IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 June 2025, accepted 23 August 2025, date of publication 2 September 2025, date of current version 8 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3605116

RESEARCH ARTICLE

TForMIX: A Method That Combines LLM and Multidimensional Modeling for Technological Foresight

GISELLE F. ROSA[1,2], **JONES O. AVELINO**[1,3], **MARIA CLAUDIA CAVALCANTI**[1], **AND JULIO CESAR DUARTE**[1]

[1]Instituto Militar de Engenharia (IME), Rio de Janeiro 22290-270, Brazil
[2]Agência de Gestão e Inovação Tecnológica (AGITEC), Rio de Janeiro 23020-470, Brazil
[3]Centro de Análise de Sistemas Navais (CASNAV), Rio de Janeiro 20091-000, Brazil

Corresponding author: Maria Claudia Cavalcanti (yoko@ime.eb.br)

This work was supported by the Financiadora de Estudos e Projetos (FINEP)/Departamento de Ciência e Tecnologia (DCT)/Fundação de Apoio à Pesquisa, Desenvolvimento e Inovação do Exército Brasileiro (FAPEB) through the "Systems of Command and Control Systems" Project ("Sistemas de Sistemas de Comando e Controle," Portuguese) under Grant 2904/20-01.20.0272.00.

ABSTRACT Technical documents, such as scientific papers and patents, are widely used as a basis for Technological Foresight (TF) processes. Typically, these analyses require identifying elements (e.g., terms) in the textual contents of these documents, which are relevant to the scientific-technological domain under investigation. Information Extraction (IE) and Natural Language Processing (NLP) techniques are useful tools to automate the identification of these elements, which is essential in TF processes that usually involve the analysis of a corpus of hundreds (and sometimes thousands) of documents. An analytical view over this corpus, based on the occurrence of those relevant elements, helps prioritize document analysis and, consequently, accelerates the whole TF process. However, building a system that provides such analytical insight is expensive. Moreover, for each domain-specific TF process, a new system would have to be built. Thus, there is a need for viable solutions to analytically explore a corpus, according to the specific requirements of each domain. This work presents Technological Foresight with Multidimensional Information eXtraction (TForMIX), a novel method for building Decision Support Systems (DSSs) that applies Named Entity Recognition (NER) and Relation Extraction (RE) while allowing multidimensional analytical exploration of entities and relations together with bibliometric data from documents. TForMIX is a flexible method that can be applied to different domains, and speeds up building DSSs for each domain. Additionally, we evaluate the applicability of the produced DSSs in TF processes by conducting a practical experiment that demonstrates that applying the method to generate DSSs, supported by IE techniques, can significantly contribute to the conduction of TF analyses. The combination of the used theories, innovative methods, and proposed practical validation highlighted the high-quality nature of the analysis in this study while offering the potential for valuable insights and contributions to the TF process.

INDEX TERMS Decision support systems, information extraction, multidimensional modeling, technological foresighting, technological forecasting.

I. INTRODUCTION

Conducting Technological Foresight (TF) processes includes several analytical studies that aim to identify and monitor scientific and technological trends. Technical documents such as scientific articles and patents are extensively used as a basis for these studies, as they are rich in technical information and widely available. Typically, when a domain-specific TF is under investigation, the collection of documents (corpus) to be analyzed is quite large, consisting of hundreds (sometimes thousands) of documents. These documents must be organized and prioritized according to aspects that are specific to the domain.

Technological trends analysis has improved over the last two decades, as recent advances in Natural Language Processing (NLP) have created new possibilities for

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

automating the exploitation of unstructured textual parts of documents [2], [3], allowing us to go beyond the traditional approach of merely analyzing structured bibliometric data.

All this variety of data, structured or unstructured, found in documents analyzed during these processes leads to an overload of information available for analysis, requiring the use of automated extraction, processing, and knowledge generation methods [1]. As a result, it is possible to take advantage of structures and systems such as Decision Support Systems (DSSs) based on dimensional models, or Data Warehouses (DWs), which are focused on facilitating access, manipulation, and exploration of data from multidimensional perspectives, with queries that allow navigating the data through dynamic views.

However, applying DW based systems for this purpose has some barriers. Even being very relevant and powerful to work with structured data, these systems still find great challenges in dealing with semi-structured or unstructured data [6]. For this purpose, one can take advantage of NLP techniques by applying them to extract and structure the rich textual information present in scientific papers and patents.

Information Extraction (IE) is one of those techniques that consists of automatically extracting structured information from unstructured or semi-structured data sources. In light of this, using IE to analyze the unstructured text, it is possible to extract relevant information, helping to identify salient facts that can compose a structured representation of the information taken from the text [4] and enrich databases or knowledge bases [5]. Named Entity Recognition (NER) is an important IE task that has the goal of locating units of information in the text and classifying them semantically according to predefined categories. Once entities are located in the text, another important task of IE is to locate and classify significant relationships between them. This task is called Relation Extraction (RE).

In this way, these IE techniques allow incorporating semantics into text analysis [15], becoming more effective in discovering richer and more representative concepts and relationships, while revealing latent structures in textual content beyond that based on the syntactic functions performed by the words. Thus, when applied in support of TF analytical processes, they can help structure the textual content of patent documents and scientific papers, by detecting useful constructions in the text for TF analyses, such as problem-solution, technology-function, function-product, technology-product, etc. [2], [10].

Although IE techniques may help identify the analytical needs for a TF process, and consequently, support the construction of the corresponding DW, they are time-consuming tasks and must be performed for each domain. Some related works [7], [8], [9] have applied NLP techniques to build domain-specific DW based systems and showed the benefits of this approach; however, they did not provide a domain-independent solution. Some other works [10], [11], [12] have applied NLP techniques to support TF processes, but they did not provide an analytical view of the corpus. Therefore, to the best of our knowledge, no work has investigated a methodology for developing a DW based system supported by NLP techniques that can be applied to any domain, as required by TF processes.

