results are irrelevant to the objectives and goals of the study and are presented in random order.

to ensure that only high-quality studies were included in the final list. Thus, 161 out of 728 studies were selected for inclusion in the qualitative synthesis of this SLR study, while the remaining 567 studies were excluded. The decision to exclude studies that did not meet the scoring criteria was not made lightly and was made following a thorough eligibility and quality assessment process.

D. STEP 4: DATA EXTRACTION AND COMPILATION OF INCLUDED STUDIES

Once the eligibility and quality assessments were completed, the list of the final selected studies and their metadata were downloaded from the scholarly databases and compiled into a spreadsheet. The information gathered from these publications includes title, authors, journal or conference name, publication venue (i.e., journal, conference, book chapter, or journal preprint), paper type (i.e., application, review, SLR, research, or survey), publication year, digital object identifier (DOI), the technique used, and the scholarly database used to retrieve relevant studies.

EndNote software was used to automatically eliminate duplicates and store offline copies of the studies for reference and citation purposes. Additionally, the list of references in each study was used to locate additional relevant references that could provide valuable information regarding IE from the textual data. This step was performed to extract knowledge that could help answer the RQs and the goals of this study.

TABLE 5. Studies retrieved in each step of systematic literature process.

Scholarly Databases	Studies Identified	Pass Screening	Studies Pass EA	Studies Included
IEEE Xplore	464	406	100	100
ScienceDirect	79	74	32	32
SpringerLink	167	150	13	13
Scopus	55	28	6	6
ACM DL	80	66	6	6
Others	4	4	4	4
Total	849	728	161	161

EA: eligibility assessment.

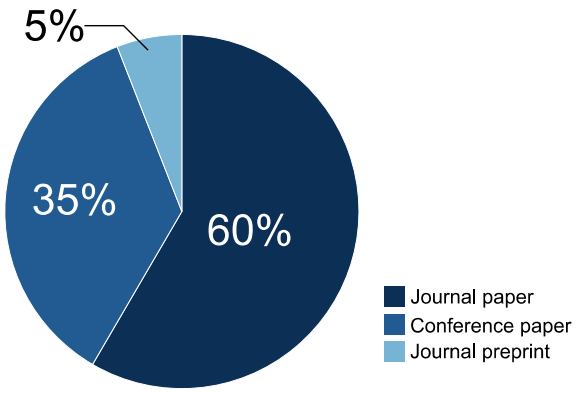

FIGURE 4. The distribution of included studies by publication venue.

Finally, 161 studies were included and dissected to answer the RQs established in this systematic review. The most valuable data exploited from these studies are related to existing IE methods, their applications, research trends, and current challenges in the domain of IE from textual data. Table 5 presents the number of studies retrieved from scholarly databases in each step of the SLR process.

Our SLR only considered studies with clear objectives and goals, provided clear methodological explanations, stated their limitations, and presented clear research findings for inclusion in the final selected studies. Therefore, the information in Table 5 leads us to the conclusion that high-quality and relevant studies can mostly be found in IEEE Xplore (100 studies), followed by ScienceDirect (32 studies), SpringerLink (13 studies), Scopus (6 studies), ACM DL (6 studies), and other sources (4 studies). Based on the included studies, high-quality materials relevant to this domain were published in journals (60%), conference proceedings (35%), and journal preprints (5%). Figure 4 depicts the distribution of the included studies by publication venue.

IV. SYNTHESIS OF DATA AND ANALYSIS

This section aims to synthesize and summarize the information from the included studies into visualization aids and answer the RQs presented earlier. The data synthesis

TABLE 6. Terms related to text-based IE.

Term	Definition
Natural language	Any language that humans use to communicate.
Natural language processing	A subfield of artificial intelligence that deals with processing vast amounts of NL data.
Closed-domain IE	It is an IE task limited to a specific domain area.
Open-domain IE	It is an IE task that does not limit to a specific domain.
Text corpus	A text corpus is a language resource consisting of a huge volume of texts in structured forms.
Knowledge graph	It is a knowledge base that utilizes a graph-structured data model to connect the data points.

TABLE 7. Comparison of closed-domain IE and open-domain IE.

Criteria	Closed-domain IE	Open-domain IE
Objective	To discover and extract topics and arguments from the text.	To discover and extract topics of interest from any text.
Text-domain specificity	Extracts only information from a specific domain.	Able to handle IE across various domains.
Extraction mechanism	Requires predefined schema/ structures for extracting information.	Does not require predefined schema/ structures for IE tasks.
Extraction output	Topic and its arguments.	Extracted knowledge clustered by topics.

presented in this section aims to help novice and experienced researchers recognize the current state of research, recent applications, and challenges related to IE from textual data.

A. RQ1: WHAT ARE THE CONCEPTS RELATED TO IE FROM TEXTUAL DATA?

Several concepts closely related to IE from textual data include but are not limited to natural language (NL), Natural Language Processing (NLP), closed-domain IE, open-domain IE, text corpus, and knowledge graph (KG). The definitions of each term are presented in Table 6.

The following subsections explain each term and concept presented in Table 6.

1) NATURAL LANGUAGE (NL) AND NATURAL LANGUAGE PROCESSING (NLP)

NL is any language humans use to communicate, and it has naturally evolved through use and repetition. Examples of natural languages include English, French, German, Spanish, and Chinese. Some languages vary according to geographical locality. For instance, the English language used in the United States is American English (US English). In contrast, the variation of the same language used in Great Britain is British English (UK English). Despite some similarities, the two languages are quite different, as their usage has evolved over the centuries. These differences can be in spelling (for example, ''color'' is the correct spelling in US English and ''colour'' is the spelling in UK English), meaning (for example, ''sidewalk'' in US English refers to a paved walkway for pedestrians, but ''footpath'' is used in UK English). Some words may not be available in other language variations (for example, ''dosh'' in UK English means cash, but it does not exist in the US English dictionary). Similarly, the Malay language used in Malaysia is Bahasa Melayu, whereas the variation of the same language used in Indonesia is Bahasa Indonesia.

It is also important to note that most languages use capitalization in written form to represent entities (e.g., names, organizations, and locations). However, capitalization is absent in written forms of languages such as Arabic, Urdu, Hindi, Tamil, Japanese, Korean, and Chinese. ThisIt challenges machines to identify entities within the written texts in such languages. In addition, the poor morphological structure of a specific language makes it more complicated for machines to understand the context of a sentence or text.

NLP is a subset of artificial intelligence (AI) and is concerned with a mechanism that allows computers to handle and analyze massive amounts of NL text. NLP employs algorithms to deduce the meaning of sentences as they are spoken and written. It has many real-world applications, including real-time translation, automatic subtitle generation, text prediction in search engines, and business intelligence. However, it is difficult to extract information from texts because NL is highly abstract, and computers must comprehend text semantics to accurately extract appropriate information accurately [36]. Therefore, IE tasks are divided into two fundamental challenges: (i) closed-domain IE and (ii) open-domain IE [23].

2) CLOSED-DOMAIN IE AND OPEN-DOMAIN IE

In general, text extraction requires a system that can handle data originating from different text sources based on a predefined structure or schema. An extraction system aims to extract specific information (e.g., characters, words, phrases, named entities, events, or relations) from a text body and fill in the information into the given schema, thus generating the output as structured information. In a closed-domain IE, extracting information from data originating from various domains may require several IE schemas to complete a task [23]. On the other hand, open-domain IE may not require a predefined schema as the main task is to extract keywords from texts and cluster similar texts to identify the topic of interest [23]. Table 7 compares the closed-domain IE to the open-domain IE.

Therefore, one of the most crucial steps in extracting information from text is identifying the task to be completed, whether it is a closed-domain IE or an open-domain IE. Several techniques, including NER, RE, EE, rule-based methods (RBM), and ML, can identify patterns and extract the desired information stored within texts.

3) TEXT CORPUS AND KNOWLEDGE GRAPH

A text corpus (or corpus, plural: corpora) is a collection of texts with semantics related to a particular subject [37]. It is used for linguistic rule validation, statistical analysis,

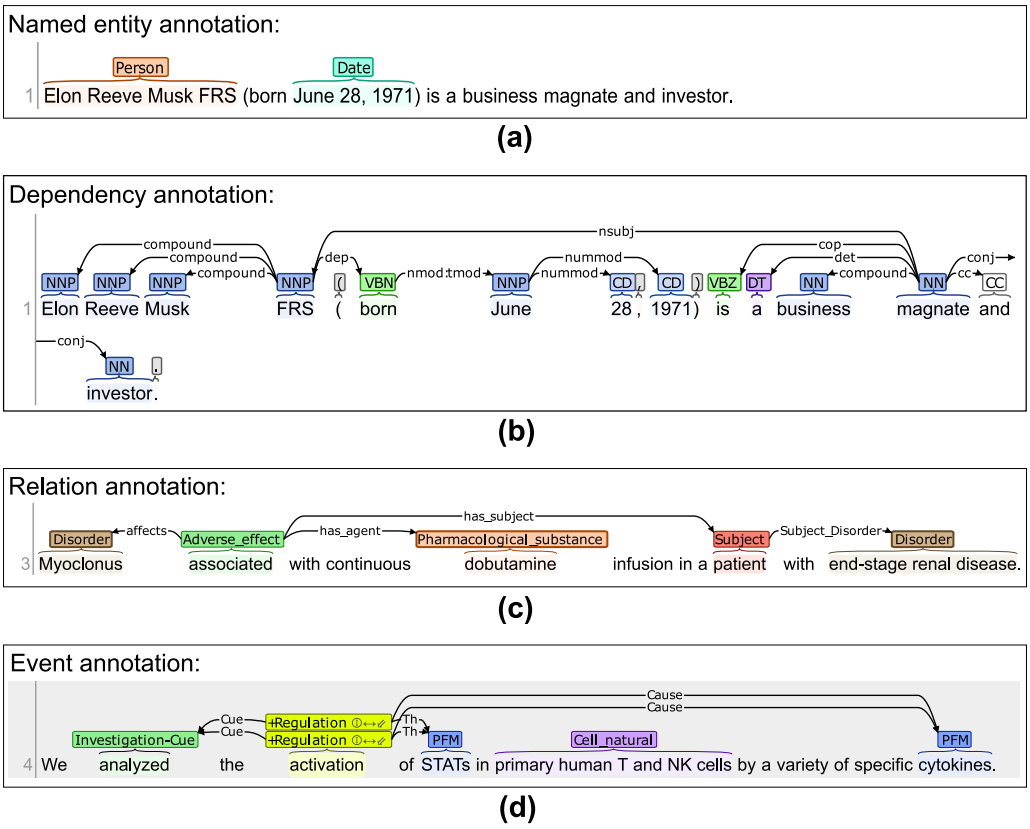

FIGURE 5. Example of corpus annotation on brat rapid annotation tool: (a) named entity annotation, (b) dependency annotation, (c) relation annotation, and (d) event annotation.

checking for occurrence, and hypothesis testing. A corpus is essential for text extraction, as it can serve as training data. An annotated version of the corpus (example shown in Figure 5) can be built from the initial corpus and supplied to an IE system to train the pattern, identify, and extract the correct information. The text corpus can be annotated and displayed using a special text annotation tool, such as the Brat Rapid Annotation Tool [38], which is available from https://brat.nlplab.org/. Manually annotating a corpus can be tedious; hence, researchers have developed automatic corpus extraction (ACE) methods to automate the corpus annotation process [39], [40]. Figure 5 (adapted from the screenshots obtained from the Brat Rapid Annotation Tool website) illustrates the annotated corpora visualized using this tool. A text corpus can also be used to build KG.

A knowledge graph (also known as a semantic graph) is a form of knowledge base that stores information as a graph-structured data model connected by relationships. Although KG is not required for IE from textual data, it can be used to visualize how information (or keywords) are related. It can help researchers understand the relationships between entities and their associated attributes or events [41]. In addition, KG can be used as a visualization and fact-check tool to reveal hidden knowledge in unstructured texts [41]. Furthermore, additional corpora or candidate facts can be used to extend the KG knowledge base. The relationship between a corpus, candidate facts, and KG is depicted in Figure 6.

B. RQ2: WHAT ARE THE CURRENT TECHNIQUES FOR EXTRACTING TEXTUAL DATA FROM UNSTRUCTURED DATA?

Based on our literature search, the current techniques used for extracting textual data are categorized as Named Entity Recognition (NER), Relation Extraction (RE), Event Extraction (EE), hybrid techniques, and other methods (e.g., rule-based, ontology-based, learning-based, DL-based methods). They are selected and applied to projects based on the valuable information that must be extracted from a project. These techniques are closely related to other NLP tasks, including part-of-speech (POS) [42], tagging, and syntactic parsing [43], [44]. The performance of these techniques can be improved or degraded, depending on the strategy used to modify or hybridize them.

IE approaches can be categorized into two different strategies: (i) learning-based methods (LBM) and rule-based methods (RBM) [6], [16], [22]. The LBMs are composed of supervised, semi-supervised, and unsupervised training models. Supervised training allows IE to be conducted based on a sample of labeled data. Semi-supervised training combines a small amount of labeled data with a large amount of unlabeled data during the training of the data model. Contrarily, unsupervised training can be used to train a model by inferring relationships from unlabeled data. To reduce errors in the extraction model, the system automatically modifies the connection weights and learning parameters throughout model training. LBMs have also been deployed as state-of-the-art

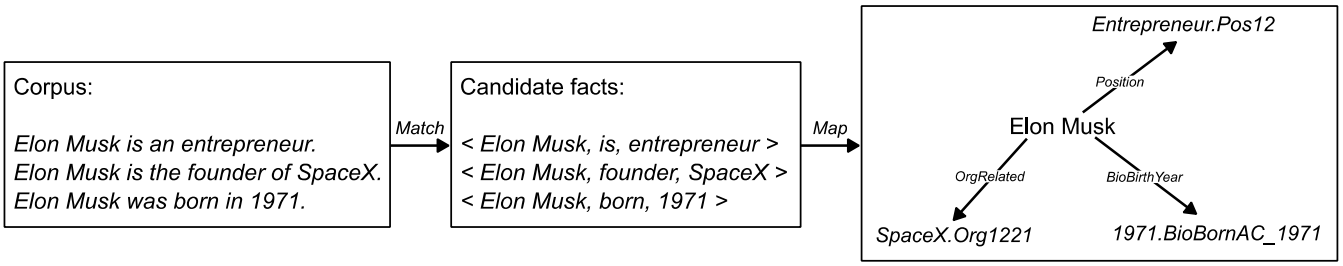

FIGURE 6. Visualization of the corpus (left) and candidate facts (middle) mapped to the knowledge graph (right).

models that feature various DL-based methods for IE from textual data [17], [18], [33], [35], [45], [46], [47].

The most popular IE technique for textual data is the NER technique, which extracts and labels the identified named entities. RE is interested in identifying attributes and relations between text entities. EE aims to detect the existence of an event reported in the text and collect all attributes related to the event. Hybrid techniques [15], [18], [28], [47], [48], [49] have been developed by combining existing techniques with other techniques to strengthen IE strategies based on individual case studies. Other methods include methods that implement rule-based, ontology-based, learning-based, or DL-based methods. Figure 7 summarizes the hierarchy of IE techniques related to textual data.

