

# Limitations of information extraction methods and techniques for heterogeneous unstructured big data

Kiran Adnan<sup>1</sup> and Rehan Akbar<sup>1</sup>

## Abstract

During the recent era of big data, a huge volume of unstructured data are being produced in various forms of audio, video, images, text, and animation. Effective use of these unstructured big data is a laborious and tedious task. Information extraction (IE) systems help to extract useful information from this large variety of unstructured data. Several techniques and methods have been presented for IE from unstructured data. However, numerous studies conducted on IE from a variety of unstructured data are limited to single data types such as text, image, audio, or video. This article reviews the existing IE techniques along with its subtasks, limitations, and challenges for the variety of unstructured data highlighting the impact of unstructured big data on IE techniques. To the best of our knowledge, there is no comprehensive study conducted to investigate the limitations of existing IE techniques for the variety of unstructured big data. The objective of the structured review presented in this article is twofold. First, it presents the overview of IE techniques from a variety of unstructured data such as text, image, audio, and video at one platform. Second, it investigates the limitations of these existing IE techniques due to the heterogeneity, dimensionality, and volume of unstructured big data. The review finds that advanced techniques for IE, particularly for multifaceted unstructured big data sets, are the utmost requirement of the organizations to manage big data and derive strategic information. Further, potential solutions are also presented to improve the unstructured big data IE systems for future research. These solutions will help to increase the efficiency and effectiveness of the data analytics process in terms of context-aware analytics systems, data-driven decision-making, and knowledge management.

## Keywords

Big data, heterogeneous data, information extraction, multimedia data, unstructured data

Date received: 20 June 2019; accepted: 4 October 2019

## Introduction

Information extraction (IE) process is used to extract structured content in the form of entities, relations, facts, terms, and other types of information that helps the data analysis pipeline to prepare the data for analysis. The efficient and accurate transformation of unstructured data leads to improved performance of data analysis and IE. Various IE approaches have been proposed to extract structured and useful information from unstructured data which eventually helps to manage, process, and analyze unstructured data. IE systems are based on natural language processing (NLP), language modeling, and structure extraction

technique. NLP and language modeling have a significant role in the IE process but not included in the scope of this review. This review mainly focused on the extraction techniques with respect to IE subtasks and different data types.

---

Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Perak, Malaysia

### Corresponding author:

Rehan Akbar, Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Perak 31900, Malaysia.  
Email: rehan@utar.edu.my



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<http://www.creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

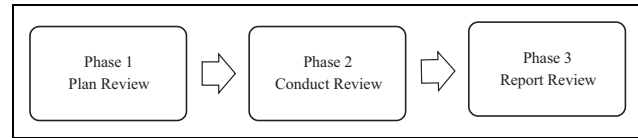
In this regard, various statistical, rule-based, and learning-based approaches for IE have been discussed.

Big data are adding more challenges to the IE process due to huge volume and variety of data (i.e. structured, semi-structured, and unstructured data). According to IDC (International Data Corporation), by 2020, unstructured data account for 95% of global data with an estimate of the compound annual growth rate of 65%.<sup>1</sup> The potential opportunities and solutions for big data are obstructed by the unstructured nature of data. The main characteristics of unstructured data are: (1) unstructured data have no schema,<sup>2,3</sup> (2) they have multiple formats,<sup>2-4</sup> (3) they come from diverse sources<sup>2-4</sup>, and (4) there is no standardization<sup>2,3</sup>, that is, different representations. The correct transformation of unstructured data improves the performance of unstructured data analytics. The complicated heterogeneity of mixed data makes it difficult to extract useful information. Due to the quality and usability issues of unstructured big data, it is important to investigate the potential and capabilities of existing IE techniques. To the best of our knowledge, this issue has not been well explored in the literature to identify the impact of unstructured big data on existing IE techniques. In this manner, the objective of this review is twofold: first, to explore the state-of-the-art techniques in IE for variety of data and second, to identify the limitations of existing IE techniques for multifaceted unstructured big data. Further, preconditions have been proposed to overcome the limitations identified in the review. The outcome of the research contributes to the identification of future directions to improve the IE systems to tackle the unstructured big data.

The rest of the article is organized as follows. In the next section, IE approaches for unstructured data have been discussed concerning IE subtasks for each data type. Further, limitations of existing IE techniques for unstructured big data analytics have been explored followed by the proposed preconditions for efficient and effective analytics. In the end, the conclusion is presented.

## Research methodology

The systematic literature review (SLR)<sup>5</sup> has been conducted to complete the present study. A literature review is a comprehensive investigation of existing literature on work presented on a specific topic.<sup>6</sup> The existing literature has been searched systematically and existing techniques and methods have been investigated thoroughly to find the limitations of the work presented on IE from unstructured data. The literature search has been made for four main data types of unstructured big data such as text, images, audio, and video. The literature has been searched based on the keywords and search strings. A SLR is the most suitable approach to identify the limitations of the existing IE systems and to meet the objectives of the study. The literature review has been completed in three phases as shown in Figure 1.



**Figure 1.** Review process.

**Table 1.** SLR process and corresponding activities.

Phases	Activities
Phase 1: Planning the review	1.1 Formulate research questions 1.2 Data sources and search strings identification 1.3 Determine inclusion and exclusion criteria
Phase 2: Conducting the review	2.1 Selection of primary studies 2.2 Quality assessment of the selected studies 2.3 Data extraction and synthesis
Phase 3: Reporting the review	3.1 Reporting the outcomes of the review

SLR: systematic literature review.

**Table 2.** Research questions and objectives.

Research questions	Objectives
1 What are the techniques used for IE for different types of unstructured data?	To explore the state-of-the-art IE techniques for a variety of unstructured data (i.e. text, images, audio, and video)
2 What are the challenges multifaceted unstructured big data bring to IE techniques?	To investigate the challenges of IE techniques to deal with volume and variety of big data
3 How these challenges can be resolved?	To suggest the preconditions to improve IE process for multifaceted unstructured big data

IE: information extraction.

Table 1 describes these phases with the corresponding activities performed during the literature review.

### Planning the review

**Research questions.** The research questions with their corresponding objectives are given in Table 2.

**Search string and data sources.** The literature, selected for this study, has been referred from renowned databases of IEEE Xplore, Springer, ACM, ScienceDirect, Web of Science, Scopus, and Google Scholar using different search strings. The searched strings used for this review are as follows: “information extraction,” “information extraction from text,” “information extraction from images,”

“information extraction from audio and video,” “multimedia information extraction,” “information extraction AND big data,” “unstructured data problems,” “challenges in big data,” “multimedia data,” “information extraction techniques,” “unstructured data types,” “challenges in information extraction,” and “issues in big data analytics.” However, the selection of the articles was based on the inclusion and exclusion criteria for this study.

**Inclusion and exclusion criteria.** The inclusion criteria to select research studies for this SLR are as follows:

- Research published between 2010 and 2018.
- Studies related to IE from text, images, audio, and video data.
- Studies discussing the IE process implemented in the domain of big data.
- Studies presented on IE from unstructured data.

The exclusion criteria for this review are as follows:

- Studies other than unstructured data.
- Studies related to unstructured big data but not related to IE.
- Duplicate studies.

### Conducting the review

The most relevant studies to the research questions and objectives have been identified using the three-step process. At the first step, research studies were searched using different relevant search strings as described earlier. After searching the studies, 348 articles were selected based on their titles. Of these, 274 articles were selected based on the abstract. After scrutinizing 274 papers, only 95 relevant papers presenting the most significant and prominent work in this area were included in the review.

International journals in the areas of IE, multimedia data, data quality and extraction, biomedical, engineering data, and many other related domains have been consulted from IEEE, ACM, Springer, and other renowned databases. The journals referred in the present study include *Principles of Big Data*, *IEEE Transactions on Multimedia*, *International Journal of computer and Information Engineering*, *Journal of Data and Information Quality*, *ACM Transactions on Knowledge Discovery from Data*, *Analysis of Images, Social Networks and Texts*, *Journal of Biomedical Informatics*, *ACM Transactions on Intelligent Systems and Technology*, *Journal of Visual Communication and Image Representation*, *Computer Speech & Language*, *Mining Text Data*, *Multisource, Multilingual Information Extraction and Summarization*, *ACM Queue*, *Procedia Computer Science*, *Expert Systems with Applications*, *International Journal of Computer Science & Engineering Technology*, *International Journal of Computer Applications*, *Technological Forecasting and Social*

*Change*, *Computational Systems for Health & Sustainability*, *Journal of Cheminformatics*, *Certified International Journal of Engineering and Innovative Technology*, *Literature review for Language and Statistics*, *Semantic Technology*, *Advances in Computing*, *Database Systems for Advanced Applications*, *Journal of Physics*, *Smart Health*, *Computers & Electrical Engineering*, *Expert Systems with Applications*, *Towards Integrative Machine Learning and Knowledge Extraction*, *International Journal of Multimedia and Ubiquitous Engineering*, *Neurocomputing*, *International Journal of Computer Science & Engineering Survey*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, *International Journal of Scientific & Engineering Research*, *International Journal of Emerging Technology and Advanced Engineering*, *International Journal of Electronics, Electrical and Computational System*, *Computer Vision and Graphics*, *Video Text Detection*, *Semantic Applications*, *International Journal of Simulation: Systems, Science and Technology*, and *PLoS Med*.

The research critically reviews the limitations of methods and techniques for IE from unstructured data. The inclusion and exclusion criteria with different search strings have been strictly followed to select quality research for this literature review. The selected article has been thoroughly studied and analyzed critically to address the research objectives. The qualitative exploratory approach has been used to determine the impact of unstructured big data on IE systems. The study also explores the state-of-the-art techniques of IE from multifaceted unstructured big data. Figure 2 shows the research process flow for the SLR.

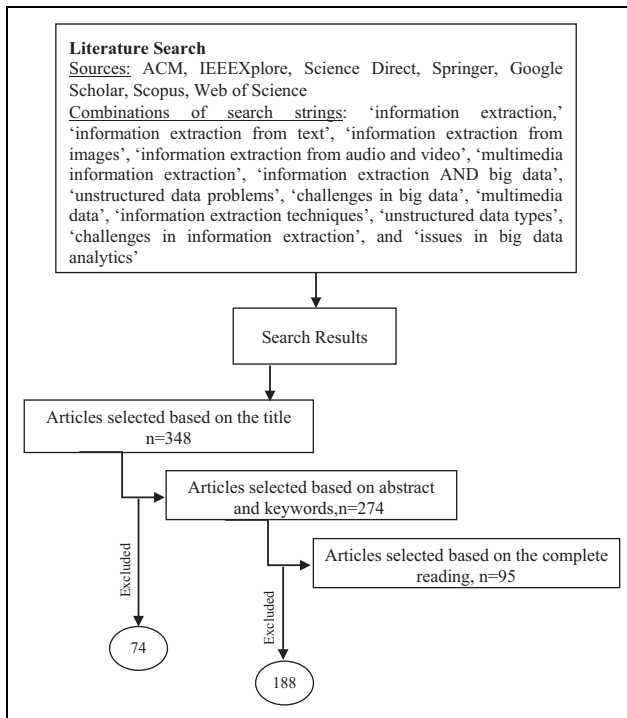
### Reporting the review

After selecting the most relevant literature, research studies have been divided into four categories with respect to data types, that is, IE from text data, IE from images, IE from audio data, and IE from visual data. Table 3 presents the distribution of selected research studies for each category, and illustrated in Figure 3.

The distinction of this literature review is presenting all four major unstructured data types about IE in a single review. The review identifies the limitations and shortcomings of the existing IE techniques for unstructured data which are necessary to understand to improve IE procedures. Eventually, the findings of this research would contribute onward to the usability enhancement of multifaceted unstructured big data to extract useful information from them.

### IE for unstructured data analysis

People and machines are producing data at a very high rate than ever before. The volume and variety of data being produced bring more challenges in identifying useful information from them. IE is a process to extract structured

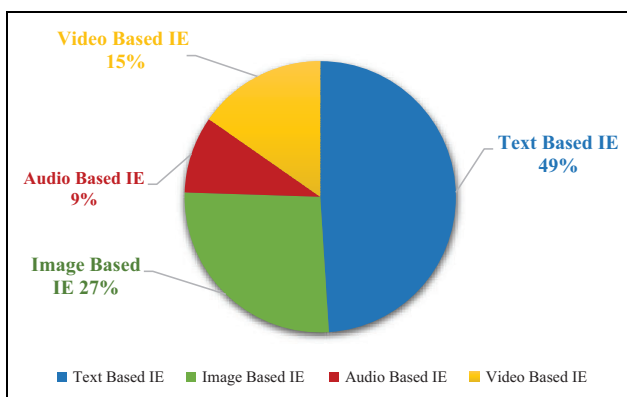


**Figure 2.** Literature selection process.

**Table 3.** Distribution of studies for each data type.

Categorization of selected articles based on data types	No. of selected articles
Text-based IE	46
IE from images	25
IE from audio data	9
IE from videos	15
Total	95

IE: information extraction.



**Figure 3.** Distribution of selected studies for each category.

information from unstructured data. As the growth rate of unstructured data is quite high as well as increasing during recent times, the significant challenges associated with IE,

mining, and analytics need to comprehend. Scalability, dimensionality, and heterogeneity of unstructured data appear as the main challenges to harvest useful information. Transformation of unstructured data into a structured format for better representation are the big questions.<sup>7</sup> Efficient and accurate transformation of unstructured data to structured data is necessary to improve analytics. Automatic IE from unstructured data is a field to identify new methods for extracting semantics and contextual information by analyzing, managing, and querying the data. This section presents IE techniques for unstructured data, that is, text, image, audio, and video formats, respectively.