This work presents, as its primary contribution, a novel approach for building DSSs that allows comprehensive multidimensional analyses and significantly enhances analytical TF processes. The DSSs are designed to facilitate faster, more effective decision-making, grounded in robust scientific and technological data analyses. In addition to the method, named Technological Foresight with Multidimensional Information eXtraction (TForMIX), and the overall framework design, this work also contributes with a conceptual metamodel tailored for the unique demands of TF. Furthermore, it presents the implementation of a Decision Support System (DSS) instance, specifically developed for a targeted study area, demonstrating its practical application. We present the proposed method and the dimensional model on which it is based, while also showcasing an implementation of the method that results in a DSS focused on the NLP area, which we submit to experimentation by TF analysts.

The organization of this article is as follows. In section II, we highlight the fundamental concepts of this work. In section III, we briefly examine some works closely related to our study. Then, we describe the proposed method in section IV and the experiments conducted to validate the methods in section V. Finally, in section VI, we present the conclusions and some suggestions for future work.

II. BACKGROUND

Our method intends to produce DSSs based on pieces of information extracted from unstructured textual data. In this section, we first briefly present concepts about Technological Foresight (subsection II-A), Decision Support Systems (subsection II-B) and Information Extraction (subsection II-C). We also introduce the Faceted Classification Theory (subsection II-D), as its key ideas are employed by the method to support the definition of the schema used to classify the information extracted from the texts.

A. TECHNOLOGY FORESIGHT

Technology Foresight consists of a set of activities focused on the creation of the Science and Technology (S&T) future, which includes predicting the characteristics of forthcoming technologies and the period of their appearance [32]. To achieve these purposes, the literature presents a wide range of tools specifically developed and consolidated over the years [34], which are commonly classified according to the foresight diamond developed by Popper [31]. In this classification, some methods are categorized as quantitative, as they define numerical parameters characterizing the studied phenomenon or the object of study [32]. Popper also highlights the importance of information technology in the ability to generate more sophisticated data analysis, increasing the capacity to generate, analyze, and communicate foresight insights.

Among the methods used in TF processes, there is a specific category that relies on analytical tools to incorporate different 'views' into the analysis models, allowing for the analytical treatment of a large amount of information [13]. *Technological Roadmap* (TRM) is one of these methods. For instance, to analyze and identify technological and market trends in levulinic acid production, the analysts [33] applied the TRM method and organized the documents according to the following categories: *Document focus* (application, technology or equipment), *Application area* (fuels, polymers and resins, fuel additives or chemical products), and *Stage* (pre-treatment, treatment, conversion or recovery).

B. DECISION SUPPORT SYSTEMS

Decision Support Systems (DSSs) are "interactive computer-based systems that help people use computer communications, data, documents, knowledge, and models to solve problems and make decisions" [22]. The focus on supporting analytical processing differentiates DSSs from transactional systems, which aim at supporting an organization in day-to-day transactional operations. DSSs must provide the right information at the right time in the right format, offering users fast and interactive information support [23]. To achieve highly efficient information synthesis, DSSs must employ robust data storage and querying methods that allow comprehensive analysis from diverse business perspectives. To accomplish this, these systems rely on DW, a data repository specially conceived to support the decision-making process, tools for data insertion, and Online Analytical Processing (OLAP) tools to query and process data from an analytical point of view.

Data models that adhere to the Third Normal Form (3NF) prioritize reducing redundancies and maintaining integrity to provide agility and security to transactions. Conversely, dimensional models focus on the arrangement of data in a structure that is optimized for complex analytical queries, rather than insertion and update transactions, presenting a lower degree of normalization than 3NF models. This denormalization simplifies queries involving pairwise joins between database tables and facilitates understanding by presenting visually simpler schemas and entities closer to the real business [24]. Therefore, dimensional modeling can be seen as the most appropriate technique to arrange analytical data, presenting understandable data to users and allowing better performance of queries [25]. One way of implementing a dimensional model is the *Star Schema*, which consists of two main components: facts and dimensions. A fact is a business observation, while dimensions represent the business entities that provide the contextual information and descriptive attributes surrounding the facts, allowing for filtering and aggregating operations [23].

An Operational Data Store (ODS) is "a subject-oriented, integrated, current, volatile collection of data used to support the tactical decision-making process for the enterprise" [26]. The data in the ODS is kept up-to-date with the data from the operational systems, with the difference that they have a certain degree of integration, closer to what happens in a DW. Thus, the ODS is used as a data source for the DW, acting as an intermediate repository that integrates current data from different sources.

C. INFORMATION EXTRACTION

Information Extraction (IE) is the process of automatically obtaining implicit structured information from unstructured or semi-structured data sources. Named Entity Recognition (NER) and Relation Extraction (RE) are two important IE tasks that seek to locate in texts entities and relationships between them, respectively.

In terms of NER, there are different strategies for conducting the task process. The gazette-based approach applies a simple lookup table containing all possible entities for each category. Another approach is rule-based: using regular expressions and linguistic information that define rules to extract and recognize the categorized entities. Additionally, it is possible to use an Machine Learning (ML)-based approach that uses texts with annotated examples to compose a training set, which generates a model to recognize and classify occurrences of entities in other texts. This approach has shown promising results in the literature, especially applying advanced ML techniques, such as Deep Learning [17], [27], yet it requires considerable manual effort for creating the training set [12].

There are also different approaches concerning the RE task. The first consists of conducting the process in two stages: Relation Identification (RI) and Relation Classification (RC). In RI, each pair of entities is fed into a binary classification to determine whether there is a semantic relationship between them. Then, in RC, positive samples are subjected to a multi-class classification process to determine their types from a predefined relationship list. There is also a second approach, which consists of conducting RE in a single step, assuming there is a relation between each pair of entities and simply determining its category. It is also possible to apply heuristics to improve the quality of extractions, such as limiting the distance between entities or specifying head and tail category entities that are not allowed for each relation.

D. FACETED CLASSIFICATION THEORY

The Faceted Classification Theory [16] is a solution to organize knowledge based on the realization that any object of knowledge can be described by arrangements of aspects, referred to as facets. Ranganathan and Gopinath [16] also proposed a classification schema based on fundamental categories applicable to all subjects: Personality, Matter, Energy, Space and Time (PMEST). *Personality* aggregates core concepts, the main objects of study of a discipline, representing its essence. In general, it denotes methods or tools used to achieve a purpose. *Matter* aggregates concepts related to properties of Personality. *Energy* groups together concepts that denote activities that result in an effect on the field of knowledge. *Space* and *Time* correspond to their usual connotations and represent concepts that translate, respectively, ideas of locations and chronological references. For example, evaluating the knowledge contained in a book titled "The Design of Wooden Houses in Southern Brazil at the beginning of the 20th Century" according to the fundamental PMEST categories would result in the following facets [29]: (a) Personality – houses; (b) Material – wood; (c) Energy – project; (d) Space – southern Brazil; (e) Time – beginning of the 20th century.