Regarding data insights, 112 of the 161 included studies were application studies. These application studies have demonstrated different IE techniques, with NER accounting for 45%, EE for 19%, RE for 17%, other methods for 10%, hybrid NER-DL for 5%, and hybrid NER-RE for 4%. Hence, NER is the most commonly applied IE technique for textual data, whereas NER-DL and NER-RE have been the most commonly used hybrid techniques over the past six years. Figure 8 shows the distribution of IE techniques applied in the application studies.

1) NAMED ENTITY RECOGNITION (NER)

NER is an IE technique commonly used for text preprocessing in a variety of NL applications. It is essential for identifying a subject of interest from a text compound [16]. One example of the application of NER is question answering [50]. Based on [22] and [50], the research on NER tasks has been around for more than 20 years since it was proposed at the Sixth Message Understanding Conference (MUC-6) in 1995 [51]. The goal of NER is to recognize and extract descriptive information regarding entity name expressions or ENAMEX (e.g., person, organization, location) and numerical expressions or NUMEX (e.g., time, currency, percentage) from unstructured texts [11], [50], [52]. Figure 9 illustrates the recognition and classification of entities from a text paragraph.

IE from text using the NER technique can be implemented using the RBM, LBM, or hybrid techniques [11]. The RBM utilizes lexicon-semantic patterns and semantic constraints to identify recurring words. By contrast, NER that applies LBM may utilize algorithms such as the Hidden Markov Model (HMM), Conditional Random Field (CRF), Naïve Bayes (NB), Neural Networks, entropy, and Support Vector Machine (SVM) [21]. The performance of NER techniques can be evaluated using accuracy (1), precision (2), recall (3), F1 score (4), and error rate (5) [53], [54], [55]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F}_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})} \tag{4}$$

$$\text{Error rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{5}$$

To clarify the abbreviations used in the above measures, we defined true positive (TP), false positive (FP), true negative (TN), and false negative (FN) using a confusion matrix, as shown in Figure 10. Note that in the case of TP, given the predicted value is 'yes' and the actual value is also 'yes'; while TN predicted 'no', and the actual value is 'no'. On the other hand, FP happens when the predicted value is 'yes', but the actual value is 'no'; and FN is when the predicted value is 'no', but the actual value is 'yes'.

In other words, accuracy is the ratio of correct predictions (TP + TN) to total observations. Precision is the ratio of correctly predicted positive observations (TP) to total predicted positive observations (TP + FP). The recall is the ratio of correctly predicted positive observations (TP) to all observations in the actual class (TP + FN). The F1 score is a weighted average of precision and recall. Finally, the error rate is the ratio of incorrect predictions (FP + FN) to the total number of observations. These metrics assist researchers in evaluating the performance of each algorithm.

In an extensive systematic study by Goyal et al. [22], the authors provided fundamental knowledge and a significant interest in the development and progress of NER research. This study concluded that the performance of NER is mainly affected by language, textual genre/domain, and entity type factors. These factors may pose challenges to NER systems, including difficulty in recognizing nested entities, ambiguity

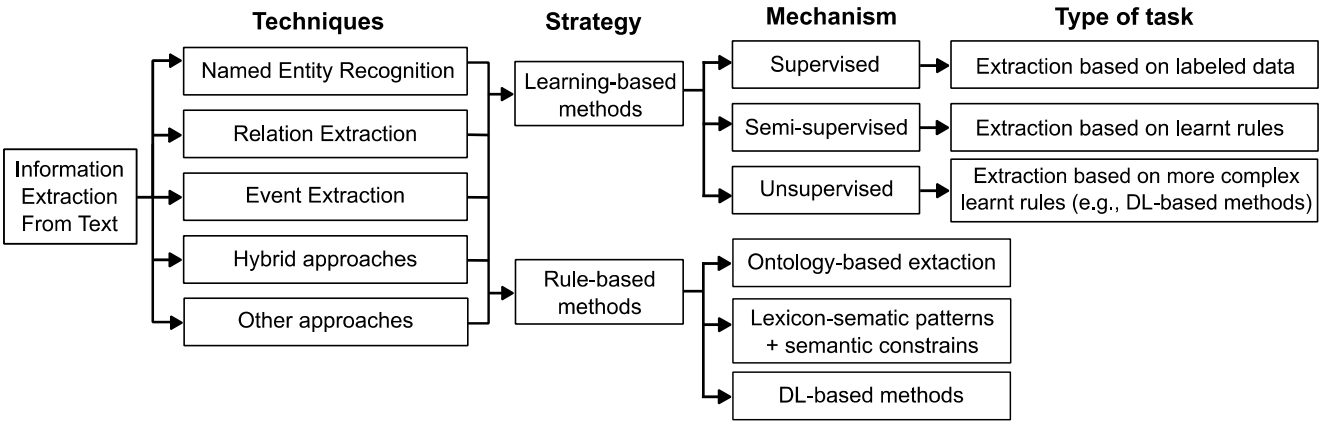

FIGURE 7. IE hierarchies according to techniques, methods, mechanisms of operation, and types of tasks.

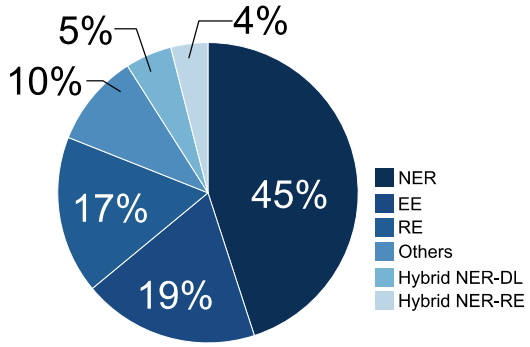

FIGURE 8. Distribution of included application studies based on the IE from textual data techniques.

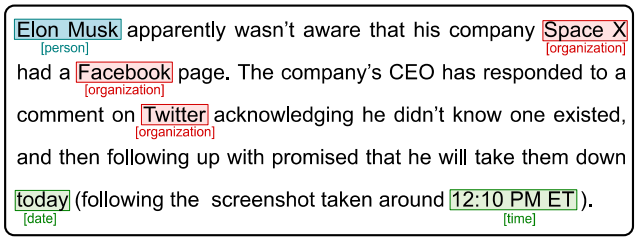

FIGURE 9. Illustration of recognition and classification of named entities using NER from the text.

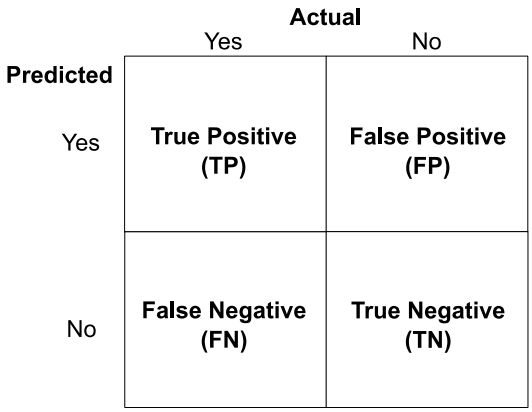

FIGURE 10. Confusion matrix to define TP, FP, TN, and FN.

in the text, the need to annotate training data, and lack of resources (large-annotated corpora and gazetteers). Therefore, these challenges must first be addressed to create more reliable NER systems. Furthermore, the authors explained that the RBM and LBM are used in existing NER systems, including base classifiers.

The research community is becoming increasingly interested in developing innovative approaches for extracting diverse named entities that can be used in various NL applications. There are several variations in NER models, which fall into four types of implementation: (i) NER implemented using the RBM, (ii) NER implemented using the LBM (e.g., supervised, semi-supervised, and unsupervised), (iii) feature-based NER, and (iv) DL-based NER [56] (Figure 7). These NER variations were used to fulfill various requirements and use cases for IE from textual data, including extracting text entities from specific languages and genres.

NER implemented using the RBM utilizes grammatical components (part-of-speech), syntactic (word precedence), orthographic features (such as capitalization), and dictionaries [57]. These elements are subsequently converted into custom rules, which serve as the framework for the NER system. An NER system using a rule-based model depends on handcrafted rules prepared by experts and does not rely on labeled data [56]. Consequently, NER system developers have full control over system identification and classification mechanisms. However, developing sound NER systems using the RBM method requires advanced programming skills and in-depth understanding of the data. In addition, handcrafting system rules is a tremendously exhausting task, especially when dealing with large amounts of data [6], [11], [16], [58].

Therefore, researchers have started to implement the LBM for NER tasks. This method utilizes ML algorithms to automatically identify and extract the named entities from text through supervised, semi-supervised, and unsupervised training [6], [11], [16], [57]. NER models based on unsupervised training do not require labeled data as training samples. However, building such systems requires high-level technical

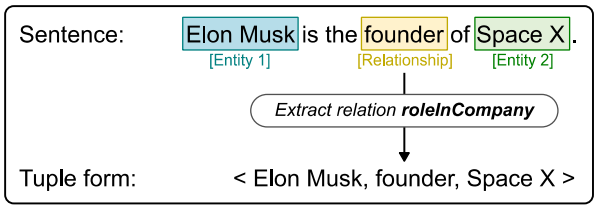

FIGURE 11. Illustration of entities and relationship extraction from text.

skills to optimize the model and achieve good results. Additionally, there are hybrid variations of NER that apply the LBM, such as the hybrid NER-EE implemented for KG construction [59] and extracting valuable knowledge from various documents [60], [61]. Typically, these hybrid models are developed to exploit the learning potential of the LBM, thus enabling them to extract named entities from labeled data accurately.

Next, the feature-based NER model combines the rule engineering of the RBM with the learning capabilities of the LBM to identify specific features from the text. This NER model is designed to identify lexical, morphological, gazetteer lookup, orthographical, POS, contextual, conjunction, and word-embedding features [58], [62]. It results in a model that learns the specific features based on a predefined set of rules. However, the first three types of NER models depend on rule-based engineering, which requires expert knowledge and significant time investment to develop such complex systems. Additionally, these systems cannot deliver good outcomes and exhibit limited robustness when presented with unseen data.

Hence, a DL-based model, the fourth type of NER model, was created to address the issues found in the earlier models. The DL-based model works on character-level, word-level, and sentence-level detection, thus allowing it to detect more complex and nested entities in the text. In addition, the DL-based model can use transfer learning from pre-trained languages and transformers, such as BERT. Henceforth, several BERT variations have been introduced and implemented, including a lite version of BERT (ALBERT), a Robustly Optimized BERT Pretraining Approach (RoBERTa), and distilled BERT (DistilBERT).

2) RELATION EXTRACTION (RE)

The RE task involves extracting attributes and semantic relations for entities in the text [35], [63], [64]. The relations between entities can be represented in a tuple format, such as *<Entity1, Relation, Entity2>*, where *Entity1* and *Entity2* are the identified entities and *Relation* is the description of the relationship between the entities [63]. For instance, given the sentence "Elon Musk is the founder of SpaceX," a relation classifier is used to identify the relation of "roleInCompany". Thus, the sentence can be represented in tuple form as *<Elon Musk, founder, SpaceX>* [65], [66], as shown in Figure 11.

All entity pairings identified from a text body must be listed to determine the crucial relationships to be extracted. A classifier determines pairs that are essential for extraction [67]. Conventional RE systems treat the extraction of relations from the text as two separate tasks in a pipeline, where: (i) the system uses NER to identify the entities present within the text, and (ii) the system extracts the semantic relations between the previously identified entities [67]. It shows that NER is a crucial part of RE. In more advanced RE systems, ACE task is utilized to automatically label a corpus using the supervised learning of relations from a small hand-labeled corpus. The performance of RE techniques is evaluated using precision (2), recall (3), and F_1-measure (4) [63], [67].

RE can be divided into two types: (i) open-domain and (ii) closed-domain. Open-domain RE extracts semantic relations from text from any domain, meaning that the extraction system must be able to build a corpus automatically based on the domain. The ACE task can be used to solve the lack of labeled training data for RE [35]. It allows the task for open-domain RE to have a faster calculation speed and higher accuracy than closed-domain RE while reducing the burden of manually labeling the copra [67]. In addition, several variations of RE implementations have been introduced to address existing issues in RE tasks. These RE implementations feature RBM, LBM, and hybrid approaches for identifying the correct information to extract.

RE systems that use RBM have a simpler design than those that use LBM. However, such systems face difficulties handling the large volume and dimensionality of unstructured data [68]. This problem arises because of the need to prepare a large set of rules for identifying relationships in large amounts of textual data. Therefore, researchers have developed a weakly supervised method that is effective for minimizing manual annotation tasks [35]. Instead, the RE system implemented by the LBM has the advantage of learning complex relations based on the presented data. However, they tend to have a more complex system design, requiring the developer to understand more complex algorithms and be familiar with advanced functions and programming skills to develop an automated RE system. The most recent RE systems utilize a supervised LBM, which requires labeled training data for RE model training [69]. As mentioned earlier, the issue of limited labeled data can be managed using ACE systems [35]. A few examples of RE using the LBM include the CNN-BiLSTM network [43], weakly supervised RE [35], and DNN unsupervised learning [45], [47], [70]. To date, research on RE is still in progress.

3) EVENT EXTRACTION (EE)

The EE mechanism includes identifying reported events in unstructured text and extracting relevant information to a structured form [6], [11], [16]. An event typically consists of two components: (i) event triggers and (ii) arguments [11]. The EE task aims to extract information regarding event triggers (or "triggers") and event arguments (or "arguments") from the text and convert them into a structured form [23].

Triggers are the main words used to identify or express the occurrence of an event in text. Therefore, trigger detection is crucial for EE tasks [71].

Concurrently, arguments are words that specify the participants or the components of a particular event.

The extracted information helps answer the "5W1H" questions (who, when, where, what, why, and how) regarding events that occurred events which come from numerous sources of text The information that is extracted assists in addressing the "5W1H" (who, when, where, what, why, and how) questions regarding events, sourced from multiple text sources [23]. EE tasks can be categorized into two types: (i) open-domain and (ii) closed-domain.

Open-domain EE systems are used to identify and extract information regarding events from various sources and activity domains (e.g., safety/security, justice, finance, biological events, and chemical reactions). Alternatively, closed-domain EE systems deal with an event originating from a specific domain and utilize a predefined event schema for the extraction task. Although such systems can only deal with data from a specific domain, they are much simpler to build and more reliable than open-domain EE systems, since they do not deal with a variety of topics/domains. If an event is detected in the text, the closed-domain EE system attempts to identify the event type, participants, and attributes, based on a predefined event schema. Next, the identified information is extracted and populated into a structured form following the schema provided earlier. The following example is adapted from [23], which visualizes predefined schemas for closed-domain EE tasks, as shown in Figure 12 (a). The mechanism for detecting and extracting relevant information (e.g., event type, trigger, and argument) from a sentence is shown in Figure 12 (b).