### Text-based IE techniques

IE procedures emerged in the 1990s with Message Understanding Conferences where several types of IE tasks were introduced over time.<sup>8</sup> During the early stages of IE, the information from text has been extracted using template filling, rule-based methods, classification model-based methods, and sequential labeling. Conversion of an unstructured text document into a structured format has been divided into two subtasks such as define the high-level structure and low-level structure. The high-level structure includes major parts of a text document such as headings and title, whereas low-level structure determines the individual elements such as entities and places.<sup>9</sup> The discovery of the low-level structure is a more important and complex task in terms of understanding the content and context of any text document. Several dimensions for low-level structuring have been identified such as named entities, events, relations, and terms. NLP, machine learning, and computational linguistics are helping IE process to bridge the semantic gap and extracting and representing the relevant information. However, a huge volume of multifaceted unstructured data make IE process more challenging.

The IE process identifies and represents structured information from natural language text. Text strings, values, or tags extracted from the text are specified in prespecified slots of user-defined structures known as templates or objects.<sup>10</sup> Segmentation, classification, association, normalization, and de-duplication are five tasks for general IE and storage.<sup>11</sup> Similarly, another IE process was proposed for Web-originated unstructured text to present structured information in XML by following data extraction, syntactic and semantic analysis, classification, and inference rules steps.<sup>12</sup> Furthermore, machine translation, auto-coding, indexing, and term extraction are the main techniques to give meanings to unstructured data in IE process where auto-coding and indexing tasks help identify terms from text.<sup>13</sup> The sections “Named entity recognition,” “Relation extraction,” and “Event extraction” present the subtasks of IE process from text data whereas the section “Limitations of text-based IE techniques”

presents all the identified limitations of the IE techniques in above sections.

**Named entity recognition.** Named entity recognition (NER) is used to extract descriptive information from identified entities such as person names, locations, organizations, and numerical and currency expressions. Modern solutions to NER are based on statistical sequence labeling algorithms, for example, maximum entropy. Relation extraction (RE) deals with finding the semantic relations between entities from text. Existing methods use carefully designed features and standard classification to solve this problem.<sup>8</sup> Entities can be generic such as a person, location, or domain-specific like proteins, chemicals, and cells. Identification of entities (named entity detection) and their classification (semantic classification) are subtasks of NER.<sup>14</sup> Extraction models in NER systems use three techniques: rule-based, machine learning, and hybrid approaches.<sup>15</sup> Rule-based methods for NER use lexico-syntactic patterns and semantic constraints to identify the occurrence of similar entities while learning-based methods use machine learning to extract named entities and their classification. Learning-based methods can be supervised, unsupervised such as clustering (hard and soft), and semi-supervised such as bootstrapping. Supervised and unsupervised approaches use a large quantity of training data to achieve high performance but semi-supervised approaches use both labeled and unlabeled corpus with a small degree of supervision.<sup>16</sup> Several hybrid approaches achieve better performance and accuracy than a single approach such as rule-based pattern extractor using link grammar parser and Stanford PoS (Part-of-Speech) tagger with semi-supervised approach using self-training algorithm for entity labeling,<sup>17</sup> maximum entropy model (MaxEnt) with language-specific rules and gazetteers for the Hindi and Bengali language text,<sup>18</sup> ChemSpot: a chemical hybrid system with CRF (Conditional Random Fields) and chemIDplus dictionary,<sup>19</sup> SVM (Support Vector Machine) with CRF for biological entities with 91% accuracy,<sup>20</sup> and a semantic and statistical model for medical entity recognition with semantic method Meta-Map, chunker-based noun phrase extraction, SVM, and supervised learning CRF.<sup>21</sup> Supervised tagger TnT and rule-based SVM for health and tourism text documents in the Malayalam language has shown 73.42% accuracy.<sup>22</sup> Combination of HMM (Hidden Markov Model) with hand rules for the Punjabi language has shown pretty notable results as compared to a supervised machine learning method that achieved the precision of 72.92% and 47.57% by using HMM only.<sup>23</sup> A semi-hybrid approach by combining HMM and some rule-based approaches for PoS tagging and entity detection from the Nepali language has been proposed to extract named entity-specific classes that include the name of the person, location, number, organization, currency, and quantifier.<sup>24</sup> Combination of dictionary-based, rule-based, and machine learning has been used to extract molecules and related properties from

the scientific literature in biomedical domain.<sup>25</sup> Combined dictionary-based approach with fuzzy matching and stemmed matching is considered more helpful in finding information as it generates a new big set of annotation from the clinical text,<sup>26</sup> but these approaches are domain-dependent due to complex and short terms of the medical domain. A domain-independent hybrid approach has shown promising results by combining SVM and HMM with some simple linguistic pre-processing methods to identify gene and protein from the text without using external knowledge base.<sup>27</sup> In this regard, a hybrid named entity recognizer was applied with manual engineered rule-based predecessor combined with lexical resources and pattern bases for semantic indexing of the Turkish text.<sup>28</sup> Combination of HMM with gazetteer method has been outperformed with tourism text in Hindi as input. The improved accuracy of combined methods was 98.37%, whereas separately the results were 40.13% and 97.3% with a gazetteer and HMM, respectively.<sup>29</sup>

It is found that machine learning approaches the best suit for NER techniques for various Indian regional languages such as Hindi, Marathi, Bengali, Punjabi, Malayalam, Bengali, Kannada, Telugu, Tamil, Urdu, and Oriya, while HMM and CRF give best results considering their limitations.<sup>30</sup> IE from human language text is different for each language. But IE is easier for rich morphological languages like Russian and English. Sazali et al.<sup>31</sup> proposed an IE technique to extract nouns using morphological rules from classical documents in the Malay language. Extracting nouns from Malay classical documents is a difficult task because Malay is not morphologically rich language. In their study,<sup>31</sup> morphological rules are used to extract nouns from the Malay language but the results still need to be evaluated by the experts because there is no complete dictionary of nouns in the Malay language.

The issues identified from literature can be categorized into entity-specific and techniques-related challenges. Traditional NER techniques are inadequate to handle the dimensionality and heterogeneity of unstructured big data. Supervised learning techniques require large annotated data for training and that is a laborious and difficult task for large-scale data sets. Weakly supervised learning is effective as compared to supervised methods due to reduced manual effort. But still, the sparsity of data make these techniques inefficient.<sup>32</sup> In this regard, reinforcement learning for IE can overcome the limitations of the above techniques. On the other hand, open nature of vocabulary, abbreviations, disambiguation, and different languages and domains are major entity-specific challenges. Further, noise (short and domain-specific text),<sup>8</sup> entity ambiguities (single entity and global entity),<sup>33</sup> and automatic labeling<sup>34</sup> are creating difficulties in identifying entities and their relations from free text big data sets. Various factors have been identified during SLR that influence the performance of NER techniques such as noise, data diversity, variation in text perspective, and data sparsity as data-related

**Table 4.** Limitations of NER techniques.

Limitations	Factors influencing NER techniques	Studies
Data-related issues	Noise	8
	Data diversity	8,26
	Variation in text perspective	9,15
Entities-related issues	Entity ambiguities	16,27,33
	Automatic labeling	25,34
	Semantics of NE	14,21
	Contextual relationship among entities	25
Domain-related issues	Domain-specific entities	15,17,19,20
Task-related issues	Selection of NER technique	30
Language-related issues	Specific language	23,24,28
	Different languages	26,29
	Poor morphological languages	22,31

NER: named entity recognition.

challenges that influence the performance of NER techniques. Similarly, all identified factors are categorized according to data, entities, domain, task, and language-related limitations, and have been presented in Table 4.

**Relation extraction.** Finding the relation between entities is one of the substantial tasks in IE. The system requires to correctly annotate the data by recognizing a piece of text having the semantic property of interest. Various techniques have been applied to extract the relation between identified entities. Most commonly used techniques are knowledge-based methods, supervised methods, and semi-supervised methods. Supervised approaches use feature-based and kernel-based (bag of features kernel and tree kernel) techniques for RE and are suitable for domain-restricted RE. Semi-supervised approaches like Dual Iterative Pattern Relation Expansion, Snowball, and KnowItAll are suitable for open-domain systems.<sup>35</sup> These approaches are limited to sentence-level RE whereas RE in paragraphs and cross documents is a way to improve accuracy.<sup>36</sup> The hybrid approach provides a solution for linguistics to extract the relation between complex terms. A filtering algorithm to extract relation between complex terms of the Arabic language using deep linguistic analysis, morph syntactic and linguistic filters has been used as a hybrid method in which term extraction from the text in the Arabic language is combined with statistical approach.<sup>37</sup> Most relevant IE from largest biomedical database Medline has been achieved by a hybrid approach where semantic and probabilistic approaches have been used together to facilitate users in terms of searching and data representation.<sup>38</sup> Traditional feature-based methods combined with convolutional and recurrent neural network resulted in state-of-the-art performance for RE and classification.<sup>39</sup>

The major focus of open IE system is to extract maximum relations from text based on the contextual decomposition of sentences. Accuracy, minimalism, and coverage

are open challenges for open IE systems in RE with high precision.<sup>40</sup> An approach for open IE system based on lexical-syntactic patterns matching without using machine learning has also been used to extract domain-independent terms.<sup>41</sup> RE still needs improvement because of language ambiguity. Most of the research work, in this regard, has been conducted for the English language.

It is challenging to extract information from highly ambiguous languages, especially without diacritics. Semantic RE from the Arabic language is a complex task due to its ambiguity. A hybrid approach that combines statistical calculus with linguistic knowledge has been used as a two-stage process in which noun phrases are extracted first and then transformed into semantic relations with 60% decreased rate.<sup>42</sup> The performance of these approaches is evaluated using precision, recall, and *f*-measure. Precision and recall are the measures for completeness and correctness, respectively,<sup>10,43</sup> whereas F1-score (also known as *f*-measure or *f*-score) is used to measure the accuracy of a system. It is a harmonic combination of precision and recalls.<sup>43</sup> Extracting relations and their associative entities from radiology reports using unsupervised way as without specifying the knowledge base scored 0.94 F1-score in terms of accuracy. The proposed hybrid approach for rule-based IE systems based on distributional semantics and clustering to find similar relations outperforms other approaches<sup>44</sup> and is limited to one language. It has been observed that automatic annotation,<sup>45,46</sup> semantic RE with appropriate features<sup>45,47</sup> are critical factors that highly influence the results. Most of the RE techniques are extracting one to one relationship between entities whereas many to many relationships among entities can be identified also. In this regard, big data sets need high computational systems to increase performance efficiency and reduce computational delay. MapReduce was used to extract many to many relationships between entities of sized 100 GB free text and a proposed solution has outperformed as compared to existing approaches.<sup>48</sup> The scalability and sparsity of unstructured big data make traditional methods ineffective.<sup>32</sup> To overcome the limitation of existing traditional methods, distant supervised learning, CNN (Convolutional Neural Network), and transfer learning have shown pretty notable results.<sup>49–51</sup>

Based on the identified critical factors, limitations of RE techniques have been identified and presented in Table 5.

**Event extraction.** An event usually consists of trigger and arguments. A trigger is a verb or normalized verb that denotes the presence of an event in the text whereas arguments are usually entities which assign semantic roles to illustrate their influence toward event description.<sup>54</sup> There are different in-practice techniques for event extraction (EE) such as data-driven (focus on specific features such as words, n-grams, and weights), knowledge-driven (lexico-syntactic patterns and lexico-semantic patterns), and hybrid approaches.<sup>55</sup> Data-driven approaches require

**Table 5.** Limitations of RE techniques.

Limitations	Factors influencing RE	Studies
Data-related issues	Data sparsity	32
	Data dimensionality	32,49
	Volume	52,53
Language issues	Language ambiguity	35
	Lack of multilingual IE	35
Relationship identification issues	Domain-specific terms and relations	37
	Semantic relationship identification	39,42,45,47
	Errors in constituent parsing	40
	Large unlabeled corpus	35
Technical issues		

IE: information extraction; RE: relation extraction.

more data as input with less domain knowledge whereas knowledge-based methods require little data but high knowledge and expertise. Hybrid approaches are compromising techniques between these two approaches to minimize the effort and to improve the performance. High expertise is required to develop a hybrid approach.<sup>55</sup> Combining semantic knowledge and statistical learning method based on temporal and spatial elements for a semantic search engine as a hybrid approach has been used to extract event information from complex Chinese text.<sup>56</sup> Rule-based combined with feature-based classifiers and machine learning (SVM) approach has been used to detect event information in the three-stage model (pre-processing, trigger detection, event detection) which has shown 50.97% accuracy. Although the results were not satisfactory, however, post-processing and modified features could increase the accuracy.<sup>57</sup>

EE from social media is more challenging than formal news article because it contains complex text in terms of informal and short words, different languages and expressions. In evaluating the combination of linguistic rules with different machine learning techniques, the comparative performance of linguistic rules (for domain-independent annotated training set and a feature set) with decision tree and SVM has shown 52% and 75.8% accuracy, respectively. Proposed hybrid techniques outperformed by 24.8% for Twitter text.<sup>58</sup> EE from Twitter is an arduous task because of valuable user-generated text in the form of tweets to extract general and specific event information. EE techniques for Twitter can be based on event type, tasks, detection method, and features.<sup>59</sup> The ambiguity of representation, noisy data, and lack of training data are open challenges to EE from tweets. Tweet replies are also helpful to identify life events.<sup>60</sup> EE from unstructured data across multiple sources is more complex than EE from text only. Meaningful visual patterns help to identify semantic attributes, arguments, and concepts from different data modalities. In this regard, a system that could search, identify, organize, and summarize events from unstructured

**Table 6.** Factors influencing EE techniques.

Factors influencing EE	Studies
Semantic event modeling	56
Limitations of ML and rule-based techniques	57
Data sparsity	54,58
Multiple languages	59
User's perspective	62
Data from diverse sources	59
Representation ambiguity	60
Noise	60
Lack of training data	60

ML: machine learning; EE: event extraction.