Broughton [30] noted the existing distinction between intra- and interfacet relationships and compared it to the distinction between paradigmatic and syntagmatic relationships. Intrafacet relationships, which occur between concepts of the same facet, are generally paradigmatic, as they are inherent to the nature of the concepts. Interfacet relationships, on the other hand, are syntagmatic as they depend not on the nature of the concept but on the interaction brought about by the context in which they occur.

III. RELATED WORK

Due to the multidisciplinary nature of the intended research, which involves TF, Text Mining, and DSS supported by multidimensional models, the search for related works was conducted on the following intersections: (i) Text Mining with Multidimensional Models (targeting papers on the insertion of information from textual sources into systems based on dimensional models); and (ii) Text Mining with TF. Works addressing the application of text mining techniques to extract information from texts stored in databases and Data Marts (DMs), as well as technology foresight in the areas of Text Mining and NLP, were discarded.

An example of the first group is the work of Chiudinelli et al. [7], who employed ontology-based NLP techniques to extract information from unstructured documents of clinical notes and integrate it into a DW. In this work, even though the extraction of the proposed system is supported by NLP, ontologies play a central role, since some properties of ontology elements are used to extract events and attributes from the text, and the storage of information in the DW also depends on the ontologies. As the authors pointed out, constructing an ontology to support the proposed method is a complex process that involves many iterations of evaluation and redesign. This can become even more challenging in TF studies that address highly complex subjects.

Another study in this direction proposes to extract unstructured information from medical records for inclusion in Clinical Data Warehouses (CDW) [8]. The authors outline the primary aim of their study, which deals with the extraction of information from texts. Their objective is to enable users to conduct queries using inclusion and exclusion criteria on these texts, akin to the manner in which structured data is typically queried within a CDW. The approach retrieves terms from texts through searches based on regular expressions, while recognizing and excluding negations. Moreover, not only does it extract concepts, but it also calculates their frequencies. Nevertheless, it is important to note that these extracted concepts are currently restricted to boolean and numerical values. Thus, the system can answer questions such as "How many patients have heart failure?" (including textual synonyms) or "How many patients have Left Ventricular Ejection Fraction (LVEF) < 45?". Extractions are performed at runtime and not during the ETL process, and the authors emphasize that this dynamic approach is more advantageous, as it better accommodates changes and requires less development effort; however, it is less accurate. In addition, they mention the advantage of not using ontologies, since it takes a long time to design and build them.

Nguyen et al. [9] focus on the joint use of NER with multidimensional environments. The authors propose the use of NER to structure real-state advertisement data to allow analysis in a DW environment. Despite not detailing the dimensional modeling, the work reinforces that the use of these models is suitable for analyzing this type of information extracted from texts.

In the second group are works that employ IE techniques to support TF processes based on the textual contents of scientific articles and patents [10], [11], [12]. In general, to facilitate deeper analyses beyond the mere extraction of technological terms from texts, these studies also prioritize the localization and identification of the functions of these terms, as well as the relationships that exist between them.

Miao et al. [10] perform semantic analysis focused on technology-relationship-technology constructions to semi-automatically extract relevant information from patent texts, aiming to help in defining analysis dimensions during the construction of Technology Roadmaps (TRM). The authors propose extractions that allow obtaining information on relationships between products, functions, and technologies, which helps to circumvent the need for an expert's opinion, making the process less costly. It should be noted, however, that the proposed method is limited to seeking relationships between technologies and performs extractions based on rules that consider noun-preposition-noun structures, which limits the applied semantics.

Puccetti et al. [11] exploit the most recent advances in NLP to extract information from patent texts. The authors show that text mining techniques for technological analysis found in the state-of-the-art focus on generic terms to investigate technologies. For this reason, the authors use NER to explore the text and identify the technologies presented. As highlighted in the document, multiple methods for NER exist, including gazetteer-based, rule-based, and deep learning-based approaches (e.g., utilizing BERT). The authors conduct a thorough comparison and integration of these three methods to achieve optimal results.

Vito et al. [12] also use NER to extract technology-related terms to evaluate the technological evolution of an area. In the proposed approach, named entities referring to technologies are extracted from a set of technical documents that include scientific articles and are also used to retrieve patent documents and evaluate their co-occurrence over time. The authors chose to use an unsupervised method to extract the entities, arguing that there is no dataset available for the task. As a result, they include a review step to filter out those that are not technologies from the entity extraction results.

In general, the literature proves the applicability of solutions based on dimensional models to store information extracted from unstructured texts to enable its processing by analytical tools. In addition, these works also point to the relevance of text mining techniques that explore the discovery of relevant terms and their semantic relationships in activities that demand obtaining structured knowledge from technical documents such as scientific articles and patents. They also highlight that the utilization of ontologies for facilitating extractions can impose a significant burden on the development process. Specifically, the use of NER and RE offers promising alternatives, as they provide useful information about the occurrence and interaction between technology, function, and product, among other relevant concepts for analyzing a technological domain.

Table 1 compares the related works. It shows that works combining DW based systems with IE techniques are aimed at constructing a single domain-specific system. On the other hand, the remaining works are domain-independent as they aimed at TF processes, and suggest using IE techniques to help analysts on each investigation. However, these works do not include support for the construction of an analytical view. To achieve this, it would be necessary to build a dimensional schema for each domain, which would allow the extractions to be analytically explored. As far as we know, none of these works presented a method to support this task.