EE systems can be trained to extract correct event-related details using the RBM or LBM [6]. The EE system employing the RBM follows a predefined event schema prepared by a domain expert as a set of rules for the extraction task. The advantage of using the RBM for EE tasks is that it allows full control of the mechanisms for event-trigger detection. However, preparing the rules for an extraction system is a tedious and time-consuming task that requires a domain expert for its completion. In contrast, an EE system employing supervised LBM does not require experts to curate the rules or extraction schemas for extraction tasks manually. Nevertheless, such a system requires huge annotated corpora or training data to build its own extraction rules [6], [11].

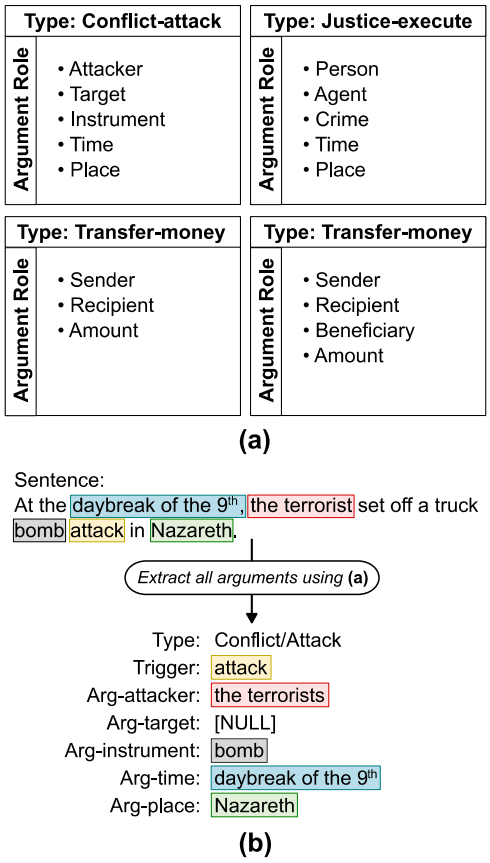

FIGURE 12. Illustration of a closed-domain EE showing: (a) the example of predefined event schemas, (b) the mechanism to extract event-related arguments from a sentence.

C. RQ3: WHAT ARE THE REAL-WORLD APPLICATIONS OF IE FROM TEXTUAL DATA?

The primary purpose of IE techniques for textual data is to identify and extract specific information from text. This study identified the primary methods for extracting information from text, including NER, RE, EE, hybrid techniques, and other methods. For simplicity, we grouped the hybrid technique and other methods (e.g., rule-based, ontology-based, learning-based, or DL-based methods) into a category called "hybrid and other methods". Each of these methods has been created to extract a certain type of target information. For example, NER can be used to identify and extract named entities and types of entities from a text body. RE is used to extract relational information between entities, whereas EE is used to identify and extract event occurrences and their arguments from text [25], [63], [67], [72].

Furthermore, based on the studies gathered, we examined the applications of various IE techniques and categorized them according to the application domain. According to the included studies, each technique has been employed in various domains and applications. Table 8 summarizes the practical applications of the existing text-based IE techniques based on domains and applications.

A thorough examination of the information in Table 8 revealed that IE techniques for textual data had been used in various tasks and application domains. Figure 13 illustrates the frequent uses and frequencies of the present text-based IE techniques.

Most applications target IE from documents from the medical or biomedical domains and documents specific to a language. Several other applications include IE from social media data, IE from (various) documents, IE for extracting corpora from text data, KG construction, IE from news or

TABLE 8. Practical applications of the existing text-based IE techniques.

Tech	Domains	Applications	Studies
NER	IE by specific task	Open IE	[65, 73]
		Corpus extraction	[39, 74-80]
		KG construction	[81]
		News IE/ summarization	[82-85]
		Language-specific task	[46, 56, 62, 73, 86-98]
		Question answering	[99-102]
		Other tasks	[103-105]
	Law and safety	Law enforcement	[49]
		Industrial safety	[44]
		Cybersecurity	[106, 107]
		Crisis/event IE	[48, 85, 108]
	Closed-domain IE by data type	Medical/biomedical	[58, 76, 89, 109-117]
		Social media data	[15, 48, 52, 94, 108, 118-120]
		Chemical documents	[18]
		Scientific documents	[81, 121]
		From other documents	[122]
RE	IE by specific task type	Open IE	[123, 124]
		Corpus extraction	[40, 80]
		KG construction	[125, 126]
		Language-specific task	[124, 127-132]
		IE from documents	[64]
		Other tasks	[133]
	Safety	Safety and security	[127]
	Closed-domain IE by data type	Medical/biomedical	[68, 72, 125, 129, 134-136]
		Economy and finance	[40, 137, 138]
EE	IE by specific task	Open IE	[139]
		Corpus extraction	[40]
		KG construction	[140, 141]
		News summarization	[142, 143]
		Language-specific task	[72, 144]
		IE from documents	[145]
	Safety	Industrial safety	[146, 147]
		Crisis/event IE	[148-150]
	Closed-domain IE by data type	Medical/biomedical	[71, 151-156]
		Social media data	[148, 150, 157-159]
		Economy and finance	[14]
Hybrid and other methods	IE by specific task	KG construction	[41, 59, 160]
		News summarization	[161]
		Language-specific task	[162]
		IE from documents	[60, 61]
		Other tasks	[163]
	Safety	Industrial safety	[17, 164, 165]
	Closed-domain IE by data type	Medical/biomedical	[166-168]
		Social media data	[166, 169]
		Scientific documents	[170]

news summarization, and others, as shown in Figure 13. Based on the top three applications of IE from textual data, we can conclude that IE from text helps resolve problems unique to domains, such as extracting information from medical and biomedical materials, documents written in a particular language, and social media data.

D. RQ4: WHAT IS THE INTENSITY OF PUBLICATIONS RELATED TO IE FROM TEXTUAL DATA?

Analyzing the intensity of recent publications is important for helping researchers identify the trends and future potential of a specific research domain area. Consequently,

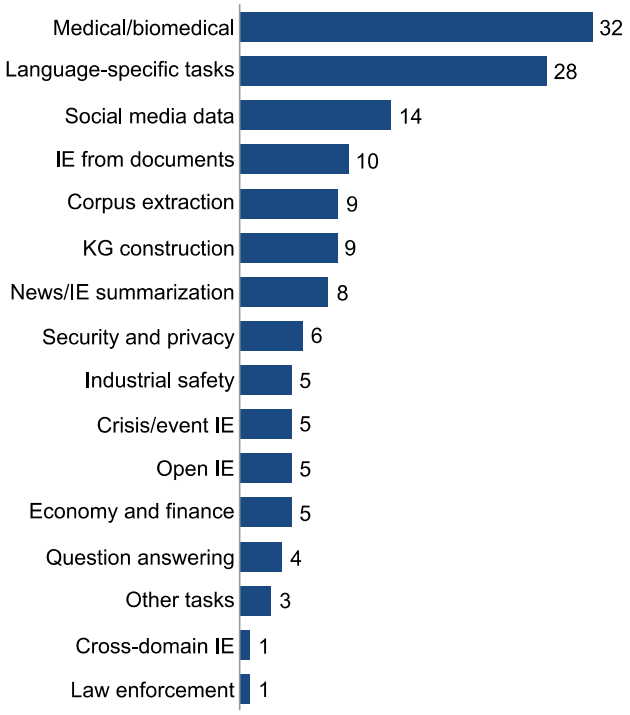

FIGURE 13. The count of included publications according to specific application/task.

we specifically designed RQ4 to highlight the intensity of publications related to IE from textual data. This allowed us to investigate the potential growth of IE techniques for textual data.

Careful and thorough analysis revealed that the domain of IE from textual data is still on the trend as the number of studies has increased over the last five years (2017–2021). However, the number of studies identified and included in 2022 is lower than that in 2021 because more studies in the current year are still in the publication process and were not yet available online when this literature search was conducted. We believe that the number of studies for 2022 will rise and cross the trendline before the end of the year based on the rise in identified studies from 2020 to 2021, which suggests that this topic is currently on the rise. Figure 14 (a) shows the overall trend in the number of identified studies, whereas Figure 14 (b) shows the publication trend of the included studies for the last six years within this domain area.

Next, the distribution of the included studies was further examined based on the search directory so that researchers could prioritize their search and discover more studies in this domain area. The results showed that in the last six years of publications in this domain, a large proportion of the included publications were from IEEE Xplore (62%), followed by ScienceDirect (20%), SpringerLink (8%), Scopus (4%), ACM Digital Library (4%), and other directories (2%). Figure 15 shows the distribution of the included studies based on search directories.

Following this, we labeled the included studies based on the search directories used and study type. According to

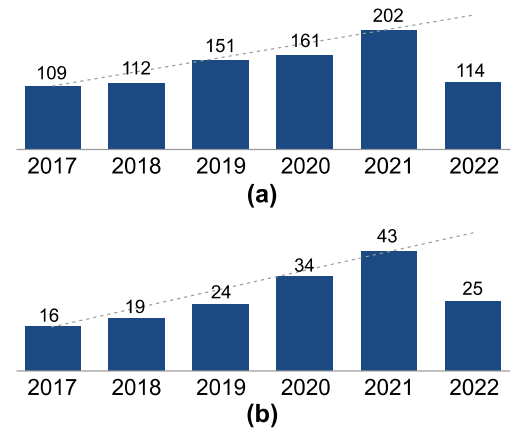

FIGURE 14. Illustration of the trend of publication counts in the recent 6 years for the: (a) count of the identified studies by publication year and (b) count of the included studies by publication year.

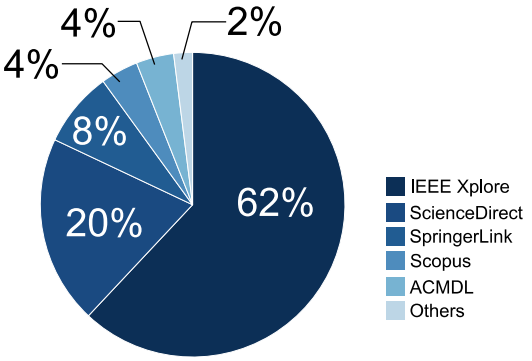

FIGURE 15. The distribution of included studies related to IE from textual data by directories.

the findings, IEEE Xplore had the most publications and the greatest variety of IE approaches covered among the five main directories and additional sources (studies listed in the reference list), followed by ScienceDirect, SpringerLink, Scopus, ACM Digital Library, and other directories. The IEEE Xplore directory contains the highest number of relevant studies compared to the other directories. IEEE Xplore and ScienceDirect cover the broadest range of IE techniques for textual data, with more than six techniques available. Therefore, when looking for content related to this domain, researchers should focus on IEEE Xplore and ScienceDirect to identify relevant materials in this domain study. The frequency of publications mentioning specific IE techniques in each directory is shown in Figure 16.

It is important to analyze the included studies according to study type to identify the most popular study designs among researchers. Based on the distribution of the included studies, most publications were focused on the application (69%), followed by research (13%), survey (6%), review (6%), and SLR studies (6%). Application studies demonstrate IE techniques for extracting information from textual data. On the other hand, research studies offer individual or group researchers an attempt to develop the latest advancements for enhancing current techniques using various methods or frameworks. Next, survey studies summarize the progress of a theory, concept, or technique from the beginning until the most recent methods. Review studies comprehensively assess and describe high-quality studies in certain domains. Finally, SLR studies gather and assess eligible studies using a systematic approach. Figure 17 shows the distribution of included publications by study type.

The earlier numbers indicate that most of the included studies were primarily concerned with applying IE techniques to their data. Following this, 13% focused on creating new techniques or improving existing IE techniques, 6% presented the results of a comprehensive survey on this topic, and 6% provided a review of high-quality studies in this area. The remaining 6% of studies were systematic reviews.

Thus, we believe that further research should be conducted, particularly to address the concerns affecting the performance of current IE approaches, as this would help other researchers to improve current research breakthroughs. In addition, comprehensive SLR studies are required to capture the most recent breakthroughs in this field from various perspectives.

V. DISCUSSION

Throughout this SLR process, we gathered sufficient data to elaborate on the fundamentals of IE from textual data. We explained the current techniques used for this purpose, practical applications, and the intensity and trend of publications in the last six years. Therefore, Section V-A analyzes the current challenges for each RQ, which must first be understood to improve IE pipeline processes. Section V-B provides recommendations for future research.

A. CURRENT CHALLENGES FOR EACH RQ

1) CHALLENGES BASED ON RQ1: DIFFICULTY IN UNDERSTANDING FUNDAMENTAL CONCEPTS AND STEEP LEARNING CURVE

The first challenge in this area of study is the steep learning curve. New researchers need to spend a significant amount of time understanding several terms and concepts related to IE from textual data. These concepts cover a wide range of knowledge that may be difficult to grasp and include natural language, natural language processing, closed-domain IE, open-domain IE, text corpora, and KG.

For example, NL and NLP fall under the umbrella of natural language processing research, where studies are conducted to understand how machines process natural languages. Although exploring the domain study of NLP is time consuming, grasping such concepts may provide new insights into comprehension and success in this challenging research area. Researchers can also learn and explore the benefits of NLP for extracting useful information from textual data by consulting existing studies and domain experts. This will eventually help us understand other concepts in this domain, such as text mining, data mining, and semantics.

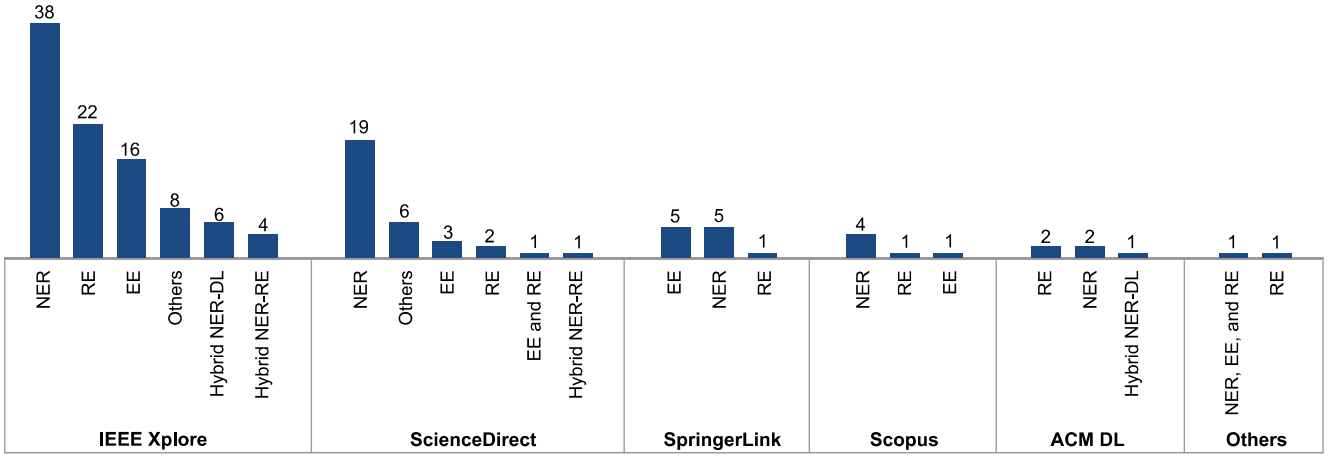

FIGURE 16. The frequency of publications mentioning specific IE techniques in each directory.