Web data has been developed using online analytical processing. Adjustment of ranking and weights given to different dimensions have been used to derive meta-paths based on user browsing feedback. The identification and summarization of events for the recommendations according to the user browsing interest incorporated a medium-level human agency.<sup>61</sup> EE techniques help to improve the efficiency and accuracy of IE from the text, but still, the research is at the infancy stage. Unstructured big data add tremendous challenges to this research due to multimodality, heterogeneity, and complexity of data. These challenges have been presented in Table 6. Identification of events and summarization in unstructured data is a state-of-the-art challenge in IE.

**Limitations of text-based IE techniques.** More sophisticated algorithms and hybrid approaches (i.e. combining supervised and unsupervised) are required to achieve high accuracy and efficiency.<sup>63</sup> Self-training can reduce over-fitting issue,<sup>64</sup> and reinforcement learning or distant supervision can perform better with small labeled data sets.<sup>50,65</sup> According to the variety and volume of big data, many questions still need to be answered using deep learning such as more resources of information will increase conflicts, hence, conflict resolution became important. How to handle timeliness and data distributions, the impact of enlarged modalities on performance,<sup>65</sup> balance among informativeness, representativeness, and diversity,<sup>66</sup> modeling performance in case of heterogeneous, dimensional, sparse and imbalance data,<sup>52</sup> and structuring the data<sup>67</sup> are the challenging issues of IE from large-scale unstructured data sets.

The most critical factors as presented in Tables 4 to 6 having strong influence on the performance of NER, RE, and EE techniques, respectively, have been combined and categorized as overall limitations of text-based IE techniques in Table 7.

### **Image-based IE—From visual to semantic extraction**

Global digitalization and social media gave exponential rise to image sharing. IE from images can be achieved by extracting elements, objects, visual concepts (low and high-



**Table 7.** Limitations of text-based IE techniques.

Limitations	Influencing factors	Studies
Unstructured data issues	Noisy data	8,60
	Data diversity	8,26
	Variation in text perspective	9,15,60,64
	Data ambiguities	16,27,33
	Data sparsity	32,54,58
	Data from diverse sources	59
	Data modeling	52,56
	Volume	52,53
Data understanding	Data dimensionality	32,49
	Semantic understanding	14,21,39,42,45,47,56
	Context Understanding	25
	Relevance from user's perspective	62
Linguistic issues	Lack of multilingual system	23,24,26,28,29,35,59
	Poor morphological languages	22,31
	Language ambiguity	35
Domain specificity	Domain-specific text	15,17,19,20,37
Technical issues	Selection of technique	30
	Lack of large labeled corpus	35
	Limitations of ML and rule-based techniques	57
	Automatic labeling	25,34

IE: information extraction; ML: machine learning.

level features), and shapes. Visual features are used to detect objects, entropy-based analysis of visual and geo-location data, structural decomposition of 3-D images, whereas data-driven approaches are used to extract a meaningful representation of facial expressions, classification, and segmentation. The IE from images is a field with great opportunities and challenges such as extracting linguistic descriptions, semantic, visual and tag features, improved scalability, and precision. Content and context-level IE from images could improve image analytics, mining, and processing. Visual relationship detection, text or face recognition from images provides contextual and useful information. The common applications of visual IE are content-based information retrieval,<sup>68</sup> visual question answering,<sup>69</sup> sentence to image retrieval,<sup>70</sup> and fine-grained recognition.<sup>71</sup> The sections “Feature extraction,” “Character recognition and text extraction,” “Dynamic scene understanding,” and “Semantic and geospatial information extraction” review the IE process from images based on extracted structure type and method used for IE whereas the section “Limitations of image-based IE techniques” presents the limitations of image-based IE techniques.

**Feature extraction.** Feature extraction and its representation is an important step to process unstructured data. Low-level and high-level visual features are used to extract semantic information from images. Visual feature extraction is used to identify the unique objects in the image. The scene in images may contain useful objects in the form of logos and signs. Such objects detection can be helpful for content-

based searching, targeted advertisements, and social network applications. Feature selection and its classification are the steps to extract features from images. Automatically detecting targets in stationary images using segmentation and SIFT (scale-invariant feature transform) for feature extraction has shown average classification accuracy up to 90.99%.<sup>72</sup> Automatic IE from HR/VHR (high resolution) remote sensing images using classification rules and correlation using prior knowledge instead of local training data set for features extraction contributes to improving the scalability.<sup>73</sup> Feature extraction is important to identify any object or target which describes information about the image. Some of the important features that can be extracted from images have been discussed as follows.

**Color features.** Several color features have been proposed like color histogram, color moments, color coherence vector, and color correlogram. The histogram is most commonly used due to its simplicity toward computation.<sup>74</sup> In a multimodal learning approach for tagging and visual feature extraction based on deep learning, the popularity of image on social media is predicted by extracting tag features using sparse word count vector and visual features using real-valued dense vector.<sup>75</sup> Descriptive visual words extraction and Hue Saturation Value color features are combined to utilize the tags of social media images to find a correlation. The user-image-tag model has been developed with the tripartite graph according to the correlation among users, images, and top-ranking tags.<sup>76</sup> For HR satellite images, a data-driven algorithm has been designed to extract water information to determine the optimal threshold of water boundary using water index and color features.<sup>77</sup> The segmentation approach was based on the Markov random fields model.

**Texture features.** According to general observation, the visual system of human beings uses texture to recognize and interpret. Texture features are different from color features as it deals with the group of pixels. Spatial and spectral feature extraction are two major texture feature extraction methods used in several domains.<sup>74</sup> Texture information could be extracted from HR satellite images using object-oriented classification approaches based on spectral and spatial heterogeneity in segmentation. It is found that the quality of segmentation is directly affected by classification accuracy when the proved accuracy of classification is 85.16%.<sup>78</sup> Edge detection, only, is not sufficient to identify an object in the image, but texture information also plays an important role. Edge detection and texture IE together have been used to extract smooth regions with poison noises in image.<sup>79</sup>

**Shape features.** The purpose of shape feature extraction is to identify objects using two main methods, that is, contour-based and region-based methods. Unsupervised change detection and multi-temporal Red Green Blue visualization are important tools to detect and visualize the



images of different fields. Traditional extraction processes based on edge detection and template matching were not performing well when the image has noise. A combination of spectral reflectance, texture measures, and shape features could be used for object extraction based on multi-scale segmentation from geographic satellite images to improve classification accuracy,<sup>75</sup> but the classification accuracy is achieved only for fine images. Image segmentation algorithm to create an image object, object-based feature extraction, and classification comprises of three steps of a hybrid approach combining object-based classification technique and multi-resolution segmentation. Scales, weighted input image layers, shape ration, and compactness ratio have been used as segmentation parameters whereas classification calculation based on spectral, texture, and geometric (such as shape index and length/width) features. Classification combined with multi-resolution segmentation showed higher accuracy as compared to object-oriented classification method.<sup>80</sup>

**Character recognition and text extraction.** A vast array of information is extracted from the content in images. In text extraction task, the text is segmented from the background for recognition and converted into a binary image. Text extraction from images can be divided into three subtasks: detection and localization, text enhancement and segmentation, and optical character recognition (OCR). But the noise, variation in font size, style, orientation, text alignment, illumination change, and complex background are adding complexities to this task.<sup>81</sup> Morphological operators, wavelet transform, artificial neural network (ANN), and histogram techniques are mostly used for text extraction,<sup>82</sup> whereas OCR is most commonly used for character recognition. The accuracy of character recognition using OCR for the Brahmi language was 91.57% and 89.75% for the Vattezhuthu.<sup>83</sup> Regions in images are recognized, classified, and converted into text with the help of classifier-based automatic IE system and ANN. ANN and a correlation algorithm were used to improve efficiency and effectiveness.<sup>84</sup> OCR combined with CNN improved the performance of visual IE from large data sets.<sup>85</sup> On the other hand, text imprints, kind of noise, make useful information unavailable. In this regard, locating the imprint location first and then applying a noise elimination technique to extract misprinted text in binary images were two main steps in IE from text imprints. Experimental results showed that Otsu's thresholding with noise elimination performed better than the K-mean clustering method.<sup>86</sup> Hence, the advantages of IE from images are efficiency, less complexity, and less time-consuming but when the image is noisy, one cannot take advantage without noise removal before IE.<sup>87</sup> In this manner, attention mechanism is the latest solution these days which uses encoder and decoder to detect, extract, and recognize the text for images.<sup>88</sup> There is a big room of improvement in these attention mechanism systems for large-scale high-

dimensional data sets. Text extraction from images has several difficulties and challenges in terms of detection and identification of text in images. Text in different languages makes this task more challenging. A single unified model to extract text from digital images for all applications is a robust task as there is no single unified model available.<sup>82</sup>

**Dynamic scene understanding.** Contextual and semantic scene understanding are two paradigms in dynamic scene understanding. Hierarchical strategies of scene understanding are top-down, bottom-up, and combined methods, while others are nonhierarchical.<sup>89</sup> In terms of scene understanding, the composition of objects and their surrounding environment plays an important role to extract some useful information from the scene. Top-down and bottom-up processing approaches are used to detect objects and features from the image. Bottom-up approach extracts visual features using the deep conventional neural network. Top-down approach converts detected images into bag-of-words feature space and then combines with visual features. SVM classifiers improve the qualitative accuracy in context understanding between object and natural scene.<sup>90</sup> Closely relevant and important feature extraction is a challenging task in dynamic scene understanding. Type and position of images, scene motion, illumination changes, static and dynamic occlusions, type speed and pose of objects, camera synchronization and handover, event complexity, and handling dynamic scenes are adding more challenges to feature extraction.<sup>89</sup>

**Semantic and geospatial IE.** Some images also contain geo-tags in metadata to represent a location in a pair of values for latitude and longitude. Low-level visual features and thematic classification methods are not adequate to identify and extract complex objects and their semantic and spatial relationship. Feature extraction, vector quantization, and latent Dirichlet allocation, however, are used for semantic labeling of a large collection of images in a semantic model. K-means and Gaussian mixture model (GMM) clustering have shown 73% classification accuracy for semantic labeling.<sup>91</sup> Structure-free document images can be represented by graphs where nodes represent dynamic semantics and edges as the attributes with spatial information.<sup>92</sup> GPS parameters are used to capture the objects to manage digital images and map databases for a semantic spatial IE system.<sup>93</sup> These parameters, however, have the limitation of nearby objects and require comparative analysis to prove the efficiency and effectiveness of the approach. Spatial features and semantic RE could also be beneficial for the content-based representation of images with the help of human-computer interaction to identify and recognize complex objects.<sup>94</sup>

**Limitations of image-based IE techniques.** A vast array of information can be extracted from images because images contain visual description of entities, events, and

**Table 8.** Factors influencing the performance of IE techniques for images.

	Influencing factors	Studies
Unstructured data issues	Noise	72,81,86,87
	Consistent and fast IE at large scale	73
	Low resolution	81,82
	Variation in text representations	82
	User's relevant content	92
Semantics understanding	Data sparsity	95,96
	Semantic feature extraction	74,91
	Semantic gap due to user's preferences	76,93
	Selection of optimal features	80
	Variation in text representation	81
Linguistic issues	Dynamic scene understanding	89,90
	Specific languages: Arabic OCR	85
Techniques-related issues	Quality of image segmentation	78

IE: information extraction; OCR: optical character recognition.

relationships. Although images are rich container of information, certain challenges are also associated with IE from images. User-generated content on social media have variations in quality,<sup>72,81,86,87</sup> resolution,<sup>81,82</sup> and information representations.<sup>82</sup> Extracting useful information from these user-generated images is helpful as well as challenging. Data sparsity,<sup>95,96</sup> extracting relevant information from user's perspective,<sup>76,92,93</sup> semantic understanding,<sup>74,80,89–91</sup> language understating, object detection, and recognition are major challenges identified in this field. Furthermore, many most critical factors influencing the performance of IE process and techniques from images have been presented in Table 8.