IV. TECHNOLOGICAL FORESIGHT WITH MULTIDIMENSIONAL INFORMATION EXTRACTION

The conducted literature review pointed to a lack of methods for building TF tools that allow analytical exploration of the textual content of technical documents, like patent and scientific papers. To fill this gap, this work proposes Technological Foresight with Multidimensional Information eXtraction (TForMIX), a method for constructing a TF DSS that employs NER and RE techniques applied to technical texts grounded in a specific dimensional model that explores the extracted information alongside other relevant metadata. Since the method is independent of ontologies for information extraction from texts, it does not require the construction of an ontology for each domain to be analyzed. It is usually a costly process involving a series of complex activities. Instead, the method proposes using generic metacategories, such as those proposed by Shiyali and Malur [16], to support the distinction, selection, and organization of the domain analysis categories. The method, thus, applies a solution that is less costly than ontology construction but still meets the requirements for building DSS according to the needs of TF processes. Figure 1 depicts the operation of the DSS based on the proposed method.

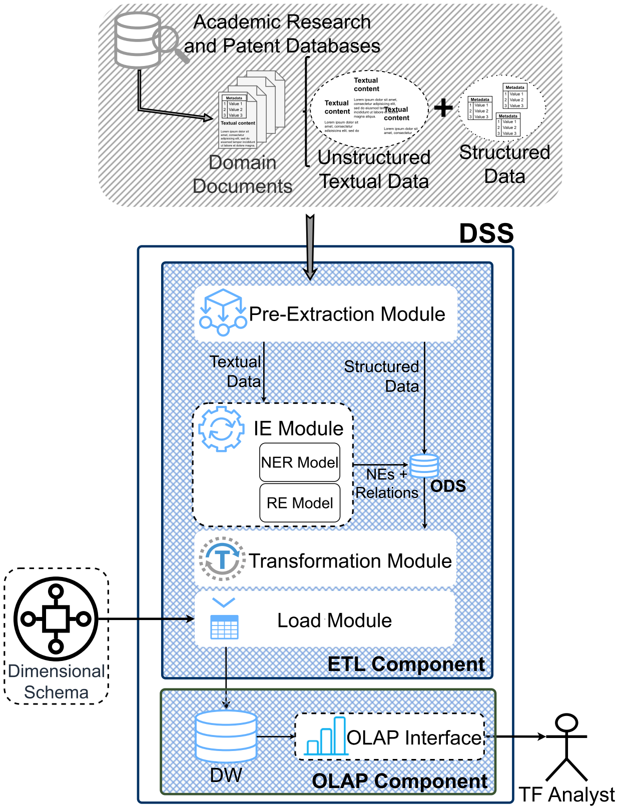


FIGURE 1. Architecture of the DSSs obtained following the proposed method.

Initially, a collection of domain-specific documents retrieved from academic research and patent databases is the input to the Pre-Extraction Module, which extracts its textual content and the structured bibliometric data of interest to the analyses. The system then stores the bibliometric data in an ODS, which is later integrated with the data extracted from the texts.

Next, the IE Module transforms the extracted textual contents into structured information. This module contains NER and RE models, which process the texts and extract Named Entities (NEs) and relations, classifying them according to the categories of interest and relation types. To achieve this, these models are customized to be able to provide meaningful extractions for the domain to be analyzed. The extracted data are also stored in the ODS.

The Transformation Module, then, operates on the bibliometric data, as well as on the NEs and relations data extracted by NER and RE models. As in a regular Extract Transform Load (ETL) process transformation step, this module performs standardization and cleaning operations to ensure data quality for analysis.

These transformed data are then fed into the Load Module, which transfers them into the DW under a dimensional schema designed to allow analytical exploration of NEs and relations extracted from texts together with the original structured data of the documents.

TABLE 1. Related work comparison.

Work	Domain	IE technique Application	Ontology usage	DW based system	Method
[7]	clinical notes	Term retrieval using regular expressions	✓	✓	
[8]	medical records	Boolean and numerical values retrieval		✓	
[9]	real-state advertisement	Named Entity Recognition (NER)		✓	
[10]	TF process	Relation extraction (RE) based on part-of-speech			
[11]	TF process	NER			
[12]	TF process	NER			
This	TF process	NER and RE		✓	✓

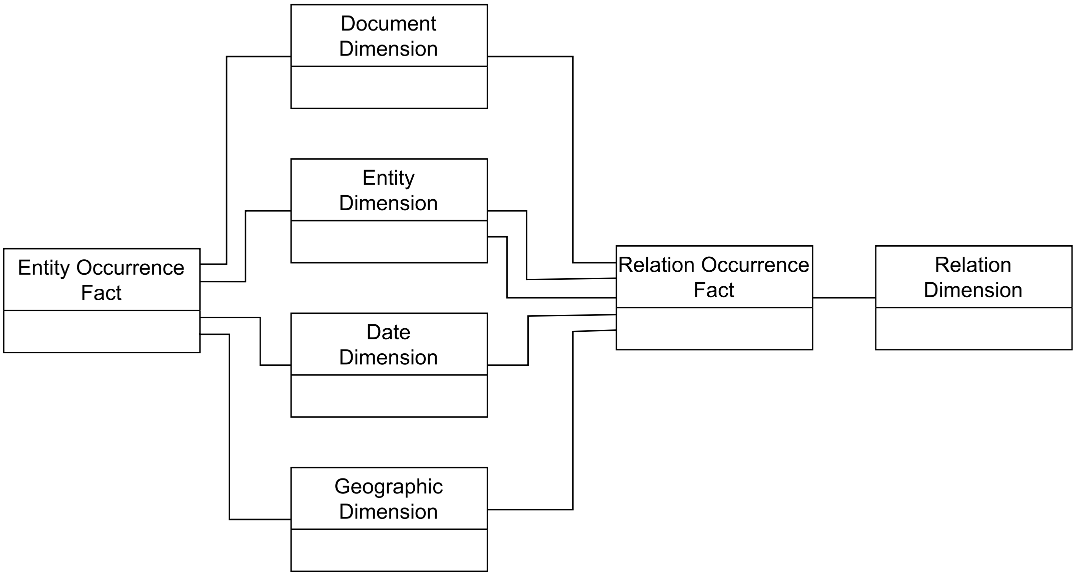


FIGURE 2. Simplified illustration of the dimensional schema that supports TForMIX.