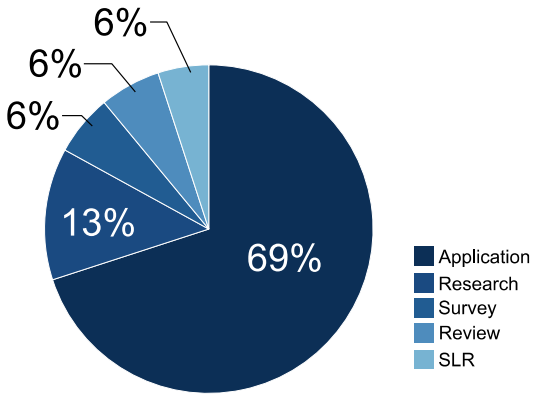

FIGURE 17. The distribution of included publications by study type.

Second, new researchers may find it challenging to select a suitable IE model that is effective for their data. This problem is common in data science. However, this becomes more distressing, mainly when dealing with the complexity of textual data originating from various sources. IE models for textual data must be able to cater to two distinct tasks: (i) closed-domain extraction and (ii) open-domain extraction. Although both tasks focused on extracting meaningful information from text, the closed-domain IE task focused more on extracting information from a particular domain (e.g., biological, chemical, or security), thus requiring an extraction schema for the system to follow.

In contrast, open-domain IE tasks are designed to identify the topic of interest from the text and thus do not require such a schema. Developing a successful extraction model may be difficult for researchers with less coding experience as they must first decide which IE task to perform and then set up their IE models following the relevant information to be extracted. In addition, building reliable systems for managing IE from language-specific documents and evaluating interactions of unstructured data requires considerable expertise.

We believe that the explanation provided in this study is sufficient for many readers to begin research in this domain. In conclusion, although this study covered the essential topics needed to comprehend this field, researchers are urged to conduct additional research, particularly by examining the cited studies to acquire a more in-depth explanation. As the body of knowledge surrounding IE from textual data grows, researchers are encouraged to remain current in recent advances and research in this area.

2) CHALLENGES BASED ON RQ2: ISSUES IN THE EXISTING IE TECHNIQUES

In the second RQ, we evaluated current techniques for extracting information from texts based on studies published in the last six years. According to the information gathered, IE from the text can be classified as NER, EE, RE, hybrid techniques, and other methods. Although researchers have collectively improved these techniques based on the essential concepts of IE from textual data, several issues remain that require other researchers to explore solutions. Table 9 presents our findings regarding the issues and challenges researchers face for each IE technique for textual data adapted from [11].

The NER technique is reliable at fulfilling its purpose of recognizing named entities when best practices are applied. However, as summarized in Table 9, NER also faces many challenging issues that affect its performance in recognizing and classifying named entities. The table categorizes these issues as being associated with data, entities, domains, tasks, and languages. Several studies have indicated that issues related to data and language are the most significant challenges that must be addressed to ensure the robustness of NER systems. The most noticeable factors influencing NER performance are IE from multilingual text bodies and IE from languages with poor morphology or low resource levels, which affect entity detection tasks in NER.

Alternatively, RE techniques face challenges related to the data, language, relationship identification, and technical issues. In RE, data-related issues such as sparsity, dimensionality, volume, and a lack of training data are more prominent. In addition, RE encounters issues with languages as each language has a different structure and ontology,

TABLE 9. Issues and challenges of the existing text-based IE techniques.

Tech	Issues	Factors Influencing the Issues	Studies
NER	Data-related	Noise (e.g., homonyms)	[6, 11, 21, 53, 62]
		Missing data	[6, 11, 171]
		Data diversity	[6, 11, 21, 85]
		Nested entities	[6, 7, 22]
		Variation in perspective	[6, 7, 21]
		Non-uniform writing/ narrative style	[21, 53, 120]
		Lack of training data	[21, 22, 25, 102, 114, 115, 171]
		Lack of resources (e.g., copra, gazetteers)	[6, 21, 22, 50, 54, 102, 172]
	Entity-related	Entities ambiguity	[6, 21, 22, 60, 114]
		Automatic labeling	[11, 22, 26]
		Semantics of NE	[11, 26, 114]
		Unseen mention of entities	[117]
	Domain-related	Domain-specific entities	[50, 54, 114, 171]
		Novel morphology patterns	[117]
	Task-related	Selection of NER technique	[26, 115]
		Generalize and define NER features	[172]
	Language	Specific language	[6, 21, 50, 56, 95, 98, 171, 173]
		Different languages	[50, 171, 173, 174]
		Language morphology	[6, 21, 50, 56, 96, 174]
		Out-of-vocabulary issue	[55, 60, 98]
		Lack of training data	[35, 50]
		Lack of capitalization	[21, 96, 174]
RE	Data-related	Noise in data	[69, 175]
		Data sparsity	[6, 11, 171]
		Data dimensionality	[6, 24, 66, 67]
		Volume	[6, 24, 25]
		Lack of training data	[24, 28, 66, 69, 171]
	Language	Language ambiguity	[25, 171]
		Lack of multilingual IE	[171]
	Identify relations	Domain-specific relations	[24, 66]
		Relationship ambiguity	[24, 66]
		Errors in constituent parsing	[66]
	Technical	Require large, annotated copra	[6, 24, 25, 28, 67, 175]
EE	Data-related	Noise	[6, 11, 158]
		Data sparsity	[71, 171]
		Data dimensionality	[6, 11, 159]
		Data sources diversity	[6, 28]
		Variation in perspective	[6, 11, 159]
		Nested/overlapping events	[155, 176]
		Representation ambiguity	[6, 158, 159, 176]
		Lack of training data	[28, 71]
	Language	Language ambiguity	[159, 171]
	Technical	Semantic event modeling	[11]
		Limitation of ML and RBM techniques	[11, 28]
Hybrid and other methods	Data-related	Data sparsity	[171]
		Lack of training data	[28, 171, 177]
	Domain-related	Domain-specific	[171, 177, 178]
	Language	Language-specific	[160, 171, 178]
		One method, different languages	[160, 179]
	Technical	Requires ontological taxonomy	[180]
		Lack of benchmark results	[179]
		Data annotation task	[168]
		Requires large, annotated copra	[28, 177, 179]

causing extracting relations from the text to be challenging. Another issue is that RE requires a large, annotated copra to process complex sentences and paragraphs, as the system needs to build many-to-many nested relations instead of simple relations.

Issues associated with EE tasks are primarily related to data, language, and technical issues. The most cited issues in EE tasks are related to data, such as noise, data sparsity, data dimensionality, data source diversity, variation in perspective, nested or overlapping events within a text, ambiguity in representation, and lack of training data. Therefore, focusing on solving such issues will improve the performance of EE models on par with other models such as NER or RE.

Finally, hybrid DL-based approaches (e.g., models based on DNN, CNN, and transformers) and other methods (e.g., RBM and ontology-based methods) encounter performance limitations in terms of data, domain, language, and technological issues. Furthermore, it is difficult for researchers to determine which model best suits their data. Researchers are still working on finding solutions to the factors affecting these issues.

3) CHALLENGES BASED ON RQ3: ISSUES RELATED TO FUTURE APPLICATIONS OF IE FROM TEXTUAL DATA

RQ3 investigated the most popular application of IE techniques for textual data. In this study, we discuss the applications of NER, EE, RE, hybrid techniques, and other methods for textual data. All the aforementioned IE techniques for textual data have immediate real-life applications. Therefore, this subsection provides information on the difficulties in adapting these strategies to real-world requirements and challenges. In the era we live in today, data are accumulated at a very high rate from various sources and gathered in huge volumes [181]. It is known as the big data problem, in which data can originate from various sources, formats, and languages.

Based on our findings, recent studies have focused on performing NER tasks on documents of specific languages. NER model struggles to extract information from documents written in multiple languages or documents in languages other than the main preset language. Thus, each NER extraction model is limited by the language it is trained to extract [6], [11]. A few examples include NER in Portuguese texts [46], [91], Chinese texts [72], [76], [89], [127], [128], [129], Indonesian texts [65], [78], [87], [119], [130], [143], Malay texts [92], and Arabic texts [21]. This demonstrates that language is a common barrier to extracting useful information from text documents.

Second, there are challenges in terms of data variety. IE techniques are used in the real world to extract information from various documents, including electronic medical records, scientific documents with chemical names and processes, law enforcement documents, safety and security documents, and financial documents. From a closer perspective, these documents share a common characteristic;

they are prepared by humans, not by sensors or machines. It is worth noting that information prepared by humans can often be much more difficult to extract and organize than information prepared by machines (such as sensor data and server logs) due to their ''inconsistent'' nature. Human-prepared documents can be written using various languages, perspectives, and writing or narrative styles. These inconsistencies can arise in structures, formats, encodings, writing or narrative styles, languages, terminologies, and measurement units. Additionally, these documents may come from various domains, such as healthcare, science, chemistry, law, safety, and finance. Consequently, creating annotated corpora or labeled data for these human-generated documents is highly resource-intensive and time-consuming as it requires many rules to be generated for IE task purposes [16], [34], [171].

Third, there are challenges of algorithm errors or limitations, particularly when extracting complex information from text. For example, Ramponi et al. [137] have successfully extracted biological information from text accurately. However, their method does not work well for complex relations, owing to algorithm errors. Because of this issue, they proposed two possible solutions: (i) relying on the RBM instead of using an ML method to have a higher degree of control over the system's behavior and (ii) enriching the system with improved representation. In another study, Luo et al. [166] reported that their algorithm required a large data sample to produce highly accurate results. Given the small size of the corpora, the algorithm cannot create a good model for correctly extracting information.

To summarize, IE techniques have been successfully implemented in various tasks in the real world, including the extraction of cause-effect relationships [24], event detection and extraction [139], [144], corpus extraction from text [39], [40], [75], KG construction, question answering [36], [99], and mining social media data [148], [150], [157]. The mechanisms of IE from textual data are continually developing as more research is being conducted to address the existing problems within each technique. Researchers still have many grounds to cover to ensure that textual data IE can be fully automated or, at the very least, semi-automated, thus removing the hassle of manually performing data preparation, which is highly time-consuming. In the future, DNN/DL models in IE from textual data have the potential to increase significantly, owing to their capacity that allow machines to learn and extract information from text accurately. Advancements in IE techniques have provided a variety of applications for solving complex real-world problems, such as identifying false news and rumors, assessing risks, and producing useful insights from news or reports.

4) CHALLENGES BASED ON RQ4: RESEARCH TRENDS RELATED TO IE FROM TEXTUAL DATA

To answer RQ4, we thoroughly examined the intensity of publications on IE from textual data within the last six years. Based on the visualizations curated from the selected studies, we believe that research in this domain is still trending (Figure 14). However, compared to other AI domains, overall research on IE from textual data is still very limited. Moreover, further in-depth analysis on Figures 8 and 16 reveals that researchers mainly focused on the NER approach compared to RE, EE, hybrid DL-based methods, and other approaches. Therefore, it is difficult to explore such approaches with limited resources.

Large corporations such as Amazon, Microsoft, Google, Facebook, and Twitter generate and host massive amounts of user-generated unstructured data on their servers. Such data are typically saved and processed to create a generalized model of how customers think and behave collectively. Businesses have been unable to fully utilize this information as textual data cannot be processed similarly to numerical data. Extracting and analyzing such data will give businesses a competitive edge, as it reveals knowledge regarding human behavior, individual or group purchasing patterns, suggestions for decision-making, human lifestyles, and many others. Researchers are encouraged to continue advancing the field of IE from textual data as this field of study is now more crucial than ever.

From a business management perspective, extracting and utilizing this information enables organizational leaders to make informed decisions and empower their decisions and directions based on data. However, the overall research in this domain area seems very limited, especially when exploring models that employ hybrid DL-based methods. Therefore, IE techniques for textual data are currently very important and there has been a constant push for further research in this domain.

B. FUTURE RESEARCH DIRECTIONS

In this study, we thoroughly evaluated existing studies in the field of IE from textual data from several perspectives, including the most recent methods, applications, and research trends, while also detailing the challenges in this field. Moving forward, it is important to highlight potential future research directions identified based on each RQ. Hence, this section presents recommendations for future research based on the challenges presented in Section V-A.

1) FUTURE RESEARCH DIRECTIONS BASED ON RQ1

The current challenges related to RQ1 are mainly based on two issues: (i) the difficulty in understanding the fundamental concepts related to IE from textual data, and (ii) the difficulty in identifying the proper IE model for their data.

To address the first issue, this study provides a thorough and straightforward explanation of several concepts related to the study area (Section IV-A). Nonetheless, researchers are encouraged to conduct more studies and gather supporting information to solidify their understanding of the underlying concepts of IE from textual data. This will help researchers to develop a thorough understanding of related concepts from various perspectives.

For the second issue, researchers may find it challenging to select a suitable IE model that will be effective for their

data or corpora, given the various IE models that exist today and the increasing popularity of DL models in recent years. It is a reasonably daunting task to explore each model. Hence, we suggest that future studies generalize the most common models (e.g., NER, RE, EE, or DL-based models) and make them available to researchers by creating test kits in the MATLAB toolbox or Python packages to facilitate easier deployment. It will assist novice researchers in understanding the operating mechanism of the model, as well as in developing and running simulations, thereby encouraging improvements to existing models. Moreover, it allows the domain to witness innovative ideas to solve the problem of accurately extracting relevant information from textual data.

2) FUTURE RESEARCH DIRECTIONS BASED ON RQ2

Based on the challenges identified from RQ2, we discussed the challenges faced by NER, RE, EE, hybrid techniques, and other methods. From this, we discovered that the most common issues faced by these techniques are related to: (i) data, (ii) language, and (iii) technical issues.

Regarding the first issue, it is important to note that the performance of an IE system may be affected by the data itself (or corpora), which may contain noise, missing data, high diversity, variance in perspective, lack of training data, and volume issues (too large or too small). Therefore, future studies should focus on developing high-quality annotated data/corpora and making them publicly and freely available to researchers for use as training datasets. The annotated corpora must have the following characteristics: decreased noise, no missing data, a clear perspective, and sufficient volume. To assist in the development and ensure that the annotated corpora are prepared correctly, a data/corpus annotation guideline and manual should be designed, as in [182], [183], [184], [185], and [186]. We believe that having high-quality and publicly downloadable corpora (either annotated or unannotated) will benefit researchers interested in studying and improving the existing IE methods for textual data.

Concerning the second issue, current IE techniques still struggle to process user-generated data owing to language-related issues caused by language morphology, out-of-vocabulary texts, lack of training data for specific languages, and lack of capitalization in some languages. Consequently, most of the existing IE techniques are limited to the language in which they are trained. Therefore, we recommend that researchers investigate appropriate strategies, such as incorporating DNN/DL or a language-detection module into the existing IE models. It ensures that future IE models can be trained and processed using data or corpora containing mixed or multiple languages. We believe such progress is significant because organizations' documents may be written in multiple languages, and future IE models should be able to extract relevant information from these documents.