### Audio content IE techniques

Companies like call centers and music industry are the major sources which generate a huge volume of audio data. Speech recognition is a process to automatically recognize spoken words. Automatic speech recognition (ASR) is mostly used to recognize speech and convert it into any other medium such as text, also known as speech to text conversion. ASR system is based on four types of speeches such as isolated, connected word, continuous, and spontaneous speech. Extraction of speech features, acoustic modeling, and recognition of words are the main steps of this process. For the first task feature extraction, acoustic feature extraction compiles a feature vector and transforms it into a compact vector. Unnecessary and redundant information is removed in this process to extract useful information. Several feature extraction techniques are linear prediction coding, mel frequency cepstral coefficient (MFCC), linear prediction cepstral coefficient, discrete wavelet transform (DWT), wavelet packet decomposition,

perceptual linear prediction, and linear discriminant analysis.<sup>97</sup> Metrical structure extraction explicitly extracts rhythm-related information from music using mid-level features instead of using low-level or high-level features with 93% accuracy as compared to baseline methods. Several music applications are using mid-level features for IE such as automatic sequencing, database navigation, mash-up generation, and complement systems.<sup>98</sup>

Speech recognition techniques are categorized into five main classes as the acoustic-phonetic approach, a pattern recognition approach, template-based approach, knowledge-based approach, and an artificial intelligence approach.<sup>99</sup> ANN-based approaches are followed in most of the research studies because these approaches can handle complex interactions and are easier to use as compared to statistical methods. A framework for event detection using SVM and neural network approaches has been proposed<sup>100</sup> but it shows only reasonable performance using the combination of these approaches for some events. Sound EE or acoustic EE is an emerging field which aims to process the continuous acoustic signals and convert them into the symbolic description. HMM frameworks have been utilized for ASR for more than 30 years and facilitating speech and language resources of big data. HMM performance is better in modeling the speech signals in ASR, and SVM classification accuracy was higher than others. HMM and SVM hybrid approach could combine the capabilities of both for speech recognition. MFDWC (Mel-Frequency Discrete Wavelet Coefficients) method is also a hybrid approach that combines MFCC techniques and DWT to increase the robustness.<sup>101</sup>

Speech fusion and recognition system could be improved through transcription correction for automatic linguistic IE from the Amharic, Hindi, and Tamil sounds using the letter to sound rule.<sup>102</sup> Automatic integrated detection and recognition technique is required to extract information from speech, for the useful analysis of speech, speaker identification, speech, and language recognition.<sup>103</sup> Semantic IE from audio is capable to extract music score and text information through classification and segmentation which are helpful to update and insert music or speech occurrence and analyze arbitrary soundtracks.<sup>104</sup> Recently, the exponential growth of unstructured big data and computational power, ASR is moving toward more advanced and challenging applications such as mobile interaction with voice, voice control in smart systems, and communicative assistance.<sup>105</sup>

**Limitations of IE techniques for audio content.** The field of acoustic IE is facing challenges such as more accurate feature selection,<sup>97,99</sup> classification of nonexclusive sound and content overlapping.<sup>104</sup> Call centers use audio data for analysis and IE in the form of conversations with clients, music, monitoring, and processing of conversations. Background noise, words overlapping, considering one single voice in crowd, and language ambiguities are open challenges in this domain. The critical factors representing

**Table 9.** Limitations of IE techniques for audio contents.

Limitations	Influencing factors	Studies
Unstructured data issues	Metrical structure	98
	Noise (overlapping sounds)	104
Acoustic feature extraction	Feature extraction	97,99
	Acoustic features	103
Linguistic issues	Language modeling	101
	Language knowledge	102,103
	Language understanding	105
	Cross languages	105
Semantic understanding	Semantic understanding	106
Technical issues	Weakly labeled data	100

IE: information extraction.

limitations of audio-based IE techniques have been presented in Table 9.

### **Audiovisual IE—Exploitation of unstructured video content**

The massive growth of video content on the Web and social media increases the demand to extract efficient, reliable, and valid information. Video summarization is a process that provides a condensed and concise summary of the video content to facilitate the users. Generally, video summaries are categorized into two groups which are keyframe-based video summarization (static video summarization) and video skimmed-based video summarization (dynamic video summarization). In keyframe video summarization, keyframe information is extracted from a video where frame extraction and feature extraction are the pre-steps. The video is divided into frames in the first step of keyframe extraction. The number of frames depends upon the size of the video. Normally the frame rate varies from 20 to 30 frames per second. Feature extraction is the second step where several visual and audio features are extracted.<sup>107</sup> Visual feature extraction through phonetic and viremic information for audiovisual speech recognition converts speech to text and vice versa.<sup>108</sup> Several methods are used in keyframe video summarization such as video summarization by clustering using Euclidean distance, perceived motion energy model, visual frame descriptor, motion attention model, multiple visual descriptor features, motion focusing, camera motion, and object motion. These methods are categorized based on semantic features and visual descriptors. As compared to static video summarization, skimmed-based summarization supports object recognition and its representativeness summarizes the video by replacing the original content.<sup>109</sup> Video skimming segments the video into smaller parts with a short duration like a movie trailer. The classification according to the static summary, dynamic summary, fixed camera, with and without knowledge about content provides information about the content. However, the selection of a method for video summarization depends on its application.<sup>110</sup>

Generally, the content extraction is divided into perceptual and semantic content extraction. Perceptual content includes attribute like color, intensity, and so on, whereas semantic content includes visual objects, events, their relationships, and so on.<sup>111</sup> To overcome the semantic gap between visual appearance and semantics, spatial and temporal association between objects and events, respectively, are identified using fuzzy ontology and rule-based model.<sup>112</sup> The proposed system achieved high precision but relatively low recall. Similarly, EE from audiovisual content consisting of CNN-based audiovisual multimodal recognition was developed and incorporated knowledge from a website using HHMM (Hierarchical Hidden Markov Model) was used to improve the efficiency. The proposed approach outperformed in terms of accuracy and concluded that CNN provides noise and occlusion robustness.<sup>113</sup>

IE from a video is the most complex task because it involves audio and video data and their synchronization. A video summarization system with face recognition is designed which decomposes facial areas using nonnegative matrix factorization and the coefficient for classification using SVM with the GMM-SVM approach, used for speaker identification.<sup>114</sup> This approach gives efficient results but performs better for one target like either face or voice. Different quality parameters have been used in temporal IE for each frame and two consecutive frames, to assess the quality of the video,<sup>115</sup> but it is limited to the video where scenes do not change frequently. Useful IE from dramatic video comprises of three modules such as face processing (face detection and recognition in the video), interaction score computing (based on interaction graphs and phonogram), and scenario IE (includes main character identification, video clip extraction for the selected characters, and visual graph construction).<sup>116</sup> Automatic subtitle generation is a process to facilitate users to understand the content of video more easily. This process includes sound extraction, ASR, and text synchronization with video.<sup>117</sup>

**Limitations of IE techniques for video content.** In the age of big data, digital videos are spreading all over the Internet at very high speed. It is not only about the size but also requires high processing power to extract useful information from this huge deluge of video data. These fast processing systems are essential for applications like crime investigation. In this regard, an extensible video processing framework was designed using Apache Hadoop to distribute processing tasks in cloud environment.<sup>118</sup> FFmpeg was employed for video coder and OpenCV was used for image processing followed by MapReduce implementation. In result, the system had shown 75% scalability. Although, the system had not shown high scalability but the findings if the research paved the way for improvements in IE from unstructured video content in big data. Furthermore, language and accent barrier, speech to text and vice versa conversion,<sup>108,113</sup> automatic subtitling and labeling for

**Table 10.** Limitations of video-based IE techniques.

Limitations	Influencing factors	Studies
Unstructured data issues	Quality of user-generated video	115,119,120
	Data diversity	121
	Data dimensionality	122
Semantics and context understanding	Personalized video summarization	107,110
	High-level semantic information from low-level audio or visual data	109
	Content-based indexing	111
	Semantic labeling	111
	Gap between context and visual object	112
	Visual features combined with acoustic features	108,113
	Face recognition	116

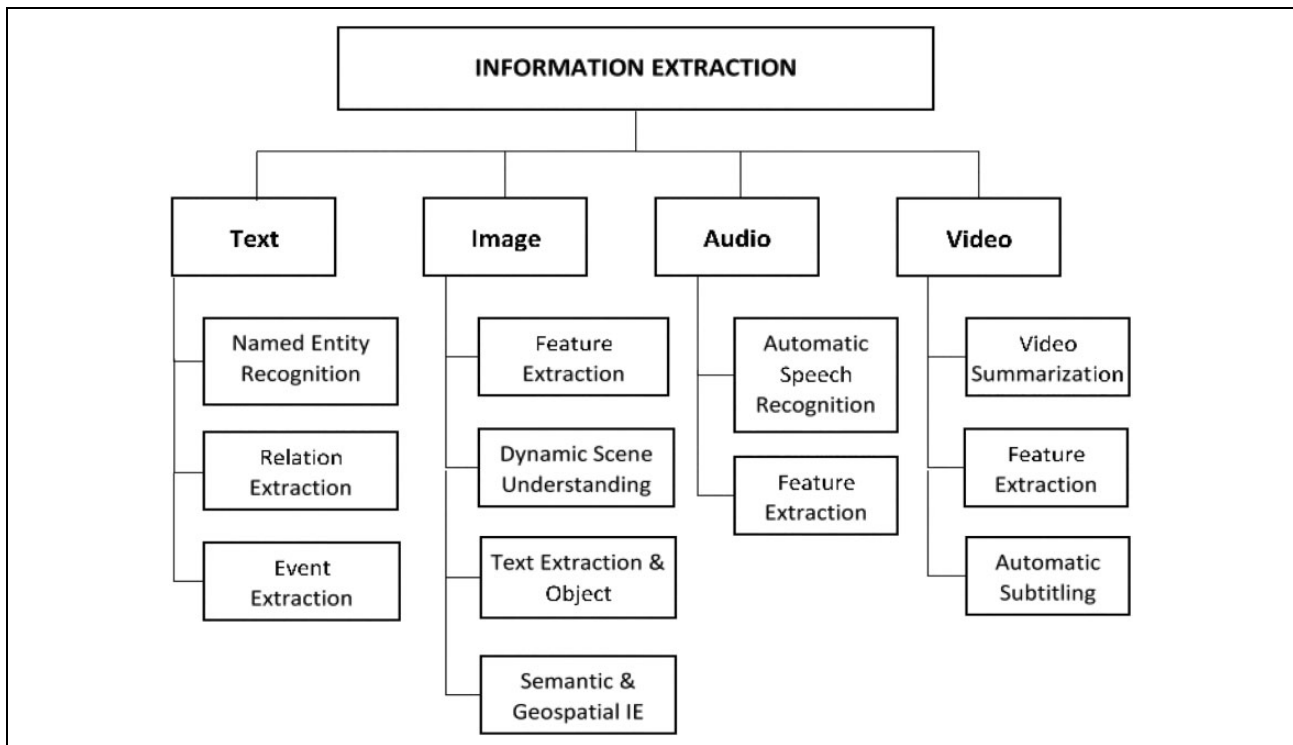
IE: information extraction.

different languages,<sup>111</sup> and noise elimination<sup>115,119,120</sup> are important challenges to IE from video content. The critical issues for IE from audiovisual content have been identified and presented in Table 10.

## Results and discussion

The review presented in this article comprehensively investigates the IE methods for different types of unstructured data such as text, image, audio, and video. It is found that

various IE methods have been used in the context of different domains and data types. Accordingly, the IE methods have been classified in the present study based on their main IE application and technique. Figure 4 shows the major classifications of IE methods used for text, images, audio, and video data. Each data type requires different methods and techniques for IE and has associated challenges and problems. The use of different techniques and methods, their varying applications in different domains and different languages have made IE procedures more complicated. It is hard to determine a standard approach for various domains and different formats of the same data type. It is evident from the study that entities, relations, and events extracted from the text use different techniques in different languages and domains. Mainly, the IE techniques on images have been applied to extract features such as color, shape and texture, scene understanding, semantic and geospatial information, character recognition, and text extraction. These techniques include several supervised, unsupervised, semi-supervised, and hybrid methods. ASR and acoustic feature extraction methods are mainly used to extract useful information from auditory data while keyframe-based and skimming-based methods are popular for video summarization for automatic video subtitling and audiovisual features. The complexities of existing IE techniques and lack of a standard technique for various formats of the same data type and domain are among the biggest challenges in IE. The unstructured data analytics and decision-making can be further improved with the help of IE.

**Figure 4.** Classification of IE techniques. IE: information extraction.

Industries and enterprises are facing challenges in finding the right information at the right time from a huge heap of data. The solution is to identify and introduce new methods to handle these challenges rather than falling back on the “drinking from a fire hose” approach, where a huge volume of unstructured data are generated every day but very less data are analyzed properly. Big data value chain illustrates the high-level activities as a series of steps required to generate value and useful information from big data where IE is concerned with the analysis. It is important to acquire raw data and transform them into useful information, but it is not well explored in literature in the context of IE. In the age of unstructured data deluge, it is necessary to make data understandable and available for analysis. IE is very helpful in this regard, but the need is to develop the advanced IE systems to facilitate analysis and mining process, and to extract useful information from different types of unstructured data. The SLR presented in this article aims to explore the limitations of existing IE techniques and identifying the improvement activities. Tables 7 to 10 present the limitations of IE techniques with respect to each data type and IE subtask, whereas Table 11 summarizes all the common and task-independent critical factors that have significant impact on the IE process due to unstructured big data.

The following subsections present the limitations of IE techniques in detail. Furthermore, some preconditions have also been proposed by critically analyzing these limitations for multifaceted unstructured big data.