This dimensional schema is based on the recognition that the facts to be explored primarily revolve around two key types of business events within documents. The first one involves the mention of an entity, where specific entities are referred to or discussed. The second one pertains to the semantic relation between two entities within the same document, indicating their connection or association. To provide context for facts, the schema presents separate dimensions to entities, relations, and documents besides the traditional date and geography dimensions. These facts and dimensions form a star schema as illustrated in Figure 2.

The OLAP Component of the DSS encompasses the DW containing the data to be explored, and the OLAP interface, specifically designed for the analyses intended for the domain. When interacting with this interface, users can filter and aggregate data to obtain proper reports and document sets that help them perform the required TF analyses.

Wrapped in dashed lines, in Figure 1, are the elements that distinguish these tailor-made systems from traditional DW-based DSSs: (1) a module that provides structured information available in unstructured texts in terms of NEs and relations; (2) a dimensional schema suitable to couple this information with other bibliometric data; and (3) an OLAP interface to support exploring the data to acquire significant insights from TF perspective.

These features allow TF analysts to conduct multidimensional analyses on documents extracted from search bases under customized perspectives for each domain. Among the available perspectives, in addition to those made possible by structured data, such as publication or invention year, country, patent classification, or research area, are those based on NEs and relations discovered in the texts, such as an entity that represents a technology; an entity that represents a problem; and a relation between them, indicating that the former is used to solve the latter. Thus, filtering and grouping analyses, especially over unstructured textual content, are not conducted solely based on keywords but by dimensions that are populated by discoveries made by NER and RE models on the texts, taking into account specific requirements within that domain.

We aim to produce DSSs with the general characteristics described but tailored to address the needs of specific TF processes. As these specificities vary in terms of the domain of knowledge to be analyzed and in terms of the interests

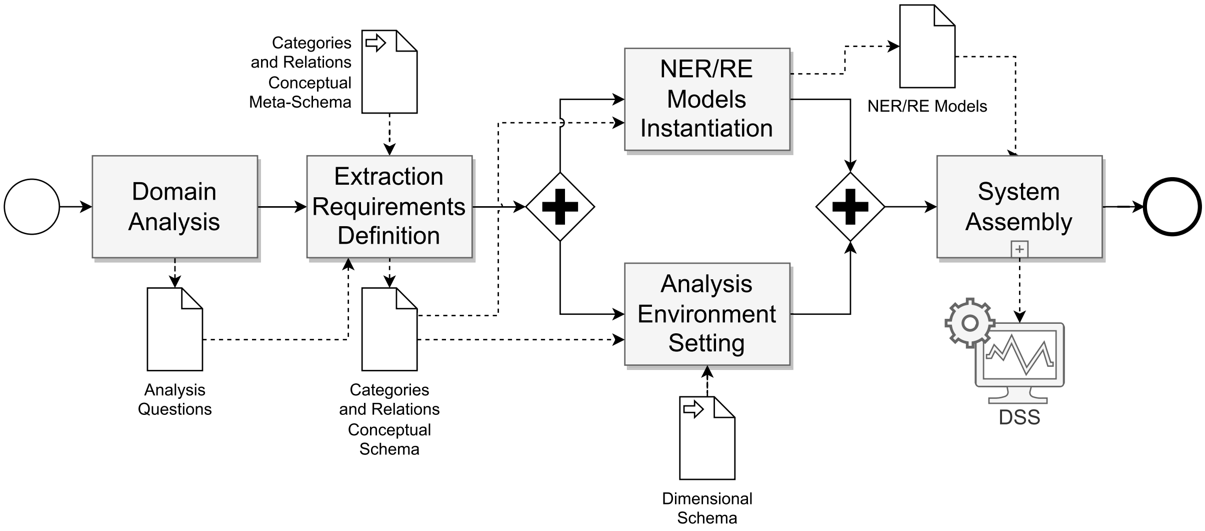

Categories and Relations Conceptual Meta-Schema
Domain Analysis
Extraction Requirements Definition
NER/RE Models Instantiation
NER/RE Models
Analysis Environment Setting
System Assembly
Analysis Questions
Categories and Relations Conceptual Schema
Dimensional Schema
DSS

FIGURE 3. TForMIX overview.

of the analyses, we propose TForMIX as a novel method to produce DSSs that are specific and customized from these two points of view. Figure 3 provides a visual representation of the proposed method.

A. DOMAIN ANALYSIS

This first activity aims to gather the specificities of the TF process that the intended DSS will support. This is accomplished through interviews with TF analysts, seeking to extract from them the intricate details of the analyses to be conducted, which will guide the entire development of the DSS. In addition to the content of technical documents that will serve as input to the system, it is crucial to consider the manual analysis process that would be carried out in the absence of a DSS. Furthermore, incorporating bibliometric data into the requirements-gathering process adds an important dimension to the analysis, which allows a more unified and systemic understanding of the domain. Once these needs are defined, they are translated into analysis questions. These questions must possess sufficient descriptive power to enable, in subsequent steps, the derivation of detailed specifications necessary for structuring the textual data, such as NE categories and relations enumerations, while capturing the details of the analytical environment of the system.

B. EXTRACTION REQUIREMENTS DEFINITION

Once the analysis questions have been established, the next step consists of specifying what must be extracted from the documents' textual content. This step of TForMIX is based on a conceptual meta-schema that, when instantiated according to the specific needs of the DSS under development, gives rise to the conceptual schema that binds how the entities extracted from texts interact with each other through relations depending on the categories to which they belong. Thereby, during the execution of the method to build a DSS for a domain, the proposed conceptual meta-schema is instantiated, resulting in a conceptual schema that, although customized, obeys the general rules stated by the meta-schema. The rules defined by the conceptual schema will constrain the training of NER and RE, performed during the upcoming NER/RE Model Instantiation Step, to generate appropriate NER and RE models.

With this approach, once the DSS is ready to be used, the information extracted from texts by the NER and RE models populates the classes provided by the conceptual schema, resulting in a collection of data that will be consistent and respect the constraints of the domain. Additionally, the construction of the OLAP Interface also respects these constraints, which allows the prior establishment of useful metrics and views that align with the established rules.