Third, existing IE models also face technical issues, such as requiring large-annotated copra, limitations of ML and RBM techniques, and a lack of benchmark results. Thus, we recommend that future studies hybridize or combine their techniques with DNN/DL, depending on the task or problem to be solved. This offers the advantage of learning hidden patterns within textual data and reasoning for the correct information to be extracted. In addition, our findings show that current state-of-the-art models employ DL to extract information from textual data. Although it may be difficult to implement a hybrid DL-based model, we believe that doing so will help to address this issue, and the results can be used as a benchmark for other studies.

3) FUTURE RESEARCH DIRECTIONS BASED ON RQ3

The challenges for RQ3 include: (i) issues related to future applications of IE from textual data, (ii) data variety issues, and (iii) limitations or issues with the existing IE algorithms.

We discovered that the most concerning issue with IE from textual data is that current models struggle to extract information from documents in multiple (or mixed) languages or documents in languages other than the main preset language. Researchers can use hybrid or DL-based IE models to investigate and propose methods for extracting useful information from documents containing multiple languages or documents of different languages (i.e., cross-language IE). This progress may still be far from implementation based on the current IE model development. Thus, pioneers need to move in this direction before this can be made possible.

Next, regarding the data variety issue, researchers may concentrate on developing a DL or DNN-based IE model, as it is a state-of-the-art model and has shown a promising outcome for generating better results than other conventional methods. Due to the complexity of human-generated textual data, two additional recommendations can be taken into account: (i) focus more on the LBM because it utilizes fewer resources, eases the burden of creating corpora/training data, and automatically generates rules based on the training model; and (ii) focus on a specific topic or theme by performing a closed-domain IE task instead of an open-domain IE task.

Regarding the third issue, regarding the limitations of existing algorithms, we recommend that researchers select a specific algorithm and address it by either creating a specific rule base for a higher degree of control or enriching the system with improved representation (additional knowledge). Researchers can also investigate suitable cross-validation and optimization techniques to improve the accuracy of the IE model after it has been trained, fine-tuned, and validated.

4) FUTURE RESEARCH DIRECTIONS BASED ON RQ4

Based on the challenges identified in RQ4, current studies have primarily focused on NER, EE, and RE, followed by hybrid techniques and other methods (Figures 8 and 16). We recommend that researchers consider using DNN/DL-based and hybrid approaches in future research. Following our findings regarding the application of IE from textual data (Figure 13), more research has focused on medical/biomedical and language-specific tasks as opposed to

other applications. This is because many corpora are available in the medical and biomedical fields, and most studies have concentrated solely on data extraction from a certain language. Hence, we recommend that future studies focus on other applications, such as security, legal, social media IE, and news IE. As shown in Figure 17, most of the included studies were application studies (69%), followed by research, surveys, reviews, and SLR studies. Following the preceding statement, we can see that despite the large number of studies presented in this domain, most focus on application studies. Hence, we recommend that researchers concentrate on improving the performance of existing IE techniques and address the issues explained in Section V-A. Finally, Figure 14 shows that the trend in this research domain is still ongoing. Still, we believe that more research is needed in this domain area, particularly to improve the accuracy and effectiveness of the existing IE techniques.

It is an open question from our perspective as to how long this domain will remain in the trend. Despite this, we believe that this research domain will continue to thrive since many challenges remain unsolved. We believe the suggestions will encourage researchers and developers to concentrate more on enhancing the accuracy and effectiveness of the current IE techniques and methodologies. Finally, this will lead to the development of an industry grade IE system for widespread use in the future.

VI. STRENGTHS AND LIMITATIONS OF THE STUDY

Section II summarized 12 survey and systematic review studies published within the domain of IE from textual data. It is important to remember that these studies focused on various aspects, and thus covered a narrow range of IE techniques used for textual data. Although studies in the current literature help us grasp the fundamentals of IE, they do not provide a comprehensive and updated analysis that focuses on the current approaches, applications, trends, and challenges of various IE techniques for textual data. Furthermore, existing studies in the literature were published a few years ago.

Therefore, this SLR compiles all studies relevant to IE from textual data in one place, presenting accurate latest figures and comprehensive materials for novices and experienced researchers. Only high-quality and relevant materials were included in this manuscript to ensure the credibility of the findings, which were achieved by adhering to SLR preparation guidelines [31], [32] and PRISMA criteria [32]. The outcome of this SLR study presents the most recent methods and highlights their applications, as well as the current research trends regarding IE from textual data with the latest facts and figures. It also describes the challenges while emphasizing suggestions for future research in this domain. To the best of our knowledge, this is the first initiative in this domain to present an SLR that focuses on various IE techniques from textual data while paying particular attention to each technique. It establishes a paradigm for future studies by summarizing the current methods, applications, research trends, and challenges of IE from textual data.

Despite our best efforts, there are certain unavoidable constraints to recognize when reporting our findings. First, this study only considered studies published in English as its source. Hence, some significant studies may not have been included due to language barriers. Second, we excluded studies with a poor methodology to maintain the quality of our sources. Third, since the scope of this study is IE from textual data, any studies that present methods for extracting text from multimedia files such as images (e.g., scanned documents, text extraction from images or handwritten documents, or using Optical Character Recognition, OCR) or videos (e.g., text extraction from subtitles, text from video clips, or text script extraction from lip-reading) are excluded. Filtering our sources based on these criteria may have influenced the final findings. However, the decision was made based on the inclusion and exclusion criteria (Table 3) and the eligibility assessment criteria (Table 4).

VII. CONCLUSION

This systematic literature study aims to present the most recent approaches to IE from textual data, their applications, and current research trends, while outlining the challenges and potential future directions for this field of study. To ensure that the manuscript complies with accepted reporting standards, the authors adhered to the systematic literature mapping procedure described in the SLR preparation guidelines [31], [32] and PRISMA standards [32]. In the literature search, 849 studies were retrieved from the five directories and authors' reference lists. After refining our list of studies using the literature study mapping process (Figure 1), 161 studies were obtained and considered for data analysis. These 161 studies met the inclusion and exclusion criteria (Table 3) and the eligibility assessment criteria (Table 4), proving them to be high-quality reference materials.

The included studies were dissected and synthesized to answer all the RQs presented at the beginning of this systematic study. Based on the collected evidence, the authors presented several fundamental terms and concepts regarding IE from textual data (Section IV-A) and current techniques widely used to extract information from textual data (Section IV-B). Next, the authors highlighted the practical applications of each IE from the textual data technique, where each is categorized by domain area (Section IV-C). The authors also presented the intensity of publications related to IE from textual data with visualized and included key highlights (Section IV-D). Finally, we discussed the current challenges and provided recommendations for future research to address the limitations found in the literature based on each RQs (Sections V-A and V-B, respectively).

Our findings support the notion that IE techniques for textual data can be divided into five categories: NER, RE, EE, hybrid techniques, and other methods (e.g., rule-based, ontology-based, learning-based, or DL-based methods). In light of the current issues in the existing IE techniques, we found that language-related issues are the most common barriers to extracting useful textual information. The second

TABLE 10. PRISMA 2020 checklist.

Section	Num	Location
Title	1	Page 1: Title
Abstract	2	Page 1: Abstract
Introduction	3	Page 3: Section I
	4	Page 3: Section I
Methods	5	Page 7: Section III-B
	6	Page 6-7: Section III-A
	7	Page 6-9: Section III-A until Section III-D
	8-15	Not applicable
Results	16a	Page 7: Section III-A
	16b	Not applicable
	17	Page 9-20: Section IV and Section V
	18-22	Not applicable
Discussion	23a	Page 9-18: Section IV
	23b	Page 18-21: Section V-A
	23c	Page 22: Section VI
	23d	Page 21-22: Section V-B
Other	24	Not applicable
Information	25	Page 1
	26-27	Not applicable

major concern is related to data, where textual data may originate from various sources, exist in various structures, or lack consistency. Third, technical issues may include algorithm errors and other factors contributing to incorrect IE processing. Other challenges related to IE from textual data are summarized in Table 9 and explained in Section V-A. A few survey studies [24], [25] have also concluded that DL-based and hybrid models are currently state-of-the-art for extracting information from text.

This literature review, therefore, aims not only to provide the most comprehensive understanding of IE from textual data and the status of research in this domain area but also to assist both novice and experienced researchers in determining the future research directions of this domain. We genuinely hope that this study will serve as a central point of reference for novice and experienced researchers to grasp fundamental knowledge regarding IE from textual data. Finally, the authors hope to inspire enthusiastic researchers to dedicate their efforts to improving the existing IE approach and investigating areas where IE from text has not been used previously. This will allow us to develop skills and broaden our knowledge.

APPENDIX A

All the literature search databases employed different search schemes. To produce equivalent search queries across all databases, we either utilized operators and wildcards to create advanced search queries, or used the advanced search form provided on each database's website. Advanced search guidelines are referred to and available from the following links, accessible on the latest search date of July 31, 2022.

1) IEEE Xplore: https://ieeexplore.ieee.org/Xplorehelp/searching-ieee-xplore/advanced-search
2) ScienceDirect: https://service.elsevier.com/app/answers/ detail/a_id/25974/
3) SpringerLink: https://link.springer.com/searchhelp
4) Scopus: https://service.elsevier.com/app/answers/detail/ a_id/11213/ (document searching) and https://service.elsevier.com/app/answers/detail/a_id/11365/ (advanced searching)
5) ACM Digital Library: https://dl.acm.org/search/advanced

Special note for ScienceDirect advanced search: The query engine automatically includes keyword results with plurals and spelling variants (e.g., ''entity'' includes ''entities'' while ''extract'' includes ''extracts'', ''extraction'', and ''extracting'').

APPENDIX B

The PRISMA 2020 Checklist template is adapted from: http://prisma-statement.org/documents/PRISMA_2020_checklist.pdf.

ABBREVIATIONS

All abbreviations used in this manuscript are defined below.

Abbreviation	Definition
ACE	Automatic corpus extraction
AI	Artificial intelligence
ALBERT	A lite BERT for supervised learning
BBMC	BERT-BiLSTM-MHATT-CRF
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional long short-term memory
CNN	Convolutional neural network
CoNLL	Conference on Computational Natural Language Learning
CRF	Conditional Random Field
DistilBERT	A distilled version of BERT
DL	Deep learning
DNN	Deep neural network
DOI	Digital object identifier
EE	Event extraction
ENAMEX	Entities such (e.g., person, organization, location)
HMM	Hidden Markov Model
IE	Information extraction
KG	Knowledge graph
LBM	Learning-based method
MHATT	Multi-head attention
ML	Machine learning
MUC-6	The Sixth Message Understanding Conference
NB	Naïve Bayes
NER	Named entity recognition
NL	Natural language
NLP	Natural language processing
NUMEX	Numerical expressions (e.g., time, currency, date)
OCR	Optical character recognition
POS	Part-of-speech
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RBM	Rule-based method
RE	Relation extraction
RoBERTa	Robustly Optimized BERT Pretraining Approach
RQ	Research question
SLR	Systematic Literature Review
SVM	Support vector machine

ACKNOWLEDGMENT

The authors would like to thank the reviewers for the valuable feedback and suggestions.

REFERENCES

[1] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.001.

[2] Ž. Krstić, S. Seljan, and J. Zoroja, "Visualization of big data text analytics in financial industry: A case study of topic extraction for Italian banks," *ENTRENOVA-ENTerprise Res. Innov.*, vol. 5, no. 1, pp. 35–43, 2019.

[3] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, "Text mining of industry 4.0 job advertisements," *Int. J. Inf. Manage.*, vol. 50, pp. 416–431, Feb. 2020.

[4] M. Selimi and A. Besimi, "A proposed model for stock price prediction based on financial news," *ENTRENOVA-ENTerprise Res. Innov.*, vol. 5, no. 1, pp. 68–75, 2019.

[5] A. Kadriu, L. Abazi, and H. Abazi, "Albanian text classification: Bag of words model and word analogies," *Bus. Syst. Res. J.*, vol. 10, no. 1, pp. 74–87, Apr. 2019.

[6] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *J. Big Data*, vol. 6, no. 1, p. 91, 2019, doi: 10.1186/s40537-019-0254-8.

[7] P. S. Jadhav, S. S. Bodhe, G. M. Borkar, and A. V. Vidhate, "Unstructured big data information extraction techniques survey: Privacy preservation perspective," in *Proc. Int. Conf. Electr., Comput., Commun. Mechatronics Eng. (ICECCME)*, Oct. 2021, pp. 1–6.

[8] M. H. A. Abdullah, N. Aziz, S. J. Abdulkadir, E. A. P. Akhir, and N. Talpur, "Event detection and information extraction strategies from text: A preliminary study using GENIA corpus," in *Proc. 2nd Int. Conf. Emerg. Technol. Intell. Syst.*, 2022, pp. 118–127.

[9] H. Baars and H.-G. Kemper, "Management support with structured and unstructured data—An integrated business intelligence framework," *Inf. Syst. Manage.*, vol. 25, no. 2, pp. 132–148, Mar. 2008, doi: 10.1080/10580530801941058.

[10] U. F. Khattak, A. Mustapha, M. Yaseen, M. A. Shah, and A. Shahzad, "Enhancing integrity technique using distributed query operation," in *Recent Trends and Advances in Wireless and IoT-Enabled Networks*, M. A. Jan, F. Khan, and M. Alam, Eds., Cham, Switzerland: Springer, 2019, pp. 139–146.

[11] K. Adnan and R. Akbar, "Limitations of information extraction methods and techniques for heterogeneous unstructured big data," *Int. J. Eng. Bus. Manag.*, vol. 11, pp. 1–23, Dec. 2019, doi: 10.1177/1847979019890771.

[12] M. F. Abdullah and K. Ahmad, "Business intelligence model for unstructured data management," in *Proc. Int. Conf. Electr. Eng. Informat. (ICEEI)*, Aug. 2015, pp. 473–477.

[13] M. F. Abdullah and K. Ahmad, "The mapping process of unstructured data to structured data," in *Proc. Int. Conf. Res. Innov. Inf. Syst. (ICRIIS)*, Nov. 2013, pp. 151–155.

[14] L. Chai, H. Xu, Z. Luo, and S. Li, "A multi-source heterogeneous data analytic method for future price fluctuation prediction," *Neurocomputing*, vol. 418, pp. 11–20, Dec. 2020, doi: 10.1016/j.neucom.2020.07.073.

[15] F. Jenhani, M. S. Gouider, and L. B. Said, "Hybrid system for information extraction from social media text: Drug abuse case study," *Proc. Comput. Sci.*, vol. 159, pp. 688–697, Jan. 2019, doi: 10.1016/j.procs.2019.09.224.

[16] K. Adnan, R. Akbar, and K. S. Wang, "Information extraction from multifaceted unstructured big data," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2, pp. 1398–1404, 2019.

[17] R. Zhang and N. El-Gohary, "A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking," *Autom. Construct.*, vol. 132, Dec. 2021, Art. no. 103834, doi: 10.1016/j.autcon.2021.103834.

[18] J. Liu, L. Gao, S. Guo, R. Ding, X. Huang, L. Ye, Q. Meng, A. Nazari, and D. Thiruvady, "A hybrid deep-learning approach for complex biochemical named entity recognition," *Knowl.-Based Syst.*, vol. 221, Jun. 2021, Art. no. 106958, doi: 10.1016/j.knosys.2021.106958.