### Limitations of existing IE techniques for unstructured data analytics

**Unstructured big data issues.** A large number of quality issues have been found in this SLR while extracting information from variety of data types. These issues are categorized as data quality, data sparsity, dimensionality, diversity, and modeling complexity. Among these issues, data sparsity, dimensionality, and diversity are more related to the data quality issues whereas the modeling complexity of unstructured big data is related to extracting and representing the structured information. The following subsections provide detailed discussion on these challenges respectively.

- a. **Data quality issues:** Noise in data creates problems in IE process. Noise elimination method as pre-processing step can improve IE but the selection of appropriate and efficient noise elimination technique for the specific problem is challenging. Unstructured data come with inherited problems of quality and incompleteness<sup>123</sup> that lead to inefficient and poor results from IE. The variety, accuracy, scalability, security, and interactivity are some of the prevalent challenges generated by unstructured data.<sup>124</sup> The problems of unstructured data are a huge barrier in deriving useful

**Table 11.** Limitations of IE techniques for multifaceted unstructured big data.

Limitations	Critical factors	Studies
Unstructured big data issues	Noise	8,60,72,81,86,87,104
	Data quality	81,82,115,119,120
	Data diversity	8,26,121
	Data dimensionality	32,49,122
	Data sparsity	32,54,58,95,96
	Data modeling	52,56,98
Unstructured data usability	Variation in perspective	9,15,60,64,81,82
	Data ambiguities	16,27,33
	Semantic understanding	14,21,39,42,45,47,56,106,108,109,113
	Context understanding	25,112
	Relevance from user's perspective	62,76,92,93,107,110
	Lack of multilingual system	23,24,26,28,29,35,59,105
Language and domain issues	Poor morphological languages	22,31
	Language ambiguity	35
	Language modeling	101
	Language knowledge	102,103
	Language understanding	105
	Domain-specific data	15,17,19,20,37
Capability issues	Volume of unstructured big data	52,53,73
	Optimal feature extraction and selection	74,80,91,97,99,103
	Automatic and semantic labeling	25,34,111
	Lack of large labeled corpus	35,100
	Limitations of ML and rule-based techniques	57
	Selection of technique	30

IE: information extraction; ML: machine learning.

information because unstructured data are noisy<sup>125</sup> and dirty (inaccurate, improper, and incomplete).<sup>126</sup> (62) Unstructured data are unverified by nature and face quality issues,<sup>127</sup> scalability and heterogeneity,<sup>128</sup> which make IE more arduous. Advanced data pre-processing techniques before IE are required to handle these quality issues of unstructured data. Standardization of data and processes, efficient tools for data cleaning, and high-level data management are critical

challenges that must be addressed to improve the IE systems.

- b. **Data modeling complexity:** Social media and smart sensors are generating a huge volume of unstructured data where streams of data are coming at a very fast rate. To analyze such fast streams of unstructured data requires very efficient IE systems to facilitate data mining and analysis techniques. The major problem for IE systems is to extract structure from unstructured data and its interpretability. High dynamicity of unstructured data as compared to structured data makes IE challenging especially for streaming data. Structural variation, different granularity levels, heterogeneity, and differences in data formats are critical challenges associated with unstructured big data.<sup>123</sup> The complicated heterogeneity of mixed data makes it difficult to analyze and extract useful information. Transformation of unstructured data into useful information is required to design the IE system for unstructured data. Unstructured data are growing very fast and to extract useful information requires to precisely specify the tasks of the IE process with efficient handling of interoperability and context understanding.<sup>129</sup>

**Unstructured data usability.** People and machines are producing data at a very high rate than ever before. The volume and variety of data being produced bring more challenges in identifying useful information from them.<sup>52,53,73</sup> Variety of big data indicated the heterogeneity of data types, different representations, and complex semantic interpretation. Usability and usefulness of unstructured big data are important dimensions for extracting useful information from variety of data types. IE process is community process that highly depends on the consumer requirements. Therefore, it is necessary to understand the user's requirements and semantics of data to extract most relevant information. Understanding the perspective is an important task where variation in user's requirements and data representations make it more complex. Understanding and handling the user requirements in IE process would improve the effectiveness of big data analysis as it directly targets the user needs. Semantic<sup>14,21,39,42,45,47,56,106,108,109,113</sup> and context understanding<sup>25,112</sup> of content are important to extract relevant information from data. It will lead to reduce the content gap. Hence, IE systems need improvement to understand the user needs, extracting the relevant data, and interpret the results in semantically and contextually rich manner.

**Domain and language issues.** It has been identified that language and domain specificity are two huge barriers in extracting useful information. These require highly

efficient and specific IE techniques that can handle the complexity of these barriers. The following subsections discuss these limitations in detail.

- a. **Lack of multilingual systems:** The complexity of natural languages is a huge barrier to IE, however, language-independent feature-based IE has been introduced to overcome it. This open multilingual IE tool and machine translation show pretty good results for the English language as compared to other languages.<sup>130</sup> These solutions are still facing language ambiguities issues, lexical and structural gaps, and grammatical issues. In this regard, rich morphological languages like English and Russian show more accurate results. Open nature of vocabulary, abbreviations, disambiguation, and specific dictionaries of domains are major challenges in general IE system for different language-independent features.
- b. **Domain specificity:** What are the steps to extract information from unstructured big data? Which techniques will play an important role in this regard? How to make data easily available for analysis? These are some questions that must be answered. These questions have different answers in different domains. The efficiency of IE techniques is highly dependent on the domain and language of unstructured text. In this regard, text in the clinical domain is different from business or other domain's text. IE from clinical text is semantically and syntactically more challenging as compared to other fields due to highly ambiguous abbreviations and acronyms. Medical terms and entities are different from other domains that make open nature of vocabulary, abbreviations, disambiguation as serious issues that need to be considered. So, IE system for the medical domain would not provide efficient and effective results for other domains.<sup>8</sup> Domain-independent solution for IE systems is one of the biggest challenging tasks

**Capability issues.** Variety of techniques for IE subtasks of each data types have been discussed in this article. These techniques can be categorized into two major IE approaches, that is, rule-based approaches and machine learning-based approaches. Hybrid of these two have also been discussed in this review. Rule-based approaches use rule languages like UIMA Ruta and dominating in industry whereas advanced learning-based approaches like distant supervision and reinforcement learning have the capabilities to overcome the limitations of traditional learning-based approaches in terms of unstructured big data. In short, rule-based and learning-based approaches in IE have their own potentials and limitations. But unstructured big

data bring more challenges to these techniques such as scalability, automatic semantic labeling, selection of appropriate techniques for the task and requirements of user, data annotation.<sup>25,30,34,35,57,100,111</sup> Hence, the emergence of advanced learning-based approaches with rule-based will improve the performance of IE systems for the huge volume and variety of big data.

- a. Optimal feature extraction and selection: Feature extraction and transformation from unstructured data are more critical for data analysis as compared to structured data due to the heterogeneity and multidimensionality of unstructured documents. Features like bag-of-words, orthographic features, lexical features, and gazetteer-related features can be extracted from the text for learning-based approaches<sup>130</sup> that improves the data analysis process. In this regard, a hybrid feature transformation technique based on iterative classification with feature weighing has been proposed for multiple domains.<sup>131</sup> Although feature transformation from heterogeneous unstructured data was achieved, a minimal loss of precision was observed. Feature extraction and transformation need advanced data preparation techniques. These techniques will help to improve the pre-processing and feature extraction tasks of heterogeneous, diverse, and multidimensional unstructured data. IE from unstructured clinical notes that contains inconsistent abbreviations and lack of structure can be achieved using matrix factorization, and multi-view learning technique in pre-processing and data modeling to handle the heterogeneous data.<sup>132</sup> While extracting features and pre-processing unstructured content, interpretability is an open quality dimension that should be considered.<sup>133</sup> Selecting the most relevant features from unstructured big data is a challenging task<sup>134</sup> that can be achieved with the help of well-defined IE systems to improve the unstructured big data analysis.
- b. Selection of technique: Many significant challenges are associated with a variety of data in terms of IE. Most importantly, data are not available in the form ready for analysis, they have to be transformed and prepared for analysis. Rather we need to identify the right IE process to pull out the required useful information from heterogeneous sources. The accuracy of the selection is a continuous challenge. Unstructured big data are not only available in the form of text, but they also include images, audio, video, animations, sequence data, and many other forms. In this scenario, IE technique selection is subjective toward its application. It is notable here that the extraction method would be different for ECG and

satellite image used for forecasting. It has been observed that IE from unstructured big data faces challenges like high dimensionality, complexity in relations identification, dynamic structures, heterogeneous distribution, and scalability. Hence, advanced IE systems, to handle a variety of data types, are the ultimate need to improve the efficacy of unstructured data analysis.

- c. Volume of unstructured big data: Extracting the useful information from huge volume of data is making IE task more complex.<sup>52,53,73</sup> The existing techniques are inadequate to handle this huge volume of data in terms of time and cost efficiency. Parallel computing and distribution using advanced tools like Apache Hadoop and Spark have capabilities to increase the time efficiency and accuracy.<sup>118</sup> Therefore, more advanced techniques are required to handle the volume of unstructured big data efficiently.

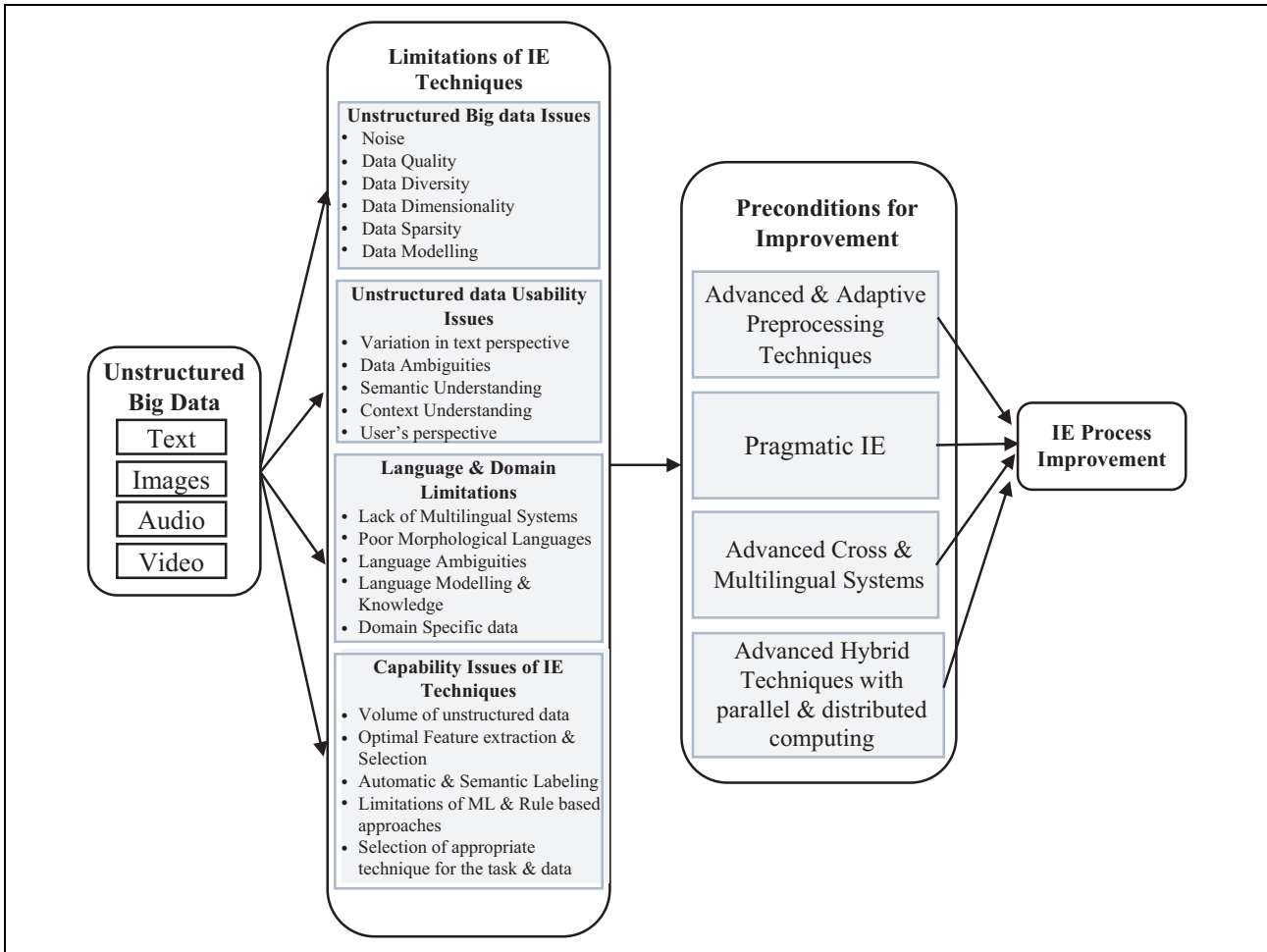
### *Preconditions to improve unstructured big data analytics*

For big data analytics, several IE approaches can be used such as statistical, machine learning, and rule-based, but interpretability, simplicity, accuracy, speed, and scalability are important characteristics that should be considered while selecting an appropriate approach for the solution.<sup>135</sup> Heterogeneity, scale, timeliness, complexity, and privacy are important challenges of big data analysis pipeline<sup>128</sup> where heterogeneity, timeliness, and complexity are more relevant to the IE from unstructured big data. Accuracy, coverage, and scalability are challenges to big data IE, whereby, accuracy and coverage are particular to IE and scalability is related to big data.<sup>136</sup> The following preconditions have been proposed to improve the efficiency of unstructured data analysis.