To achieve this intention, we formulated the conceptual meta-schema enclosed within the central dotted frame (b) of Figure 4. The NEs and relations that occur in the texts of documents are the elements of reality that the meta-schema intends to represent. The Representative superclass maps all these items generically to being specialized as an Entity or a Relation. This modeling approach offers expressiveness and flexibility to the schema, enabling the classification of mentions that occur in the text based on the role that best suits each domain in question, while allowing for a more accurate representation of NEs and relations.

As the NER task requires categories to classify the extracted entities, the metaclass Entity must be specialized according to these domain-specific categories. However, instead of specializing the Entity metaclass directly into classes that represent these NE categories, we envisioned an intermediate specialization layer to both expand the analytical exploration possibilities of the DSS and guide the TF experts in defining the NE categories they should choose for that domain. To accomplish this, we added

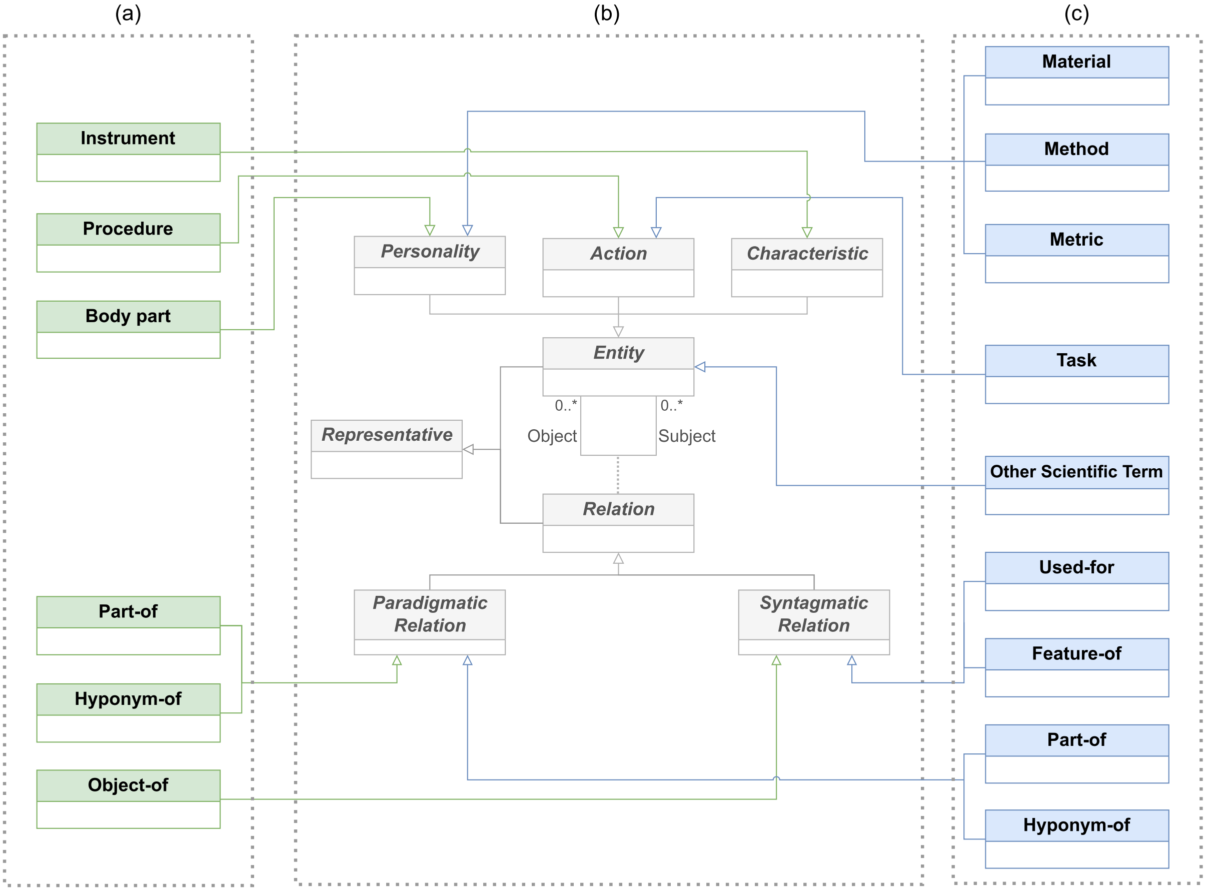


FIGURE 4. The proposed conceptual metaschema (b) and two different specializations: human medicine domain (a) and ML for NLP (c).

intermediate metaclasses to the metaschema that help to better structure and organize the specialized conceptual schema. Based on the Faceted Classification Theory and inspired by its fundamental categories, we created the following metaclasses: Personality, Action (derived from Energy), and Characteristic (derived from Matter). These metaclasses may, then, be specialized into the classes that represent the NE domain categories.

Accordingly, the *Personality* metaclass generalizes categories of entities that, in the context of the domain to be studied, congregate core concepts that represent the main objects of study of the discipline, expressing its essence. Instances of this class are mostly methods or tools used to achieve a purpose. The *Action* metaclass, on the other hand, generalizes categories whose entities denote activities that result in an effect for that domain. The instances of such categories do not necessarily refer to the occurrence of events but rather to the conceptual idea of a planned, potential activity. Instances of this class are mostly verbs or action nouns that denote the achievement of some purpose. Finally, the *Characteristic* metaclass generalizes categories whose entities denote qualities of a Personality or Action. Instances of this class are mostly terms that work as adverbial or adnominal adjuncts, adding extra meaning to instances of other classes, like specifying the instrument of an Action or the size of a Personality instance.

We also exploited the distinctions considered by Broughton [30] when analyzing how categorized terms can relate to each other and assumed that these relationships can have a syntagmatic or paradigmatic nature. The Relation metaclass is an association class between entities, with the first entity assuming the role of the subject and the second assuming the role of the object. Following the metamodel Pattern proposed by Muller [21], relations encompass several possibilities of relationships and are specialized by two metaclasses: the first representing the paradigmatic relations, which pertain to relationships between entities of the same category, and the second representing syntagmatic relations, which involve relationships between entities that may not necessarily belong to the same category. To maintain clarity in the conceptual metaschema diagram (Figure 4), the restriction imposed by the paradigmatic relation class of only associating instances of the same class is not represented. These relation restrictions are one of the benefits of the metaschema, as mentioned before, and are also inherited by the specialized schema.