[19] N. Talpur, S. J. Abdulkadir, and M. H. Hasan, "A deep learning based neuro-fuzzy approach for solving classification problems," in *Proc. Int. Conf. Comput. Intell. (ICCI)*, Oct. 2020, pp. 167–172.

[20] N. Talpur, S. J. Abdulkadir, H. Alhussian, M. H. Hasan, and M. H. A. Abdullah, "Optimizing deep neuro-fuzzy classifier with a novel evolutionary arithmetic optimization algorithm," *J. Comput. Sci.*, vol. 64, Oct. 2022, Art. no. 101867, doi: 10.1016/j.jocs.2022.101867.

[21] W. Etaiwi, A. Awajan, and D. Suleiman, "Statistical Arabic name entity recognition approaches: A survey," *Proc. Comput. Sci.*, vol. 113, pp. 57–64, Jan. 2017, doi: 10.1016/j.procs.2017.08.288.

[22] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: A systematic review," *Comput. Sci. Rev.*, vol. 29, pp. 21–43, Aug. 2018, doi: 10.1016/j.cosrev.2018.06.001.

[23] W. Xiang and B. Wang, "A survey of event extraction from text," *IEEE Access*, vol. 7, pp. 173111–173137, 2019, doi: 10.1109/ACCESS.2019.2956831.

[24] A. Akkasi and M.-F. Moens, "Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey," *J. Biomed. Informat.*, vol. 119, Jul. 2021, Art. no. 103820, doi: 10.1016/j.jbi.2021.103820.

[25] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: State-of-the-art," *ACM Comput. Surv.*, vol. 54, no. 1, p. 20, 2021, doi: 10.1145/3445965.

[26] M. Albared, M. G. Ocana, A. Ghareb, and T. Al-Moslmi, "Recent progress of named entity recognition over the most popular datasets," in *Proc. 1st Int. Conf. Intell. Comput. Eng. (ICOICE)*, Dec. 2019, pp. 1–9.

[27] M. P. Bach, Ž. Krstić, S. Seljan, and L. Turulja, "Text mining for big data analysis in financial sector: A literature review," *Sustainability*, vol. 11, no. 5, p. 1277, Feb. 2019.

[28] 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, Jan. 2020, doi: 10.1016/j.aiopen.2021.02.004.

[29] G. Frisoni, G. Moro, and A. Carbonaro, "A survey on event extraction for natural language understanding: Riding the biomedical literature wave," *IEEE Access*, vol. 9, pp. 160721–160757, 2021, doi: 10.1109/ACCESS.2021.3130956.

[30] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 50–70, Jan. 2022, doi: 10.1109/TKDE.2020.2981314.

[31] B. A. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ., Durham Univ., U.K., Joint Rep. EBSE 2007-001, 2007. [Online]. Available: https://www.elsevier.com/data/promis_misc/525444systematicreviewsguide.pdf and https://www.bibsonomy.org/bibtex/aed0229656ada843d3e3f24e5e5c9eb9

[32] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Systematic Rev.*, vol. 10, no. 1, p. 89, Dec. 2021, doi: 10.1186/s13643-021-01626-4.

[33] N. Talpur, S. J. Abdulkadir, H. Alhussian, H. Hasan, N. Aziz, and A. Bamhdi, "A comprehensive review of deep neuro-fuzzy system architectures and their optimization methods," *Neural Comput. Appl.*, vol. 34, no. 3, pp. 1837–1875, Feb. 2022, doi: 10.1007/s00521-021-06807-9.

[34] N. Talpur et al., "Deep neuro-fuzzy system application trends, challenges, and future perspectives: A systematic survey," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 865–913, Feb. 2023.

[35] Y. Y. Huang and W. Y. Wang, "Deep residual learning for weakly-supervised relation extraction," IN *Proc. Conf. Empirical Methods Natural Lang. Process.* Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 1803–1807. [Online]. Available: https://aclanthology.org/D17-1191/

[36] T. Al-Moslmi, M. G. Ocana, A. L. Opdahl, and C. Veres, "Named entity extraction for knowledge graphs: A literature overview," *IEEE Access*, vol. 8, pp. 32862–32881, 2020, doi: 10.1109/ACCESS.2020.2973928.

[37] R. A. Bridges, K. M. T. Huffer, C. L. Jones, M. D. Iannacone, and J. R. Goodall, "Cybersecurity automated information extraction techniques: Drawbacks of current methods, and enhanced extractors," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 437–442.

[38] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. I. Tsujii, "BRAT: A web-based tool for NLP-assisted text annotation," in *Proc. Demonstrations 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2012, pp. 102–107.

[39] M. Akmal and A. Romadhony, "Corpus development for Indonesian product named entity recognition using semi-supervised approach," in *Proc. Int. Conf. Data Sci. Appl. (ICoDSA)*, Aug. 2020, pp. 1–5.

[40] H. Wu, Q. Lei, X. Zhang, and Z. Luo, "Creating a large-scale financial news corpus for relation extraction," in *Proc. 3rd Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2020, pp. 259–263.
[41] C. B. Wang, X. Ma, J. Chen, and J. Chen, "Information extraction and knowledge graph construction from geoscience literature," *Comput. Geosci.*, vol. 112, pp. 112–120, Mar. 2018, doi: 10.1016/j.cageo.2017.12.007.
[42] G. Park, H.-G. Lee, and H. Kim, "Named entity recognition model based on neural networks using parts of speech probability and gazetteer features," *Adv. Sci. Lett.*, vol. 23, no. 10, pp. 9530–9533, Oct. 2017, doi: 10.1166/asl.2017.9740.
[43] B. Yin, Y. Wang, D. Pei, and Y. Yan, "Research on the extraction of entity relationships from fusion syntactic information," in *Proc. IEEE Int. Conferences Ubiquitous Comput. Commun. (IUCC) Data Sci. Comput. Intell. (DSCI) Smart Comput., Netw. Services (SmartCNS)*, Oct. 2019, pp. 258–265.
[44] Y. Choi, M. D. Nguyen, and T. N. Kerr, "Syntactic and semantic information extraction from NPP procedures utilizing natural language processing integrated with rules," *Nucl. Eng. Technol.*, vol. 53, no. 3, pp. 866–878, Mar. 2021, doi: 10.1016/j.net.2020.08.010.
[45] Y. Chen, K. Wang, W. Yang, Y. Qing, R. Huang, and P. Chen, "A multi-channel deep neural network for relation extraction," *IEEE Access*, vol. 8, pp. 13195–13203, 2020, doi: 10.1109/ACCESS.2020.2966303.
[46] I. Fernandes, H. L. Cardoso, and E. Oliveira, "Applying deep neural networks to named entity recognition in Portuguese texts," in *Proc. 5th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Oct. 2018, pp. 284–289.
[47] Y. Pang, J. Liu, L. Liu, Z. Yu, and K. Zhang, "A deep neural network model for joint entity and relation extraction," *IEEE Access*, vol. 7, pp. 179143–179150, 2019, doi: 10.1109/ACCESS.2019.2949086.
[48] G. Chen, W. Mao, Q. Kong, and H. Han, "Joint learning with keyword extraction for event detection in social media," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Nov. 2018, pp. 214–219.
[49] D. Ji, P. Tao, H. Fei, and Y. Ren, "An end-to-end joint model for evidence information extraction from court record document," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102305, doi: 10.1016/j.ipm.2020.102305.
[50] P. Sun, X. Yang, X. Zhao, and Z. Wang, "An overview of named entity recognition," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2018, pp. 273–278.
[51] R. Grishman and B. Sundheim, "Message understanding conference-6: A brief history," in *Proc. 16th Conf. Comput. Linguistics*, 1996, pp. 1–6.
[52] M. Orellana, C. Farez, and P. Cardenas, "Evaluating named entities recognition (NER) tools vs algorithms adapted to the extraction of locations," in *Proc. Int. Conf. Digit. Transformation Innov. Technol. (Incodtrin)*, Oct. 2020, pp. 123–128.
[53] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Information extraction from scientific articles: A survey," *Scientometrics*, vol. 117, no. 3, pp. 1931–1990, Dec. 2018, doi: 10.1007/s11192-018-2921-5.
[54] G. Popovski, B. K. Seljak, and T. Eftimov, "A survey of named-entity recognition methods for food information extraction," *IEEE Access*, vol. 8, pp. 31586–31594, 2020, doi: 10.1109/ACCESS.2020.2973502.
[55] J. Li, B. Chiu, S. Feng, and H. Wang, "Few-shot named entity recognition via meta-learning," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 9, pp. 4245–4256, Sep. 2022, doi: 10.1109/TKDE.2020.3038670.
[56] J. Wang, S. Li, E. Agyemang-Duah, X. Feng, C. Xu, Y. Ji, and J. Liu, "Fine-grained Chinese named entity recognition based on MacBERT-Attn-BiLSTM-CRF model," in *Proc. IEEE 12th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2022, pp. 0125–0131.
[57] A. Anandika and S. P. Mishra, "A study on machine learning approaches for named entity recognition," in *Proc. Int. Conf. Appl. Mach. Learn. (ICAML)*, May 2019, pp. 153–159.
[58] L. A. Mady, Y. M. Afify, and N. L. Badr, "Enhancing performance of biomedical named entity recognition," in *Proc. 10th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, Dec. 2021, pp. 467–472.
[59] K. Zaporojets, J. Deleu, C. Develder, and T. Demeester, "DWIE: An entity-centric dataset for multi-task document-level information extraction," *Inf. Process. Manage.*, vol. 58, no. 4, Jul. 2021, Art. no. 102563, doi: 10.1016/j.ipm.2021.102563.
[60] R. Chaniago and M. L. Khodra, "Information extraction on novel text using machine learning and rule-based system," in *Proc. Int. Conf. Innov. Creative Inf. Technol. (ICITech)*, Nov. 2017, pp. 1–6.
[61] C. A. Aguirre, S. Gullapalli, M. F. D. L. Torre, A. Lam, J. L. Weese, and W. H. Hsu, "Learning to filter documents for information extraction using rapid annotation," in *Proc. Int. Conf. Mach. Learn. Data Sci. (MLDS)*, Dec. 2017, pp. 85–90.
[62] O. Ooban, S. A. Ozel, and A. Inan, "Named entity recognition over FBNER: A new Facebook dataset in Turkish," in *Proc. Innov. Intell. Syst. Appl. Conf. (ASYU)*, Oct. 2021, pp. 1–6.
[63] Q. Zhang, M. Chen, and L. Liu, "A review on entity relation extraction," in *Proc. 2nd Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, Dec. 2017, pp. 178–183.
[64] X. Han and L. Wang, "A novel document-level relation extraction method based on BERT and entity information," *IEEE Access*, vol. 8, pp. 96912–96919, 2020, doi: 10.1109/ACCESS.2020.2996642.
[65] Y. Gultom and W. C. Wibowo, "Automatic open domain information extraction from Indonesian text," in *Proc. Int. Workshop Big Data Inf. Secur. (IWBIS)*, Sep. 2017, pp. 23–30.
[66] F. Alshuwaier, A. Areshey, and J. Poon, "A comparative study of the current technologies and approaches of relation extraction in biomedical literature using text mining," in *Proc. 4th IEEE Int. Conf. Eng. Technol. Appl. Sci. (ICETAS)*, Nov. 2017, pp. 1–13.
[67] P. Abdurehim, T. Tohti, and A. Hamdulla, "A short review of relation extraction methods," in *Proc. 13th Int. Conf. Intell. Comput. Technol. Autom. (ICICTA)*, Oct. 2020, pp. 18–22.
[68] A. Gupta, I. Banerjee, and D. L. Rubin, "Automatic information extraction from unstructured mammography reports using distributed semantics," *J. Biomed. Informat.*, vol. 78, pp. 78–86, Feb. 2018, doi: 10.1016/j.jbi.2017.12.016.
[69] D. Christou and G. Tsoumakas, "Improving distantly-supervised relation extraction through BERT-based label and instance embeddings," *IEEE Access*, vol. 9, pp. 62574–62582, 2021, doi: 10.1109/ACCESS.2021.3073428.
[70] L. Xue, S. Qing, and Z. Pengzhou, "Relation extraction based on deep learning," in *Proc. IEEE/ACIS 17th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2018, pp. 687–691.
[71] Y. Chen, "A transfer learning model with multi-source domains for biomedical event trigger extraction," *BMC Genomics*, vol. 22, no. 1, p. 31, Dec. 2021, doi: 10.1186/s12864-020-07315-1.
[72] C. Ruan, Y. Wu, G. Sheng Luo, Y. Yang, and P. Ma, "Relation extraction for Chinese clinical records using multi-view graph learning," *IEEE Access*, vol. 8, pp. 215613–215622, 2020, doi: 10.1109/ACCESS.2020.3037086.
[73] R. Saheb-Nassagh, M. Asgari, and B. Minaei-Bidgoli, "RePersian: An efficient open information extraction tool in Persian," in *Proc. 6th Int. Conf. Web Res. (ICWR)*, Apr. 2020, pp. 93–99.
[74] A. Kumar and B. Starly, "'FabNER': Information extraction from manufacturing process science domain literature using named entity recognition," *J. Intell. Manuf.*, vol. 33, pp. 2393–2407, Jun. 2021, doi: 10.1007/s10845-021-01807-x.
[75] G. Puccetti, F. Chiarello, and G. Fantoni, "A simple and fast method for named entity context extraction from patents," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115570, doi: 10.1016/j.eswa.2021.115570.
[76] Y. Chen, C. Zhou, T. Li, H. Wu, X. Zhao, K. Ye, and J. Liao, "Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training," *J. Biomed. Informat.*, vol. 96, Aug. 2019, Art. no. 103252, doi: 10.1016/j.jbi.2019.103252.
[77] K. Englmeier, "Named entities and their role in creating context information," *Proc. Comput. Sci.*, vol. 176, pp. 2069–2076, Jan. 2020, doi: 10.1016/j.procs.2020.09.243.
[78] J. Santoso, E. I. Setiawan, C. N. Purwanto, E. M. Yuniarno, M. Hariadi, and M. H. Purnomo, "Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory," *Expert Syst. Appl.*, vol. 176, Aug. 2021, Art. no. 114856, doi: 10.1016/j.eswa.2021.114856.
[79] C. Chantrapornchai and A. Tunsakul, "Information extraction based on named entity for tourism corpus," in *Proc. 16th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, Jul. 2019, pp. 187–192.
[80] S. Suravee, T. Stoev, D. Schindler, I. Hochgraeber, C. Pinkert, B. Holle, M. Halek, F. Kruger, and K. Yordanova, "Annotation scheme for named entity recognition and relation extraction tasks in the domain of people with dementia," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops Other Affiliated Events (PerCom Workshops)*, Mar. 2022, pp. 236–241.

[81] H. Zhao, Y. Pan, and F. Yang, "Research on information extraction of technical documents and construction of domain knowledge graph," *IEEE Access*, vol. 8, pp. 168087–168098, 2020, doi: 10.1109/ACCESS.2020.3024070.

[82] S. Venkatachalam, L. P. Subbiah, R. Rajendiran, and N. Venkatachalam, "An ontology-based information extraction and summarization of multiple news articles," *Int. J. Inf. Technol.*, vol. 12, no. 2, pp. 547–557, Jun. 2020, doi: 10.1007/s41870-019-00367-x.