**Precondition 1: Advanced data pre-processing.** Advanced data pre-processing techniques before IE are required to handle the quality issues of unstructured big data like data dimensionality, diversity, sparsity, and noise. Standardization of data and processes, efficient tools for data cleaning, high-level data management, and context-aware transformation of unstructured data are important issues that must be addressed to improve the IE systems. More advanced and adaptive data pre-processing techniques are required to overcome the limitations of IE systems for unstructured big data.

**Precondition 2: IE should meet pragmatics.** Pragmatic IE is related to the usefulness and usability of data. This will help to identify various dimensions of unstructured data to solve the problem at hand. As most of the IE systems are task and domain-dependent but still pragmatic IE from





**Figure 5.** IE process improvement model for multifaceted unstructured big data. IE: information extraction.

unstructured big data is an open issue due to the volume and variety challenges of big data. In this regard, identification of pragmatic characteristics according to the type of data and problem is hard to achieve for unstructured big data. The emergence of pragmatics with semantics in IE systems will ultimately improve the efficiency of unstructured data analysis, although it is one of the critical challenges for IE from unstructured big data. Big data analysis using IE systems is based on NLP, language modeling, and structure extraction method. These systems are facing challenges that are already discussed in the previous section. Extracting the contextually relevant information will lead to improved unstructured data analysis. Although extracting contextually relevant information is not easy for unstructured big data due to quality, heterogeneity and dimensionality issues. Contextually and semantically enriched IE systems are the ultimate requirement for improved unstructured data analytics.

**Precondition 3: Advance cross and multilingual systems.** Advanced algorithms and solutions are required to increase the efficiency of IE systems as these systems belongs to NLP. NLP brings language complexity, ambiguities,

language modeling, and understanding issues to IE systems. Meanwhile, domain-specific terms or images also require domain-specific solutions. With the limitations of language and domain specificity, new multilingual systems with reduced domain specificity can reduce the limitations of existing IE techniques.

**Precondition 4: Advanced hybrid IE techniques.** IE from big data is a complex process due to a large amount of variety of data. It performs an important role in the big data analysis pipeline. In this regard, a detailed discussion on IE approaches has been carried out to identify the current status and challenges. It has been concluded that hybrid approaches are performing more efficiently as these approaches can take more advantages from IE techniques by considering the cost and benefit measures. However, computational cost, accuracy, and scalability are important key factors for large-scale data IE.<sup>137</sup> It has been observed that deep neural networks RNN, CNN are performing better in IE field with certain limitations. However, taking advantage of both approaches (rule-based and machine learning-based) as hybrid or emerged approaches to mitigate the limitations of individuals can make IE systems

more reliable and accurate. Specifically for unstructured big data, parallel and distributed computing, using Apache Hadoop and Spark, can be incorporated with these techniques that has a big room of improvement in IE from unstructured big data.

### *Proposed IE process improvement model for multifaceted unstructured big data*

The outcome of this literature review proposed a model to overcome the limitations of existing IE techniques as depicted in Figure 5. The identified challenges of IE from unstructured big data, as presented in Table 11, and corresponding preconditions that would help to improve the IE process are followed in the proposed model to improve the IE process in big data environment.

## Conclusion

The exponential growth of multifaceted unstructured big data is creating challenges in context-aware analytics, data-driven decision-making, and data management. IE techniques are important to extract useful information from unstructured data that improve the effectiveness of data analytics. In this regard, a structured review was conducted to investigate the limitations of existing IE techniques for unstructured big data analysis. For this reason, state-of-the-art IE subtasks and their techniques from different types of data (text, images, audio, and video) have been discussed briefly. This review also investigated the effectiveness of existing IE techniques for unstructured big data. It has been observed that variety of big data is creating numerous challenges for traditional IE systems in terms of accuracy, scalability, generalizability, and usability which leads us to a new era of advanced IE approaches with new opportunities and challenges. It has also been concluded that hybrid approaches (i.e. combination of learning-based and rule-based) achieved better performance in IE whereas the quality of data has a significant impact on the effectiveness of results. However, many challenges are still associated with these hybrid approaches such as language barriers, domain issues, and appropriate method selection for the task at hand. These challenges are specific to IE process but scalability, quality, heterogeneity, and interoperability are critical factors associated with IE from unstructured big data. To overcome the limitations of existing IE techniques, more advanced and adaptive pre-processing techniques are required to remove the quality and usability issues. Further, some suggestions have also been provided in this review by critically analyzing the literature and limitations of existing solutions. Our analysis finds that there is a significant potential to improve the analysis process in terms of context-aware analytics systems. Advanced techniques and methods for IE systems, particularly for multifaceted unstructured big data sets, are the utmost requirement. Existing approaches and methods do not apply to all

domains and varying types of data, even in a single data type. There is a need to develop new techniques and refine existing techniques for pre-processing stage of data that can help to significantly reduce the problems in data sets, later used for IE, knowledge discovery, and decision-making.

## Limitations of the study

The SLR investigates the existing IE techniques and presents the limitations of these techniques for multifaceted unstructured big data. Various recognized data sources have been explored during the study, however, sparsity of literature on this topic and different data types, data formats, and standards made the identification and selection of articles very time-consuming and tedious. Limited but prominent literature could be selected and it restricted the investigations to the limitations of the IE techniques. It was really hard to find the right article and techniques whereby several works have presented many techniques in context of data quality, data extraction, data transformation, and information retrieval. However, it has been managed by careful and proper selection of the articles meeting the inclusion and exclusion criteria.

## Future work

The existing literature presents the techniques and methods for IE from unstructured big data but does not present any model or framework to improve the IE procedures. It is also needed to present a very structured approach and devise methods for specific data types rather than presenting general techniques. Developing a theoretical model based on the comparative analysis of existing techniques (similarities and differences), defining rules for data extraction and transformation is a potential research problem to address the data quality improvement issues.


## Declaration of conflicting interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Kiran Adnan  <https://orcid.org/0000-0002-2474-8844>

Rehan Akbar  <https://orcid.org/0000-0002-3703-5974>

## References

1. Gantz J and Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. <http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf> (2012, accessed 8 May 2019).
2. Wang Y, Kung LA, and Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare

- organizations. *Technol Forecast Soc Change* 2018; 126: 3–13.
3. Lomotey RK and Deters R. Topics and terms mining in unstructured data stores. In: *2013 IEEE 16th international conference on computational science and engineering* (ed L O’Conner), Sydney, Australia, 3–5 December 2013, pp. 854–861. IEEE Computer Society and Conference Publishing Services (CPS).
4. Srinidhi SB, Bhandi V, and Balaji S. Impact of big data and emerging research trends. In: *Proceedings of the international conference, “computational systems for health & sustainability”*, Department of Maths, CSE, ISE, RV College of Engineering, Bangalore, Karnataka, 17–18 April 2015, pp. 14–17.
5. Glauber R and Claro DB. A systematic mapping study on open information extraction. *Exp Syst Appl* 2018; 112: 372–387.
6. Aveyard H *Doing a literature review in health and social care: a practical guide*. 2nd ed. UK: McGraw-Hill Education, 2014.
7. Zhu W, Cui P, Wang Z, et al. Multimedia big data computing. *IEEE Multi Med* 2015; 22: 96–105.
8. Jiang J. Information extraction from text. In: CC Aggarwal and CX Zhai (eds) *Mining text data*. Boston, MA: Springer, pp. 11–41.
9. Bouckaert RR. Low level information extraction: a Bayesian network based approach. In: *Proc of the workshop on text learning (TextML-2002)* (ed ICML), Sydney, Australia, 22 April 2002, pp. 194–202. Technical Report TR98-1702.
10. Piskorski J and Yangarber R. Information extraction: past, present and future. In: T Poibeau, H Saggion, J Piskorski, et al. (eds) *Multi-source, multilingual information extraction and summerization*. Berlin, Heidelberg: Springer, 2013, pp. 23–49.
11. McCallum A Information extraction: distilling structured data from unstructured text. *ACM Queue* 2005; 3: 48–57.
12. Rusu O, Halcu I, Grigoriu O, et al. Converting unstructured and semi-structured data into knowledge. In: *Networking in education and research, roedunet international conference (RoEduNet)*, 11th ed., Siania, Romania, 17–19 January 2013. IEEE Curran Associates.
13. Berman JJ. Providing structure to unstructured data. In: *Principles of big data: Preparing, sharing, and analyzing complex information* (ed H Scherer), 1st ed., San Francisco, CA, USA, 2013, pp. 1–14. Elsevier Morgan Kaufmann.
14. Marrero M, Urbano J, Sánchez-Cuadrado S, et al. Named entity recognition: fallacies, challenges and opportunities. *Comp Stand Interfac* 2013; 35: 482–489.
15. Abdallah ZS, Carman M, and Haffari G. Multi-domain evaluation framework for named entity recognition tools. *Comp Spee Lang* 2017; 43: 34–55.
16. Kanya N and Ravi T. Modelings and techniques in named entity recognition: an information extraction task. In: *International conference on sustainable energy and intelligent systems (SEISCON 2012)*, 3rd ed., Tiruchengode, India, 27–29 December 2012, pp. 104–108. IET IEEE.
17. Sari Y, Hassan MF, and Zamin N. Rule-based pattern extractor and named entity recognition: a hybrid approach. In: *Proceedings of information technology (ITSim), 2010 international symposium*, Kuala Lumpur, Malaysia, 15–17 June 2010, pp. 563–568. IEEE.
18. Plu J and Rizzo G. A hybrid approach for named entity recognition for Indian languages. In: *Proceedings of workshop on NER for south and south east Asian languages, IJCNLP-08*, Hyderabad, India, 12 January 2008, pp. 17–24. Hyderabad: IIIT.
19. Rocktäschel T, Weidlich M, and Leser U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* 2012; 28: 1633–1640.
20. Zhu F and Shen B. Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing. *PLoS ONE* 2012; 7: e39230.
21. Asma Ben A and Abacha PZ. Medical entity recognition: a comparison of semantic and statistical methods. In: *Proceedings of BioNLP 2011 workshop* (ed SIGBIOMED), Portland, Oregon, USA, 23–24 June 2011, pp. 56–64. Association for Computational Linguistics.
22. Jisha PJ, Rajeev RR, and Sherly E. A hybrid statistical approach for named entity recognition for Malayalam language. In: *Workshop on Asian language resources* (ed P Bhattacharyya), 11th ed., Nagoya, Japan, 14–18 October 2013, pp. 58–63. Asian Federation of Natural Language Processing (AFNLP).
23. Singh Bajwa K. Hybrid approach for named entity recognition. *Int J Comp Appl* 2015; 118: 36–41.
24. Dey A, Paul A, and Purkayastha S. Named entity recognition for Nepali language: a semi hybrid approach. *Certif Int J Eng Innovat Technol* 2014; 9001: 2277–3754.
25. Eltyeb S and Salim N. Chemical named entities recognition: a review on approaches and applications. *J Cheminform* 2014; 6: 17.
26. Quimbaya AP, Múnera AS, Rivera RAG, et al. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Comput Sci* 2016; 100: 55–61.
27. Atkinson J and Bull V. A multi-strategy approach to biological named entity recognition. *Exp Syst Appl* 2012; 39: 12968–12974.
28. Küçük D and Yazıcı A. A hybrid named entity recognizer for Turkish. *Exp Syst Appl* 2012; 39: 2733–2742.
29. Jahan N, Morwal S, and Chopra D. Named entity recognition in Indian languages using gazetteer method and hidden Markov model: a hybrid approach. *Int J Comput Sci Eng Tech (IJCSET)* 2012; 3: 621–628.
30. Kale S and Govilkar S. Survey of named entity recognition techniques for various Indian regional languages. *Int J Comput Appl* 2017; 164: 37–43.
31. Sazali SS, Rahman NA, and Bakar ZA. Information extraction: evaluating named entity recognition from classical Malay documents. In: *International conference on information retrieval and knowledge management, CAMP*, 3rd ed.,