It is worth mentioning that, when specializing the proposed metaschema to obtain a conceptual schema for a specific purpose, not all metaclasses must be specialized, as the analysis questions may not require having categories and relations of all the proposed supertypes. By adding the items outside the central dotted frame (b) of Figure 4, we obtain a conceptual schema that exemplifies the specialization of the metaschema to the Human medicine study domain, as shown in the left dotted frame (a) of Figure 4. Parts of the human body express this domain's essence, so Body Part is a class that specializes 'Personality'. `Heart` and `aortic valve` are examples of terms that can be located in texts as entities that belong to the entity category Body Part, so they are instances of the class Body Part. Conversely, medical procedures denote activities that result in an effect for this domain, so 'Procedure' is an Action. Examples of entities that instantiate this class are `surgery`, `aortic valve repair`, and `orthopedic surgery`. An instrument (e.g., `laser`) or instrumental adjunct (e.g., `Robotic-Assisted`) denotes a procedure quality, so 'Instrument' is viewed as a characteristic of the procedures in this domain. Regarding relations, for this domain, we specify the `Part-of` and `Hyponym-of` relations, which relate entities of the same category and are then specialized as Paradigmatic Relations. On the other hand, in this domain, we also have the `Object-of` relation, which is categorized as a syntagmatic relation, since it can occur between entities belonging to distinct categories.

Figure 5 shows four instances (1 to 4) allowed by the schema presented in frames (a) and (b) of Figure 4. For example, Procedures like `orthopedic surgery` and `aortic valve repair` may relate to the Procedure `surgery` by the `hyponym-of` relation. Also, it is possible to represent that a Body Part, like `aortic valve`, is part of another Body Part, `heart`, and is the `object-of aortic valve repair`. Conversely, the fifth instance is not allowed, since an instance of Body Part cannot be related to a Procedure by `part-of`.

The conceptual metaschema is an important instrument for TForMIX, as it allows arranging NE categories in a hierarchical structure that is essential to impose restrictions on the semantic relations to be extracted from texts in the IE Module of the DSS. Moreover, the OLAP Component operations can explore both the hierarchy of categories and relations. Besides the benefit of assisting in selecting the final categories, the metaschema helps TF analysts and developers in other steps of the system's formulation, as it supports not only the system's concept comprehension but also plays a crucial role in its validation. Furthermore, it is a convenient documentation to guide text annotation that helps create training datasets to train NER and RE models. Thus, at TForMIX's Extraction Requirement Definition step, the analytical needs of the system, specified by the analysis questions, constrain the specification of the NE categories and relations that are used, respectively, in the NER and RE tasks and also support the analysis environment construction.

C. NER AND RE MODELS INSTANTIATION

The conceptual schema presented in the previous subsection specifies the lists of NE categories and relations of interest to a specific domain. It also restrains the categories of NE that can play the subject and object roles in each relation. To extract information from the textual contents based on that, it is necessary to have customized NER and RE models.

In the scenario where pre-existing models that meet the defined constraints are not available, training new models becomes mandatory. On the other hand, it is possible to find datasets that can be used to train one or both models. Also, it is worth mentioning that the possibility of reusing a pre-trained model or a training dataset may lead to changes in the categories chosen at step Extraction Requirements Definition to make it possible to take advantage of the available resources and shorten the development process.

Once we obtain the NER and RE models – pre-trained or through training – they can be used to extract NE and relations present in a test dataset. The extractions must be validated by the TF analysts, and, if one or both models have proved inadequate, the training must be redone, or the ready-made models must be replaced.

D. ANALYSIS ENVIRONMENT SETTING

The OLAP Interface module is designed to allow users to query and analyze data under the established domain during the Domain Analysis step, which is described by the Analysis Questions. To achieve this level of integration, these questions must be mapped into functionalities in the OLAP Interface.

The Analysis Environment Setting step consists of specifying the necessary settings of the OLAP tool to both functionally and visually enable the DSS to support analysts in answering all the analysis questions. This task includes creating adequate metrics, data visions, and dashboards to allow the required filtering, aggregations, and report production.

To ensure that data is manipulated according to the requirements, the analysis environment must be instantiated considering the conceptual schema formulated for that domain. Also, as the configuration of the environment depends on the data schema according to which the information will be stored in the DW, the proposed dimensional schema must be considered at this step.

As illustrated in Figure 2 and detailed in Figure 6, the proposed dimensional schema is a two-fact star schema, which will store the output of the IE Module. The `Entity Dimension` will store each recognized entity (`Entity name`) in the text, and its corresponding category and supercategory (`Entity category` and `Entity supercategory`). Instances of the `Entity category` attribute, for example, correspond to the named classes specified in the examples given in Figure 4 (a), while instances

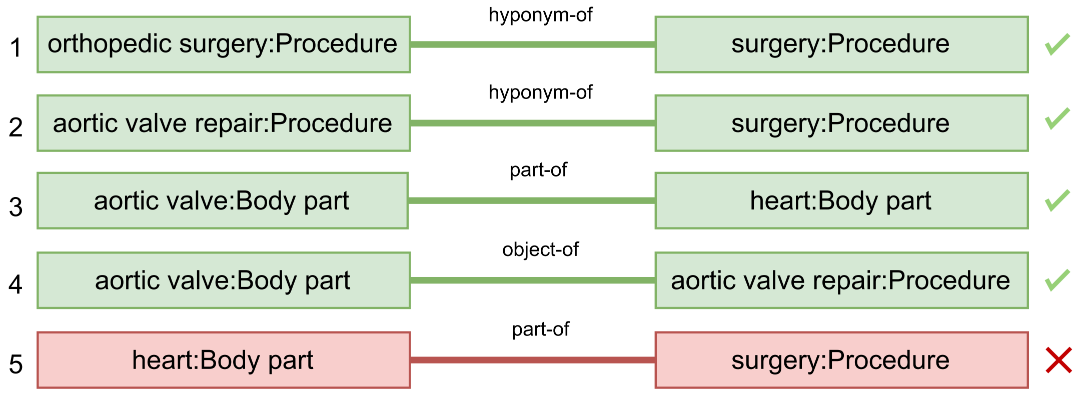


FIGURE 5. Instantiation examples for the human medicine study domain.

of the `Entity supercategory` attribute are always the labels of the corresponding superclasses (Figure 4(b)).