[83] D. Feng and H. Chen, "A small samples training framework for deep learning-based automatic information extraction: Case study of construction accident news reports analysis," *Adv. Eng. Informat.*, vol. 47, Jan. 2021, Art. no. 101256, doi: 10.1016/j.aei.2021.101256.

[84] E. H. M. D. Silva, J. Laterza, M. P. P. D. Silva, and M. Ladeira, "A proposal to identify stakeholders from news for the institutional relationship management activities of an institution based on named entity recognition using BERT," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 1569–1575.

[85] Y. Norouzi, "Spatial, temporal, and semantic crime analysis using information extraction from online news," in *Proc. 8th Int. Conf. Web Res. (ICWR)*, May 2022, pp. 40–46.

[86] A. Wosiak, "Automated extraction of information from Polish resume documents in the IT recruitment process," *Proc. Comput. Sci.*, vol. 192, pp. 2432–2439, Jan. 2021, doi: 10.1016/j.procs.2021.09.012.

[87] W. Gunawan, D. Suhartono, F. Purnomo, and A. Ongko, "Named-entity recognition for Indonesian language using bidirectional LSTM-CNNs," *Proc. Comput. Sci.*, vol. 135, pp. 425–432, Jan. 2018, doi: 10.1016/j.procs.2018.08.193.

[88] F. Y. Azalia, M. A. Bijaksana, and A. F. Huda, "Name indexing in Indonesian translation of Hadith using named entity recognition with Naïve Bayes classifier," *Proc. Comput. Sci.*, vol. 157, pp. 142–149, Jan. 2019, doi: 10.1016/j.procs.2019.08.151.

[89] X. Li, H. Zhang, and X.-H. Zhou, "Chinese clinical named entity recognition with variant neural structures based on BERT methods," *J. Biomed. Informat.*, vol. 107, Jul. 2020, Art. no. 103422, doi: 10.1016/j.jbi.2020.103422.

[90] D. C. Wintaka, M. A. Bijaksana, and I. Asror, "Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF," *Proc. Comput. Sci.*, vol. 157, pp. 221–228, Jan. 2019, doi: 10.1016/j.procs.2019.08.161.

[91] R. de Aquino Silva, L. da Silva, M. L. Dutra, and G. M. de Araujo, "An improved NER methodology to the Portuguese language," *Mobile Netw. Appl.*, vol. 26, no. 1, pp. 319–325, Feb. 2021, doi: 10.1007/s11036-020-01644-x.

[92] M. S. Salleh, S. A. Asmai, H. Basiron, and S. Ahmad, "A Malay named entity recognition using conditional random fields," in *Proc. 5th Int. Conf. Inf. Commun. Technol. (ICoIC7)*, May 2017, pp. 1–6.

[93] S. Sukardi, M. Susanty, A. Irawan, and R. F. Putra, "Low complexity named-entity recognition for Indonesian language using BiLSTM-CNNs," in *Proc. 3rd Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Nov. 2020, pp. 137–142.

[94] I. S. Azarine, M. A. Bijaksana, and I. Asror, "Named entity recognition on Indonesian tweets using hidden Markov model," in *Proc. 7th Int. Conf. Inf. Commun. Technol. (ICoICT)*, Jul. 2019, pp. 1–5.

[95] B. T. Nguyen, T. T. N. Doan, S. T. Huynh, K. Q. Tran, A. T. Nguyen, A. T.-H. Le, A. M. Tran, N. Ho, T. T. Nguyen, and D. T. Huynh, "An end-to-end named entity recognition platform for Vietnamese real estate advertisement posts and analytical applications," *IEEE Access*, vol. 10, pp. 87681–87697, 2022, doi: 10.1109/ACCESS.2022.3195496.

[96] N. Alsaaran and M. Alrabiah, "Classical Arabic named entity recognition using variant deep neural network architectures and BERT," *IEEE Access*, vol. 9, pp. 91537–91547, 2021, doi: 10.1109/ACCESS.2021.3092261.

[97] J. Diao, Z. Zhou, and G. Shi, "Leveraging integrated learning for open-domain Chinese named entity recognition," *Int. J. Crowd Sci.*, vol. 6, no. 2, pp. 74–79, Jun. 2022, doi: 10.26599/IJCS.2022.9100015.

[98] G. Kim, J. Son, J. Kim, H. Lee, and H. Lim, "Enhancing Korean named entity recognition with linguistic tokenization strategies," *IEEE Access*, vol. 9, pp. 151814–151823, 2021, doi: 10.1109/ACCESS.2021.3126882.

[99] L. Qiu, D. Ru, Q. Long, W. Zhang, and Y. Yu, "QA4IE: A question answering based system for document-level general information extraction," *IEEE Access*, vol. 8, pp. 29677–29689, 2020, doi: 10.1109/ACCESS.2020.2970119.

[100] P. Banerjee, K. K. Pal, M. Devarakonda, and C. Baral, "Biomedical named entity recognition via knowledge guidance and question answering," *ACM Trans. Comput. Healthcare*, vol. 2, no. 4, p. 33, 2021, doi: 10.1145/3465221.

[101] K. Won, Y. Jang, H.-D. Choi, and S. Shin, "Design and implementation of information extraction system for scientific literature using fine-tuned deep learning models," *ACM SIGAPP Appl. Comput. Rev.*, vol. 22, no. 1, pp. 31–38, Mar. 2022, doi: 10.1145/3530043.3530047.

[102] A. Iovine, A. Fang, B. Fetahu, O. Rokhlenko, and S. Malmasi, "CycleNER: An unsupervised training approach for named entity recognition," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2916–2924.

[103] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "Improving named entity recognition in noisy user-generated text with local distance neighbor feature," *Neurocomputing*, vol. 382, pp. 1–11, Mar. 2020, doi: 10.1016/j.neucom.2019.11.072.

[104] D. Ameta and P. M. Jat, "Information extraction from Wikipedia articles using DeepDive," in *Proc. Int. Conf. Commun. Inf. Comput. Technol. (ICCICT)*, Feb. 2018, pp. 1–6.

[105] C. Liu, Y. Yu, X. Li, and P. Wang, "Named entity recognition in equipment support field using tri-training algorithm and text information extraction technology," *IEEE Access*, vol. 9, pp. 126728–126734, 2021, doi: 10.1109/ACCESS.2021.3109911.

[106] P. Ma, B. Jiang, Z. Lu, N. Li, and Z. Jiang, "Cybersecurity named entity recognition using bidirectional long short-term memory with conditional random fields," *Tsinghua Sci. Technol.*, vol. 26, no. 3, pp. 259–265, Jun. 2021, doi: 10.26599/TST.2019.9010033.

[107] G. B. Herwanto, G. Quirchmayr, and A. M. Tjoa, "A named entity recognition based approach for privacy requirements engineering," in *Proc. IEEE 29th Int. Requirements Eng. Conf. Workshops (REW)*, Sep. 2021, pp. 406–411.

[108] P. K. Putra, D. B. Sencaki, G. P. Dinanta, F. Alhasanah, and R. Ramadhan, "Flood monitoring with information extraction approach from social media data," in *Proc. IEEE Asia–Pacific Conf. Geosci., Electron. Remote Sens. Technol. (AGERS)*, Dec. 2020, pp. 113–119.

[109] P. R. Deshmukh and R. Phalnikar, "Information extraction for prognostic stage prediction from breast cancer medical records using NLP and ML," *Med. Biol. Eng. Comput.*, vol. 59, no. 9, pp. 1751–1772, Sep. 2021, doi: 10.1007/s11517-021-02399-7.

[110] E. Yehia, H. Boshnak, S. AbdelGaber, A. Abdo, and D. S. Elzanfaly, "Ontology-based clinical information extraction from physician's free-text notes," *J. Biomed. Informat.*, vol. 98, Oct. 2019, Art. no. 103276, doi: 10.1016/j.jbi.2019.103276.

[111] X. Liu, Y. Zhou, and Z. Wang, "Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network," *J. Vis. Commun. Image Represent.*, vol. 60, pp. 1–15, Apr. 2019, doi: 10.1016/j.jvcir.2019.02.001.

[112] L. Zhou, J. Li, Z. Gu, J. Qiu, B. B. Gupta, and Z. Tian, "PANNER: POS-aware nested named entity recognition through heterogeneous graph neural network," *IEEE Trans. Computat. Social Syst.*, early access, Mar. 30, 2022, doi: 10.1109/TCSS.2022.3159366.

[113] L.-H. Lee and Y. Lu, "Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 7, pp. 2801–2810, Jul. 2021, doi: 10.1109/JBHI.2020.3048700.

[114] U. Naseem, M. Khushi, V. Reddy, S. Rajendran, I. Razzak, and J. Kim, "BioALBERT: A simple and effective pre-trained language model for biomedical named entity recognition," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2021, pp. 1–7.

[115] N. Liu, Q. Hu, H. Xu, X. Xu, and M. Chen, "Med-BERT: A pretraining framework for medical records named entity recognition," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5600–5608, Aug. 2022, doi: 10.1109/TII.2021.3131180.

[116] R. Ramachandran and K. Arutchelvan, "Named entity recognition on biomedical literature documents using hybrid-based approach," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 2, pp. 1–10, Feb. 2023, doi: 10.1007/s12652-021-194703078-z.

[117] H. Kim and J. Kang, "How do your biomedical named entity recognition models generalize to novel entities?" *IEEE Access*, vol. 10, pp. 31513–31523, 2022, doi: 10.1109/ACCESS.2022.3157854.

[118] C. Suman, S. M. Reddy, S. Saha, and P. Bhattacharyya, "Why pay more? A simple and efficient named entity recognition system for tweets," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114101, doi: 10.1016/j.eswa.2020.114101.

[119] V. Rachman, S. Savitri, F. Augustianti, and R. Mahendra, "Named entity recognition on Indonesian Twitter posts using long short-term memory networks," in *Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS)*, Oct. 2017, pp. 228–232.

[120] M. E. Barachi, S. S. Mathew, and M. AlKhatib, "Combining named entity recognition and emotion analysis of tweets for early warning of violent actions," in *Proc. 7th Int. Conf. Smart Sustain. Technol. (SpliTech)*, Jul. 2022, pp. 1–6.

[121] S. Dai, Y. Ding, Z. Zhang, W. Zuo, X. Huang, and S. Zhu, "GrantExtractor: Accurate grant support information extraction from biomedical full-text based on Bi-LSTM-CRF," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 1, pp. 205–215, 2021, doi: 10.1109/TCBB.2019.2939128.

[122] A. Lyra, C. E. Barbosa, Y. Lima, H. Salazar, and J. Souza, "NERMAP: Collaborative building of technological roadmaps using named entity recognition," in *Proc. IEEE 25th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2022, pp. 986–991.

[123] A. O. Shelmanov, D. A. Devyatkin, V. A. Isakov, and I. V. Smirnov, "Open information extraction from texts: Part II. Extraction of semantic relationships using unsupervised machine learning," *Sci. Tech. Inf. Process.*, vol. 47, no. 6, pp. 340–347, Dec. 2020, doi: 10.3103/S0147688220060076.

[124] A. Romadhony, A. Purwarianti, and D. H. Widyantoro, "Rule-based Indonesian open information extraction," in *Proc. 5th Int. Conf. Adv. Inform., Concept Theory Appl. (ICAICTA)*, Aug. 2018, pp. 107–112.

[125] D. Sousa and F. M. Couto, "Biomedical relation extraction with knowledge graph-based recommendations," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 4207–4217, Aug. 2022, doi: 10.1109/JBHI.2022.3173558.

[126] M. S. Parniani and M. Z. Reformat, "Relation extraction with sentence simplification process and entity information," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Dec. 2021, pp. 635–640.

[127] J. Hou, X. Li, H. Yao, H. Sun, T. Mai, and R. J. I. A. Zhu, "Bert-based Chinese relation extraction for public security," *IEEE Access*, vol. 8, pp. 132367–132375, 2020, doi: 10.1109/ACCESS.2020.3002863.

[128] B. Kong, S. Liu, F. Wei, L. Jia, and G. Wang, "Chinese relation extraction using extend softword," *IEEE Access*, vol. 9, pp. 110299–110308, 2021, doi: 10.1109/ACCESS.2021.3102225.

[129] K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, and P. He, "Fine-tuning BERT for joint entity and relation extraction in Chinese medical text," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 892–897.

[130] M. N. Nityasya, R. Mahendra, and M. Adriani, "Hypernym-hyponym relation extraction from Indonesian Wikipedia text," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2018, pp. 285–289.

[131] Y. Li, "Research on Chinese entity relation extraction method based on deep learning," in *Proc. Int. Conf. Commun., Inf. Syst. Comput. Eng. (CISCE)*, May 2021, pp. 731–735.

[132] S. Qi, L. Zheng, and F. Shang, "Dependency parsing-based entity relation extraction over Chinese complex text," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 4, pp. 1–34, Jul. 2021, doi: 10.1145/3450273.

[133] M. Zhou, D. Ji, and F. Li, "Relation extraction in dialogues: A deep learning model based on the generality and specialty of dialogue text," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2015–2026, 2021, doi: 10.1109/TASLP.2021.3082295.

[134] V. Suárez-Paniagua, R. M. R. Zavala, I. Segura-Bedmar, and P. Martínez, "A two-stage deep learning approach for extracting entities and relationships from medical texts," *J. Biomed. Informat.*, vol. 99, Nov. 2019, Art. no. 103285, doi: 10.1016/j.jbi.2019.103285.

[135] S. Yadav, S. Ramesh, S. Saha, and A. Ekbal, "Relation extraction from biomedical and clinical text: Unified multitask learning framework," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 2, pp. 1105–1116, Mar./Apr. 2020, doi: 10.1109/TCBB.2020.3020016.

[136] Y. Hu, H. Shen, W. Liu, F. Min, X. Qiao, and K. Jin, "A graph convolutional network with multiple dependency representations for relation extraction," *IEEE Access*, vol. 9, pp. 81575–81587, 2021, doi: 10.1109/ACCESS.2021.3086480.

[137] A. Ramponi, S. Giampiccolo, D. Tomasoni, C. Priami, and R. Lombardo, "High-precision biomedical relation extraction for reducing human curation efforts in industrial applications," *IEEE Access*, vol. 8, pp. 150999–151011, 2020, doi: 10.1109/ACCESS.2020.3014862.

[138] Q. Lai, S. Ding, J. Gong, J. Cui, and S. Liu, "A Chinese multi-modal relation extraction model for internet security of finance," in *Proc. 52nd Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*, Jun. 2022, pp. 123–128.

[139] S. Sahnoun, S. Elloumi, and S. Ben Yahia, "Event detection based on open information extraction and ontology," *J. Inf. Telecommun.*, vol. 4, no. 3, pp. 383–403, Jul. 2020, doi: 10.1080/24751839.2020.1763007.

[140] R. Pradeep Kumar and P. Aswathi, "Extraction of causality and related events using text analysis," in *Proc. 2nd Int. Conf. Intell. Comput., Instrum. Control Technol. (ICICICT)*, Jul. 2019, pp. 1448–1453.