- Bandar Hilir, Malaysia, 22–23 August 2016, pp. 48–53. Malaysia: IEEE.
32. Li P, Wang H, Li H, et al. Employing semantic context for sparse information extraction assessment. *ACM Trans Knowl Discov Data* 2018; 12: 1–36.
  33. Wang J, Yu Y, Yan J, et al. A probabilistic method for linking BI provenances to open knowledge base. In: *International conference on brain informatics* (eds Y Yao, R Sun, et al.), Brain Informatics 2010 edition, Toronto, Ontario, Canada, 28–30 August 2010, pp. 367–376. Berlin, Heidelberg: Springer.
  34. Chou CL and Chang CH. Named entity extraction via automatic labeling and tri-training: comparison of selection methods. In: *Asia information retrieval symposium (AIRS 2014)* (eds A Jaafar, et al.), Lecture Notes in Computer Science, Vol. 8870, Kuching, Malaysia, 5 December 2014, pp. 244–255. Cham: Springer.
  35. Konstantinova N. Review of relation extraction methods: what is new out there? In: *Analysis of images, social networks and texts* (eds D Ignatov, M Khachay, A Panchenko, N Konstantinova and R Yavorsky), Yetkarinburg, Russia, 15 February 2017, pp. 15–28. Cham: Springer.
  36. Bach N and Badaskar S. A review of relation extraction. *Lit Rev Lang Stat II* 2007; 2: 1–15.
  37. Lamrani EK, Ben Lahmar EH, Marzak A, et al. Mixed method for extraction of domain terminology from text: linguistic and statistical filtering. In: *Third international colloquium in information science and technology (CIST 2014)* (eds I Jellouli, et al.), Tetuan, Morocco, 22 October 2014, pp. 291–295. Morocco: IEEE.
  38. Mannai M and Karaa WBA. Bayesian information extraction network for medline abstract. In: *World congress on computer and information technology (WCCIT 2013)* (ed NJ Piscataway), Sousse, Tunisia, 22–24 June 2013, pp. 1–3. Lebanon: IEEE.
  39. Nguyen TH and Grishman R. Combining neural networks and log-linear models to improve relation extraction. *Computer Science*. Epub ahead of print 18 Nov 2015. arXiv preprint arXiv:1511.05926
  40. Bast H and Haussmann E. Open information extraction via contextual sentence decomposition. In: *International conference on semantic computing* (ed R Bilof), Irvine, CA, USA, 16–18 September 2013, pp. 154–159. Piscataway: IEEE.
  41. Xavier CC, de Lima VLS, and Souza M. Open information extraction based on lexical-syntactic patterns. In: *2013 Brazilian conference on intelligent systems* (eds AT Ramirez Pozo, H de Arruda Camargo, V Furtado and V Pinheiro), Fortaleza, Brazil, 19–24 October 2013, pp. 189–194. Piscataway: IEEE.
  42. Lahbib W, Bounhas I, Elayeb B, et al. A hybrid approach for Arabic semantic relation extracion. In: *Proceedings of the twenty-sixth international Florida artificial intelligence research society conference, FLAIRS 2013*, St. Pete Beach, FL, USA, 22–24 May 2013, pp. 315–320. AAAI Press.
  43. Goutte C and Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: *European conference on information retrieval* (eds DE Losada and JM Fernández-Luna), Advances in Information Retrieval, ECIR 2005, Lecture Notes in Computer Science, Vol. 3408, Spain, 20–31 March 2005, pp. 345–359. Berlin, Heidelberg: Springer.
  44. Gupta A, Banerjee I, and Rubin DL. Automatic information extraction from unstructured mammography reports using distributed semantics. *J Biomed Inform* 2018; 78: 78–86.
  45. Wang K and Shi Y. User information extraction in big data environment. In: *3rd IEEE international conference on computer and communications (ICCC)* (eds Y Qu, S-U Guan, T Li, Y Ishibashi and Y Wang), Chengdu, China, 13–16 December 2017, pp. 2315–2318. IEEE.
  46. Guo X and He T. Leveraging Chinese encyclopedia for weakly supervised relation extraction. In: *Semantic technology: 5th joint international conference, JIST* (eds G Qi, K Kozaki, J Pan and S Yu), Yichang, China, 11–13 November 2015, pp. 127–140. Cham: Springer.
  47. Torres JP, de Piñerez Reyes RG, and Bucheli VA. Support vector machines for semantic relation extraction in Spanish language. In: *Advances in Computing, CCC 2018* (eds CJ Serrano and J Martínez-Santos), Communications in Computer and Information Science, Vol. 885, Cartagena, Colombia, pp. 326–337. Cham: Springer.
  48. Li J, Cai Y, Wang Q, et al. Entity relation mining in large-scale data. In: *Database systems for advanced applications, DASFAA 2015* (eds A Liu, Y Ishikawa, T Qian, S Nutanong and M Cheema), Lecture Notes in Computer Science, Vol. 9052, Hanoi, Vietnam, 20–23 April 2015. Cham: Springer.
  49. Wang C, Song Y, Roth D, et al. World knowledge as indirect supervision for document clustering. *ACM Trans Knowl Discov Data* 2016; 11: 1–36.
  50. Gao H, Gui L, and Luo W. Scientific literature based big data analysis for technology insight. *J Phys Con Ser* 2019; 1168: 032007.
  51. Liu Z, Tong J, Gu J, et al. A semi-automated entity relation extraction mechanism with weakly supervised learning for Chinese medical webpages. In: *Smart Health, ICSH 2016* (eds C Xing, Y Zhang and Y Liang), Lecture Notes in Computer Science, Vol. 10219, Haikou, China, pp. 44–56. Cham: Springer.
  52. Feldman K, Faust L, Wu X, et al. Beyond volume: the impact of complex healthcare data on the machine learning pipeline. In: *Canada, Towards integrative machine learning and knowledge extraction* (eds A Holzinger, R Goebel, M Ferri and V Palade), Lecture Notes in Computer Science, Vol. 10344, pp. 150–169. Cham: Springer.
  53. Wang P, Hao T, Yan J, et al. Large-scale extraction of drug-disease pairs from the medical literature. *J Associat Inform Sci Tech* 2017; 68: 2649–2661.
  54. Miwa M, Thompson P, Korkontzelos I, et al. Comparable study of event extraction in newswire and biomedical domains. In: *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers* (eds J Tsujii and J Hajic), Dublin, Ireland, 23–29

- August 2014, pp. 2270–2279. Dublin City University and Association for Computational Linguistics.
55. Hogenboom F, Frasincar F, Kaymak U, et al. An overview of event extraction from text. In: *Workshop on detection, representation, and exploitation of events in the semantic web (DeRiVE 2011) at tenth international semantic web conference (ISWC 2011)* (L Aroyo, C Welty, H Alani, J Taylor, A Bernstein, L Kagal, N Noy and E Blomqvist), Vol. 779, Bonn, Germany, pp. 48–57. Semantic Web Science Association.
  56. Jiana B, Tingyu L, and Tianfang Y. Event information extraction approach based on complex Chinese texts. In: *2012 international conference on Asian language processing* (eds Hust and COLLIPS), Hanoi, Vietnam, 13–15 November 2012, pp. 61–64. Piscataway: IEEE Computer Society.
  57. Pham XQ, Le MQ, and Ho BQ A hybrid approach for biomedical event extraction. In: *Proceedings of BioNLP shared task 2013 workshop* (eds C Nédellec, JD Kim, S Pyysalo, S Ananiadou and P Zweigenbaum), Sofia, Bulgaria, 9 August 2013, pp. 121–124. Association for Computational Linguistics.
  58. Jenhani F, Gouider MS, and Ben L. A hybrid approach for drug abuse events extraction from Twitter. *Procedia Comput Sci* 2016; 96: 1032–1040.
  59. Atefeh F and Khreich W. A survey of techniques for event detection in Twitter. *Computat Intell* 2015; 31: 132–164.
  60. Li J, Ritter A, Cardie C, et al. Major life event extraction from Twitter based on congratulations/condolences speech acts. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (ed L Màrquez), Doha, Qatar, 25–29 October 2014, pp. 1997–2007. Association for Computational Linguistics.
  61. Lu D, Voss CR, Tao F, et al. Cross-media event extraction and recommendation. In: *Proceedings of NAACL-HLT 2016* (eds J DeNero, M Finlayson and S Reddy), San Diego, CA, USA, 17 June 2016, pp. 72–76. Association for Computational Linguistics.
  62. Mezhar A, Ramdani M, and Elmezabi A. A novel approach for open domain event schema discovery from Twitter. In: *10th international conference on intelligent systems: theories and applications (SITA)* (ed ENSIAS), Rabat, Morocco, 20–21 October 2015, pp. 1–7. IEEE.
  63. Jan B, Farman H, Khan M, et al. Deep learning in big data analytics: a comparative study. *Comput Elect Eng* 2017; 75: 275–287.
  64. Li P and Mao K. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Exp Syst Appl* 2019; 115: 512–523.
  65. Gheisari M, Wang G, and Bhuiyan MZA. A survey on deep learning in big data. In: *2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)* (ed CSE & EUC), Guangzhou, China, 21–24 July 2017, pp. 173–180. IEEE Computer Society.
  66. Reyes O and Ventura S. Evolutionary strategy to perform batch-mode active learning on multi-label data. *ACM Trans Intell Syst Technol* 2018; 9: 1–26.
  67. Liu X, Zhou Y, and Wang Z. Recognition and extraction of named entities in online medical diagnosis data based on a deep neural network. *J Vis Comm Image Repres* 2019; 60: 1–15.
  68. Lu C, Krishna R, Bernstein M, et al. Visual relationship detection with language priors. In: *Computer vision – ECCV 2016* (eds B Leibe, J Matas, N Sebe and M Welling), Lecture Notes in Computer Science, Vol. 9905, Berlin, Germany, 15 September 2016, pp. 852–869. Cham: Springer.
  69. Antol S, Agrawal A, Lu J, et al. VQA: visual question answering. In: *Proceedings of international conference on computer vision* (ed ICCV 2015), Santiago, Chile, 7–13 December 2015, pp. 2623–2631. Washington, DC: IEEE Computer Society.
  70. Ma L, Lu Z, Shang L, et al. Multimodal convolutional neural networks for matching image and sentence. In: *Proceedings of international conference on computer vision* (ed ICCV 2015), Santiago, Chile, 7–13 December 2015, pp. 2623–2631. Washington, DC: IEEE Computer Society.
  71. Yatskar M, Zettlemoyer L, and Farhadi A. Situation recognition: visual semantic role labeling for image understanding. In: *Proceedings of CVPR 2016, conference on computer vision and pattern recognition (CVPR)*, Casino, Las Vegas, Nevada, USA, 26 June 2016, pp. 5534–5542. IEEE Computer Society.
  72. Agrawal A, Mangalraj P, and Bisherwal MA. Target detection in SAR images using SIFT. In: *Proceedings of ISSPIT 2015, international symposium on signal processing and information technology (ISSPIT)*, Abu Dhabi, UAE, 4 September 2015, pp. 90–94. Washington, DC: IEEE Computer Society.
  73. Gueguen L and Pesaresi M. Interscale learning and classification for global HR/VHR image information extraction. In: *Proceedings of IGARSS 2014, geoscience and remote sensing symposium*, Quebec, Canada, 13–18 July 2014, pp. 1481–1484. IEEE.
  74. Ping Tian D. A review on image feature extraction and representation techniques. *Int J Multimed Ubiq Eng* 2013; 8: 385–396.
  75. Fan S, Liu Z, and Hu Y. Extraction of building information using geographic object-based image analysis. In: *Proceedings of EORSA 2016, 4th international workshop on earth observation and remote sensing applications (EORSA)*, Guangzhou, China, 4–6 July 2016, pp. 140–144. Piscataway: IEEE.
  76. Zhang J, Yang Y, Tian Q, et al. Personalized social image recommendation method based on user-image-tag model. *IEEE Trans Multimed* 2017; 19: 2439–2449.
  77. Li S, Wang S, Zheng Z, et al. A new algorithm for water information extraction from high resolution remote sensing imagery. In: *Proceedings of ICIP 2016, international conference on image processing (ICIP)*, Phoenix, AZ, USA, 25–28 September 2016, pp. 4359–4363. USA: IEEE.
  78. Kuldeep T and Garg PK. Texture based information extraction from high resolution images using object based classification approach. In: *Proceedings of EORSA 2016 third*