[141] C. Chang, Y. Tang, Y. Long, K. Hu, Y. Li, J. Li, and C.-D. Wang, "Multi-information preprocessing event extraction with BiLSTM-CRF attention for academic knowledge graph construction," *IEEE Trans. Computat. Social Syst.*, early access, Jul. 4, 2022, doi: 10.1109/TCSS.2022.3183685.

[142] A.-U.-N. Fatima, H. Ahmad, M. Ahmad, W. Ahmad, and N. Faisal, "Extraction of temporal Events' frequency from online news channels," in *Proc. 30th Int. Conf. Comput. Theory Appl. (ICCTA)*, Dec. 2020, pp. 109–116.

[143] F. Rahma and A. Romadhony, "Rule-based crime information extraction on Indonesian digital news," in *Proc. Int. Conf. Data Sci. Appl. (ICoDSA)*, Oct. 2021, pp. 10–15.

[144] X. Gao, Z. Diao, K. Wei, Y. Yang, and L. Li, "Event extraction via rules and machine learning," in *Proc. IEEE 6th Int. Conf. Cloud Comput. Intell. Syst. (CCIS)*, Dec. 2019, pp. 41–46.

[145] S. Agarwal, V. Aggarwal, A. R. Akula, G. B. Dasgupta, and G. Sridhara, "Automatic problem extraction and analysis from unstructured text in IT tickets," *IBM J. Res. Develop.*, vol. 61, no. 1, pp. 4:41–4:52, Jan. 2017, doi: 10.1147/JRD.2016.2629318.

[146] J. Liu and X. Huang, "Forecasting crude oil price using event extraction," *IEEE Access*, vol. 9, pp. 149067–149076, 2021, doi: 10.1109/ACCESS.2021.3124802.

[147] B. Yu, "Hazard information extraction and classification based on domain ontology," in *Proc. 11th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2018, pp. 1–5.

[148] R. Interdonato, J.-L. Guillaume, and A. Doucet, "A lightweight and multilingual framework for crisis information extraction from Twitter data," *Social Netw. Anal. Mining*, vol. 9, no. 1, p. 65, Dec. 2019, doi: 10.1007/s13278-019-0608-4.

[149] M. M. Mirończuk, "Information extraction system for transforming unstructured text data in fire reports into structured forms: A Polish case study," *Fire Technol.*, vol. 56, no. 2, pp. 545–581, 2020, doi: 10.1007/s10694-019-00891-z.

[150] D. Liu, Q. Fu, C. Wan, X. Liu, T. Jiang, G. Liao, X. Qiu, and R. Liu, "Suicidal ideation cause extraction from social texts," *IEEE Access*, vol. 8, pp. 169333–169351, 2020, doi: 10.1109/ACCESS.2020.3019491.

[151] Y. Wu, H. Sun, and C. Yan, "An event timeline extraction method based on news corpus," in *Proc. IEEE 2nd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2017, pp. 697–702.

[152] N. Viani, C. Larizza, V. Tibollo, C. Napolitano, S. G. Priori, R. Bellazzi, and L. Sacchi, "Information extraction from Italian medical reports: An ontology-driven approach," *Int. J. Med. Informat.*, vol. 111, pp. 140–148, Mar. 2018, doi: 10.1016/j.ijmedinf.2017.12.013.

[153] W. J. Hou and B. Ceesay, "Domain transformation on biological event extraction by learning methods," *J. Biomed. Informat.*, vol. 95, Jul. 2019, Art. no. 103236, doi: 10.1016/j.jbi.2019.103236.

[154] X. Yu, W. Rong, J. Liu, D. Zhou, Y. Ouyang, and Z. Xiong, "LSTM-based end-to-end framework for biomedical event extraction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 6, pp. 2029–2039, Nov. 2020, doi: 10.1109/TCBB.2019.2916346.

[155] K. Espinosa, P. Georgiadis, F. Christopoulou, M. Ju, M. Miwa, and S. Ananiadou, "Comparing neural models for nested and overlapping biomedical event detection," *BMC Bioinf.*, vol. 23, no. 1, p. 211, Dec. 2022, doi: 10.1186/s12859-022-04746-3.

[156] F. Su, Y. Zhang, F. Li, and D. Ji, "Balancing precision and recall for neural biomedical event extraction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 1637–1649, 2022, doi: 10.1109/TASLP.2022.3161146.

[157] G. B. Herwanto and D. P. Dewantara, "Traffic condition information extraction from Twitter data," in *Proc. Int. Conf. Electr. Eng. Informat. (ICELTICs)*, Sep. 2018, pp. 95–100.

[158] S. S. Burramsetty, N. P. Gonugunta, S. Uppu, and S. K. Nunna, "Event extraction from Telugu-english code mixed social media text," in *Proc. 7th Int. Conf. Commun. Electron. Syst. (ICCES)*, Jun. 2022, pp. 946–951.

[159] T. Kolajo, O. Daramola, and A. A. Adebiyi, "Real-time event detection in social media streams through semantic analysis of noisy terms," *J. Big Data*, vol. 9, no. 1, p. 90, Dec. 2022, doi: 10.1186/s40537-022-00642-y.

[160] F. B. Rodrigues, W. F. Giozza, R. de Oliveira Albuquerque, and L. J. G. Villalba, "Natural language processing applied to forensics information extraction with transformers and graph visualization," *IEEE Trans. Computat. Social Syst.*, early access, Apr. 5, 2022, doi: 10.1109/TCSS.2022.3159677.

[161] S. S. Naik and M. N. Gaonkar, "Extractive text summarization by feature-based sentence extraction using rule-based concept," in *Proc. 2nd IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, May 2017, pp. 1364–1368.

[162] A. Felicetti, D. Williams, I. Galluccio, D. Tudhope, and F. Niccolucci, "NLP tools for knowledge extraction from Italian archaeological free text," in *Proc. 3rd Digit. Heritage Int. Congr. (DigitalHERITAGE) Held Jointly With 24th Int. Conf. Virtual Syst. Multimedia (VSMM)*, Oct. 2018, pp. 1–8.

[163] P. P. Shelke and A. A. Pardeshi, "Review on candidate feature extraction and categorization for unstructured text document," in *Proc. 4th Int. Conf. Comput. Methodolog. Commun. (ICCMC)*, Mar. 2020, pp. 88–92.

[164] A. Ayadi, M. Auffan, and J. Rose, "Ontology-based NLP information extraction to enrich nanomaterial environmental exposure database," *Proc. Comput. Sci.*, vol. 176, pp. 360–369, Jan. 2020, doi: 10.1016/j.procs.2020.08.037.

[165] K. Liu and N. El-Gohary, "Ontology-based semi-supervised conditional random fields for automated information extraction from bridge inspection reports," *Autom. Construct.*, vol. 81, pp. 313–327, Sep. 2017, doi: 10.1016/j.autcon.2017.02.003.

[166] X. Luo, P. Gandhi, S. Storey, and K. Huang, "A deep language model for symptom extraction from clinical text and its application to extract COVID-19 symptoms from social media," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 4, pp. 1737–1748, Apr. 2022, doi: 10.1109/JBHI.2021.3123192.

[167] S. Wang, M. Pang, C. Pan, J. Yuan, B. Xu, M. Du, and H. Zhang, "Information extraction for intestinal cancer electronic medical records," *IEEE Access*, vol. 8, pp. 125923–125934, 2020, doi: 10.1109/ACCESS.2020.3005684.

[168] Y. Sun, J. Wang, H. Lin, Y. Zhang, and Z. Yang, "Knowledge guided attention and graph convolutional networks for chemical-disease relation extraction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Dec. 28, 2021, doi: 10.1109/TCBB.2021.3135844.

[169] D. Hemavathi, M. Kavitha, and N. B. Ahmed, "Information extraction from social media: Clustering and labelling microblogs," in *Proc. Int. Conf. IoT Appl. (ICIOT)*, May 2017, pp. 1–10.

[170] G. Zaman, H. Mahdin, K. Hussain, J. Abawajy, and S. A. Mostafa, "An ontological framework for information extraction from diverse scientific sources," *IEEE Access*, vol. 9, pp. 42111–42124, 2021, doi: 10.1109/ACCESS.2021.3063181.

[171] Z. Hong, L. Ward, K. Chard, B. Blaiszik, and I. Foster, "Challenges and advances in information extraction from scientific literature: A review," *JOM*, vol. 73, no. 11, pp. 3383–3400, Nov. 2021, doi: 10.1007/s11837-021-04902-9.

[172] F. Zhao, X. Gui, Y. Huang, H. Jin, and L. T. Yang, "Dynamic entity-based named entity recognition under unconstrained tagging schemes," *IEEE Trans. Big Data*, vol. 8, no. 4, pp. 1059–1072, Aug. 2022, doi: 10.1109/TBDATA.2020.2998770.

[173] S. Alves, J. Costa, and J. Bernardino, "Information extraction applications for clinical trials: A survey," in *Proc. 14th Iberian Conf. Inf. Syst. Technol. (CISTI)*, Jun. 2019, pp. 1–6.

[174] M. Zhong, G. Liu, J. Xiong, and J. Zuo, "DualNER: A trigger-based dual learning framework for low-resource named entity recognition," *IEEE Intell. Syst.*, vol. 37, no. 4, pp. 79–87, Jul. 2022, doi: 10.1109/MIS.2022.3167168.

[175] Y. Yu, K. He, and J. Li, "Adversarial training for supervised relation extraction," *Tsinghua Sci. Technol.*, vol. 27, no. 3, pp. 610–618, Jun. 2022, doi: 10.26599/TST.2020.9010059.

[176] X. Xu, T. Gao, Y. Wang, and X. Xuan, "Event temporal relation extraction with attention mechanism and graph neural network," *Tsinghua Sci. Technol.*, vol. 27, no. 1, pp. 79–90, Feb. 2022, doi: 10.26599/TST.2020.9010063.

[177] G. Alfattni, N. Peek, and G. Nenadic, "Extraction of temporal relations from clinical free text: A systematic review of current approaches," *J. Biomed. Informat.*, vol. 108, Aug. 2020, Art. no. 103488, doi: 10.1016/j.jbi.2020.103488.

[178] G. Suganya and R. Porkodi, "Ontology based information extraction—A review," in *Proc. Int. Conf. Current Trends towards Converging Technol. (ICCTCT)*, Mar. 2018, pp. 1–7.

[179] R. Glauber and D. B. Claro, "A systematic mapping study on open information extraction," *Expert Syst. Appl.*, vol. 112, pp. 372–387, Dec. 2018, doi: 10.1016/j.eswa.2018.06.046.

[180] A. Konys, "Towards knowledge handling in ontology-based information extraction systems," *Proc. Comput. Sci.*, vol. 126, pp. 2208–2218, Jan. 2018, doi: 10.1016/j.procs.2018.07.228.

[181] K. Hussain, M. N. M. Salleh, S. Talpur, and N. Talpur, "Big data and machine learning in construction: A review," *Int. J. Soft Comput. Metaheuristics*, to be published.

[182] J.-D. Kim, T. Ohta, Y. Tateisi, and J. I. Tsujii, "GENIA corpus manual," Tsujii Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. TR-NLP-UT-2006-1, 2006.

[183] J.-D. Kim, T. Ohta, Y. Tateisi, and J. I. Tsujii, "GENIA ontology," Tsujii Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. TR-NLP-UT-2006-2, 2006.

[184] J.-D. Kim and J. I. Tsujii, "GENIA corpus curation framework," Tsujii Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. TR-NLP-UT-2006-3, 2006.

[185] T. Ohta, J.-D. Kim, and J. I. Tsujii, *Guidelines for Event Annotation*. Tokyo, Japan: Tsujii Laboratory, Univ. Tokyo, 2007.

[186] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "GENIA corpus—A semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. 1, pp. i180–i182, Jul. 2003, doi: 10.1093/bioinformatics/btg1023.

MOHD HAFIZUL AFIFI ABDULLAH (Graduate Student Member, IEEE) received the B.Sc. degree in CS and the M.Sc. degree in IT from Universiti Tun Hussein Onn Malaysia. He is currently pursuing the Ph.D. degree with Universiti Teknologi PETRONAS. He has two years of experience in the industry and seven years in academic institutions. He has coauthored various journal papers, conference papers, and book chapters related to information extraction, predictive analytics, personalized modeling, spiking neural networks, neuro fuzzy systems, and gene expression clustering. His research interests include predictive analytics, big data analytics, artificial intelligence, and machine learning. He serves as a Reviewer for *Computers, Materials and Continua*, *Computer Systems Science and Engineering*, and *Intelligent Automation and Soft Computing*.

NORSHAKIRAH AZIZ received the Ph.D. degree in electronic business—supply chain management integration and collaboration. She has a total experience of 18 years both in academic institutions and in the industry. Her industry working experience is related to business intelligence, e-business, and IT project management. She is currently an Associate Professor with the Computer and Information Sciences Department, Universiti Teknologi PETRONAS (UTP), Malaysia. She is currently a Researcher with the UTP Centre for Research in Data Sciences (CeRDaS) and the Data Governance Leader with the High-Performance Cloud Computing Data Centre (HPCCC). Her research interests include business intelligence, data analytics, data governance, and digital addiction. She is currently acting as an Exco Committee Member of the Data Science Association.

SAID JADID ABDULKADIR (Senior Member, IEEE) received the B.Sc. degree in computer science from Moi University, the M.Sc. degree in computer science from Universiti Teknologi Malaysia, and the Ph.D. degree in information technology from Universiti Teknologi PETRONAS. He is currently an Associate Professor with the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS. His current research interests include supervised machine learning and predictive and streaming analytics. He is currently serving as a Journal Reviewer for *Artificial Intelligence Review*, IEEE Access, and *Knowledge-Based Systems*.

HITHAM SEDDIG ALHASSAN ALHUSSIAN received the B.Sc. and M.Sc. degrees in computer science from the School of Mathematical Sciences, Khartoum University, Sudan, and the Ph.D. degree from Universiti Teknologi PETRONAS, Malaysia. He is currently an Associate Professor with the Department of Computer and Information Sciences and a Core Research Member of the Centre for Research in Data Science (CeRDaS), Universiti Teknologi PETRONAS. His main research interests include real-time parallel distributed systems, cloud computing, big data mining, machine learning, and secure computer-based management systems.

NOUREEN TALPUR received the master's degree in IT from Universiti Tun Hussein Onn Malaysia (UTHM). She is currently pursuing the Ph.D. degree with Universiti Teknologi PETRONAS (UTP), Malaysia. She has coauthored 14 papers in the areas of adaptive neuro-fuzzy inference system (ANFIS), membership functions, neuro-fuzzy systems, metaheuristic algorithms, and optimization. Her research publications include research papers and articles in well-known journals and conference proceedings. Her research interests include machine learning, deep learning, classification, metaheuristics, optimization, and neuro-fuzzy systems, such as ANFIS and deep neuro-fuzzy systems. She is currently serving as a Journal Reviewer for *Artificial Intelligence Review*, *The Imaging Science Journal*, *International Journal of Data Mining and Bioinformatics*, and *Concurrency and Computation: Practice and Experience*.