- international workshop on earth observation and remote sensing applications* (ed W Qihao), Changsha, China, 11–14 June 2016, pp. 299–303. Piscataway: IEEE.
79. Mingliang H and Yuran L. Study of information extraction algorithm of Poisson noise images based on fractional order differentiation. In: *ICCIS 2013 proceedings, international conference on computational and information sciences*, Shiyang, China, 21 June 2013, pp. 766–769. Washington, DC: IEEE Computer Society.
  80. Xu Y and Duan F. Color space transformation and object oriented based information extraction of aerial images. In: *Proceedings of 21st international conference on geoinformatics*, Kaifeng, Henan, China, 20 June 2013, pp. 1–4. IEEE GRSS (Geoscience & Remote Sensing Society).
  81. Zhang H, Zhao K, Song YZ, et al. Text extraction from natural scene image: a survey. *Neurocomput* 2013; 122: 310–323.
  82. Sumathi C, Santhanam T, and Devi GG. A survey on various approaches of text extraction in images. *Int J Comp Sci Eng Surv (IJCSSES)* 2012; 3: 27.
  83. Vellingiriraj EK, Balamurugan M, and Balasubramanie P. information extraction and text mining of ancient vattezhuthu characters in historical documents using image zoning. In: *Proceedings of IALP 2016, international conference on asian language processing*, Tainan, Taiwan, 21–23 November 2016, pp. 37–40. USA: IEEE.
  84. de Vasconcelos LEG, Kusumoto AY, Leite NPO, et al. Automated extraction information system from HUDs images using ANN. In: *Proceedings of ITNG 2015, 12th international conference on information technology – new generations*, Las Vegas, NV, USA, 13–15 April 2015, pp. 657–661. USA: IEEE.
  85. Younis KS and Alkhateeb AA. A new implementation of deep neural networks for optical character recognition and face recognition. In: *Proceedings of NTIT 2017, new trends in information technology* (eds M Sawalha, A Sleit, B Hammo, N Obaid, H Hiary, KE Sabri, L AlNemer, I Al-Jarrah, J Al-Sakran and H Al-Sawalqahthe), Jordan, 25–27 April 2017, pp. 157–162. University of Jordan.
  86. Suntronsuk S and Ratanotayanon S. Automatic text imprint analysis from pill images. In: *Proceedings of KST 2017, 9th international conference on knowledge and smart technology*, Pattaya, Chon Buri, Thailand, 1–4 February 2017, pp. 288–293. IEEE International publishing Inc.
  87. Deivalakshmi S, Poreddy R, Palanisamy P, et al. Information extraction and unfilled-form structure retrieval from filled-up forms. In: *Proceedings of ICRTIT 2013, international conference on recent trends in information technology*, Chennai, India, 25–27 July 2013, pp. 297–300. Piscataway: IEEE.
  88. Xie H, Fang S, Zha ZJ, et al. Convolutional attention networks for scene text recognition. *ACM Trans Multimed Comput Appl* 2019; 15: 1–17.
  89. Aarthi S and Chitrakala S. Scene understanding—a survey. In: *Proceedings of ICCSP 2017, international conference on computer, communication and signal processing*, Chennai, India, 10–11 January 2017, pp. 1–4. Piscataway: IEEE.
  90. Khosla D, Uhlenbrock R, and Chen Y. Automated scene understanding via fusion of image and object features. In: *Proceedings of HST 2017, IEEE international symposium on technologies for homeland security (HST)*, Waltham, MA, USA, 25–26 April 2017, pp. 1–4. IEEE.
  91. Gleason S, Ferrell R, Cheriyyadat A, et al. Semantic information extraction from multispectral geospatial imagery via a flexible framework. In: *Proceedings of iGARSS 2010, international geoscience and remote sensing symposium* (ed S Yueh), Honolulu, Hawaii, USA, 25–30 July 2010, pp. 166–169. USA: IEEE.
  92. Santosh KC and Belaid A. Document information extraction and its evaluation based on client's relevance. In: *Proceedings of ICDAR' 2013, 12th international conference on document analysis and recognition*, Washington, DC, USA, 25–28 August 2013, pp. 35–39. Washington, DC: IEEE Computer Society.
  93. Foukarakis M, Ragia L, and Christodoulakis S. A digital library system for semantic spatial information extraction from images. In: *Proceedings of GITSAM 2015, 1st international conference on geographical information systems theory, applications and management* (eds C Grueau and JG Rocha), Barcelona, Spain, 28–30 April 2015, pp. 165–169. Lda, Portugal: SCITEPRESS – Science and Technology Publications.
  94. Markowska-Kaczmar U, Szymanska A, and Culer L. Automatic information extraction from heatmaps. In: *Proceedings of IISA 2014, 5th international conference on information, intelligence, systems and applications* (eds NG Bourbakis, GA Tsihrintzis and M Maria Virvou), Chania, Crete, Greece, 7–9 July 2014, pp. 267–272. Piscataway, NJ: IEEE.
  95. Jung J and Park J. Visual relationship detection with language prior and softmax. In: *Proceedings of IPAS 2018, international conference on image processing, applications and systems*, Sophia Antipolis, France, 12–14 December 2018, pp. 143–148. IEEE.
  96. Liang X, Lee L, and Xing EP. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In: *Proceedings of CVPR 2017, IEEE conference on computer vision and pattern recognition*, Honolulu, Hawaii, USA, 21–26 July 2017, pp. 4408–4417. IEEE Computer Society.
  97. Ghule KR and Deshmukh RR. Feature extraction techniques for speech recognition: a review. *Int J Sci Eng Res* 2015; 6: 143–147.
  98. Quinton E, Sandler M, and Harte C. Extraction of metrical structure from music recordings. In: *Proceedings of the 18th international conference on digital audio effects (DAFx)*, Trondheim, Norway, 3 December 2015.
  99. Desai N, Dhameliya K, and Desai V. Feature extraction and classification techniques for speech recognition: a review. *Int J Emerg Tech Adv Eng* 2013; 3: 367–371.
  100. Kumar A and Raj B. Audio event detection using weakly labeled data. In: *Proceedings of MM '16, 24th international conference on multimedia*, The Netherlands, 15–19 October 2016, pp. 1038–1047. NY, USA: ACM Press.

101. Cutajar M, Gatt E, Grech I, et al. Comparative study of automatic speech recognition techniques. *IET Sign Process* 2013; 7: 25–46.
102. Hailemariam S and Prahallad K. Extraction of linguistic information with the aid of acoustic data to build speech systems. In: *Proceedings of ICASSP'07, International conference on acoustics, speech and signal processing*, Honolulu, Hawaii, USA, 15–20 April 2007, pp. IV-717–IV-720. USA: IEEE.
103. Lee CH and Siniscalchi SM. An information-extraction approach to speech processing: analysis, detection, verification, and recognition. *Proceed IEEE* 2013; 101: 1089–1115.
104. Mohammed DY, Duncan PJ, Al-Maathidi MM, et al. A system for semantic information extraction from mixed soundtracks deploying MARSYAS framework. In: *Proceedings of INDIN' 2015, 13th international conference on industrial informatics*, Cambridge, UK, 22–24 July 2015, pp. 1084–1089. IEEE International Publishing Inc.
105. He X and Deng L. Speech-centric information processing: an optimization-oriented approach. *Proceed IEEE* 2013; 101: 1116–1135.
106. Londhe ND and Kshirsagar GB. Chhattisgarhi speech corpus for research and development in automatic speech recognition. *Int J Speech Tech* 2018; 21: 193–210.
107. Rajendra SP. A survey of automatic video summarization techniques. *Int J Electro Elect Comput Syst* 2014; 3(1): 1–6.
108. Biswas A, Sahu PK, and Chandra M. Multiple camera in car audio–visual speech recognition using phonetic and visemic information. *Comp Elect Eng* 2015; 47: 35–50.
109. El khattabi Z and Youness Tabii AB. Video summarization: techniques and applications. *Int J Comp Inform Eng* 2015; 9: 928–933.
110. Ajmal M, Ashraf MH, Shakir M, et al. Video summarization: techniques and classification. In: *Computer vision and graphics, ICCVG 2012* (eds L Bolc, R Tadeusiewicz, LJ Chmielewski and K Wojciechowski), Lecture notes in computer science, Vol. 7594, Warsaw, Poland, 24–26 September 2012, pp. 16–28. Berlin, Heidelberg: Springer.
111. Lu T, Palaiahnakote S, Tan CL, et al. Introduction to video text detection. In: T Lu, S Palaiahnakote, CL Tan, et al. (eds) *Video text detection*. London, England: Springer, 2014, pp. 1–18.
112. Manju A and Valarmathie P. Organizing multimedia big data using semantic based video content extraction technique. In: *Proceedings of ICSNS'15, international conference on soft computing and networks security*, Coimbatore, India, 25–27 February 2015, pp. 1–4. IEEE International Publishing Inc.
113. Kojima R, Sugiyama O, and Nakadai K. Audio-visual scene understanding utilizing text information for a cooking support robot. In: *Proceedings of IROS' 15, IEEE/RSJ international conference on intelligent robots and systems* (eds LF D'Haro, AI Niculescu and A Vijayalingam), Hamburg, Germany, 2 October 2015, pp. 4210–4215. IEEE Inc.
114. Lee YS, Hsu CY, Lin PC, et al. Video summarization based on face recognition and speaker verification. In: *Proceedings of ICIEA' 15, 10th IEEE conference on industrial electronics and applications*, Auckland, New Zealand, 15–17 June 2015, pp. 1821–1824. Singapore: IEEE Industrial Electronics Society.
115. Zhang Z and Shi H. No-reference video quality assessment based on temporal information extraction. In: *Proceedings of IMSNA' 13, 2nd international symposium on instrumentation and measurement, sensor network and automation*, Toronto, Ontario, Canada, 23–24 December 2013, pp. 925–927. Piscataway, NJ: IEEE.
116. Zeng C. Automatic extraction of useful scenario information for dramatic videos. In: *Proceedings of ICICS' 13, 9th international conference on information, communications & signal processing*, Tainan, Taiwan, 10–13 December 2013, pp. 1–5. IEEE International publishing.
117. Mathur A, Saxena T, and Krishnamurthi R. Generating subtitles automatically using audio extraction and speech recognition. In: *Proceedings of CICT' 15, international conference on computational intelligence & communication technology*, Gaziabad, India, 13–14 February 2015, pp. 621–626. Piscataway, NJ: IEEE.
118. Ryu C, Lee D, Jang M, et al. Extensible video processing framework in Apache Hadoop. In: *Proceedings of Cloud-Com' 13, 5th international conference on cloud computing technology and science*, Bristol, United Kingdom, 2–5 December 2013, pp. 305–310. Piscataway, NJ: IEEE.
119. Potapov D, Douze M, Harchaoui Z, et al. Category-specific video summarization. In: *Computer Vision – ECCV 2014. ECCV 2014: Lecture notes in computer science*, (eds D Fleet, T Pajdla, B Schiele and T Tuytelaars), Vol. 8694. September 2014. pp. 540–555. Cham: Springer.
120. Mahasseni B, Lam M, and Todorovic S. Unsupervised video summarization with adversarial LSTM networks. In: *Proceedings of CVPR' 17, IEEE conference on computer vision and pattern recognition*, Honolulu, Hawaii, USA, 26 July 2017, pp. 2982–2991. USA: IEEE Computer Society.
121. Gong B, Chao WL, Grauman K, et al. Diverse sequential subset selection for supervised video summarization. In: *Advances in neural information processing systems 27 (NIPS 2014)* (eds Z Ghahramani, M Welling, C Cortes, ND Lawrence and KQ Weinberger), Montreal, Canada, 8–13 December 2014, pp. 2069–2077. Curran Associates Inc.
122. Wang X, Jiang Y, Yang S, et al. End-to-end scene text recognition in videos based on multi frame tracking. In: *Proceedings of ICDAR 2017, 14th IAPR international conference on document analysis and recognition* (eds K Kise, D Lopresti and S Marinai), Kyoto, Japan, 9–15 November 2017, pp. 1255–1260. Piscataway, NJ: IEEE.
123. Mayer W, Grossmann G, Selway M, et al. Variety management for big data. In: T Hoppe, B Humm and A Reibold (eds) *Semantic Applications*. Berlin, Heidelberg: Springer, 2018, pp. 47–62.
124. Che D, Safran M, and Peng Z. From big data to big data mining: challenges, issues, and opportunities. In: *Database*



- systems for advanced applications, DASFAA 2013*, (eds B Hong, X Meng, L Chen, W Winiwarter and W Song), Lecture Notes in Computer Science, Vol. 7827, 22–25 April 2013, pp. 1–15. Berlin, Heidelberg: Springer.
125. EY GM, Ke W, and Peng T. Big data changing the way businesses. *Int J Simulat Syst Sci Tech* 2014; 16: 28.
  126. Vashisht P and Gupta V. Big data analytics techniques: a survey. In: *ICGCIOT '15 Proceedings of the 2015 International conference on green computing and internet of things*, Noida, India, 8–10 October 2015, pp. 264–269. Washington, DC: IEEE Computer Society.
  127. Gao J and Koronios A. Unlock the value of unstructured data in EAM. In: *Proceedings of the 7th world congress on engineering asset management (WCEAM 2012)* (eds W Lee, B Choi, L Ma and J Mathew), Lecture Notes in Mechanical Engineering, 30 September 2014, pp. 265–275. Cham: Springer.
  128. Jaseena KU and David JM. Issues, challenges, and solutions: big data mining. In: *Computer science & information technology (computer science conference proceeding CSCP)* (eds N Meghanathan, et al.), Bangalore, India, 13–14 September 2014, pp. 131–140. India: CS & IT-CSCP.
  129. Bellot P, Bonnefoy L, Bouvier V, et al. Large scale text mining approaches for information retrieval and extraction. In: C Faucher and L Jain (eds) *Innovations in Intelligent Machines-4: Studies in computational intelligence*, Vol. 514, 2014, pp. 3–45. Cham: Springer.
  130. Liu X, Wei F, Zhang S, et al. Named entity recognition for tweets. *ACM Trans Intell Syst Tech (TIST)* 2013; 4: 3.
  131. Rajbabu K, Srinivas H, and Sudhab S. Industrial information extraction through multi-phase classification using ontology for unstructured documents. *Comput Indust* 2018; 100: 137–147.
  132. Huddar V, Desiraju BK, Rajan V, et al. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access* 2016; 4: 7988–8001.
  133. Shankaranarayanan G and Blake R. From content to context: the evolution and growth of data quality research. *J Data Inform Qual (JDIQ)* 2017; 8: 9.
  134. Bolón-Canedo V, Sánchez-Marroño N, and Alonso-Betanzos A. Recent advances and emerging challenges of feature selection in the context of big data. *Know-Bas Syst* 2015; 86: 33–45.
  135. Sedkaoui S. Data analytics process: there's great work behind the scenes. In: *Data analytics and big data, information systems, web and pervasive computing series*. London: John Wiley & Sons Inc., 2018, pp. 77–99.
  136. Williams K, Wu J, Choudhury SR, et al. Scholarly big data information extraction and integration in the citeseer<sup>x</sup> digital library. In: *Proceedings of ICDE 2014, 30th international conference on data engineering workshops* (eds FC Isabel, E Ferrari, Y Tao, E Bertino and G Trajcevski), Chicago, IL, USA, 31 March–4 April 2014, pp. 68–73. IEEE Computer Society.
  137. Peng C, Cheng J, and Cheng Q. A supervised learning model for high-dimensional and large-scale data. *ACM Trans Intell Syst Tech (TIST)* 2017; 8: 30.