The Relation Dimension will store each predefined relation (`Relation name`), and its corresponding category (`Relation type`). For example, the `Relation name` attribute will assume values corresponding to the names of the classes in Figure 4(a), which are specializations of Relation, while the `Relation type` attribute will assume values according to the relation corresponding superclass name, i.e., `paradigmatic` or `syntagmatic` (b).

The remaining dimensions present attributes to appropriately describe and provide important TF context for facts from the document, date, and geographic perspective.

At the Analysis Environment Setting step, the dimension attributes may be adjusted to match the specific demands of the domain. For instance, if additional information regarding date units is available and relevant to the intended analyses, attributes such as day or month must be added to the Date Dimension. Likewise, if the analyses demand to consider factors such as the population or climate of the countries, the Geographic Dimension must be augmented with the corresponding attributes.

E. SYSTEM ASSEMBLY

The main objective of the System Assembly step is to implement the DSS system for the TF process in focus. As illustrated in Figure 1, an under-construction DSS encompasses some software components: a Pre-Extraction Module, an IE Module, a Transformation Module, a DW, and an OLAP Interface. Thus, in addition to instantiating the appropriate NER and RE models, the TForMIX method includes developing and integrating other parts to obtain the complete system suited to the needs of each TF. Before presenting the activities involved in these implementations and to achieve a better understanding of what they require, we present in more detail the functioning of the DSS modules.

The Pre-Extraction Module extracts the technical documents' textual content and bibliometric data. It delivers the former to the NER and RE models and the latter, via ODS, to the Transformation Module. The Extraction Module receives textual data as input and applies the NER and RE models to obtain information about entities, relations, their occurrences in texts, and related attributes as their categories and metacategories. This information is stored in the ODS, which is formatted according to the requirements expected by the Transformation Module.

The Transformation Module needs to adapt all structured data in the ODS, both bibliometric and extracted by NER and RE models, for storage in the DW. It relies on the dimensional schema and adapts data to conform to it. Also, before providing data to the DW loading phase, this module performs cleansing and aggregation tasks, which include identifying which terms recognized as entities in different occurrences must be represented as a single entry in the Entity Dimension. The Load module is responsible for transferring the information extracted during the use of the DSS into the DW. The module is tasked with loading data, which are instances of the Entity and Document Dimensions, as well as the Fact tables. As the Transformation Module delivers these data with the necessary cleaning already carried out and in a table format very close to that used in the DW, the Load Module is only responsible for executing the load routines consistent with the DW technology chosen for the DSS being implemented. The OLAP component encompasses the DW that stores data according to the proposed dimensional schema and the OLAP Interface, which had its functionality requirements listed during the **Analysis Environment Setting stage**. Once the DSS functionality has been clarified, the System Assembly macroprocess must be detailed. It comprises a series of activities aimed at developing and integrating all DSS modules according to the characteristics presented, producing as output a fully functional DSS. The first activity aims to develop the Pre-Extraction Module, which includes analyzing the data formats delivered by data sources and implementing a tool to convert data to the input format expected by NER and RE models. In the second activity,

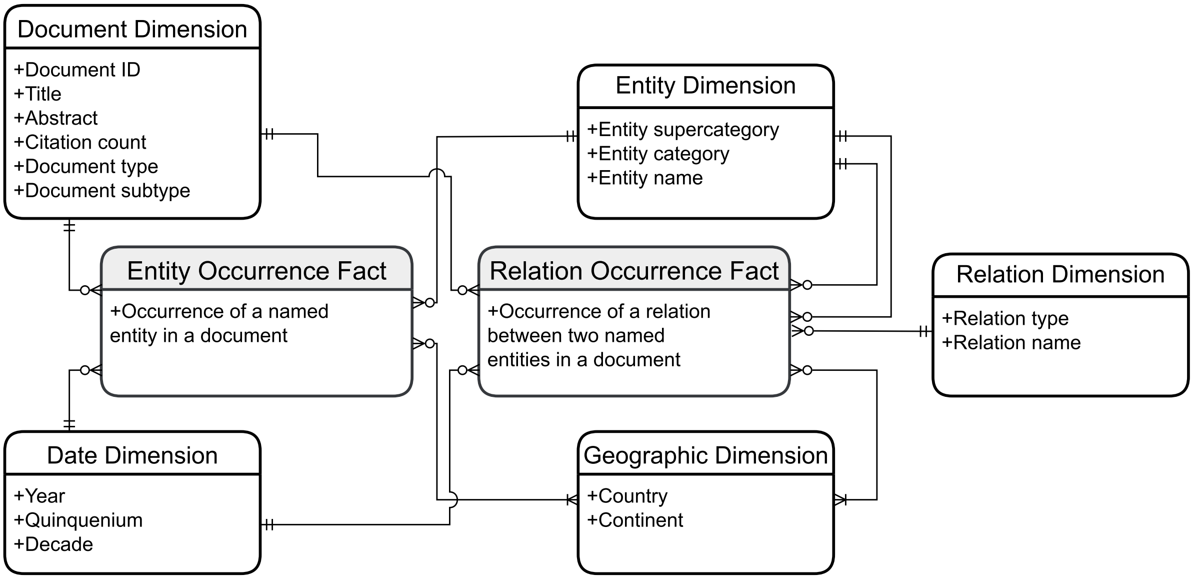


FIGURE 6. Dimensional Schema of data stored in the DW of the DSS.

the instantiated NER and RE models are wrapped together to complete the development of the Extraction Module. As the RE model depends on the entities recognized by the NER model, this wrapping may demand assembling custom language processing pipelines and creating routines to iterate over pairs of entities extracted by the NER model to infer and classify relations between them. This activity also includes developing the routines to format and load the extracted data into the ODS.

The third activity consists of developing the Transformation Module. As a primary task, it includes developing cleansing and aggregation routines to detect different extracted terms that represent the same concept, so they must refer to the same entry in the Entity Dimension corresponding table in the DW. This may demand procedures to transform terms to their singular forms, treat acronyms, and compile a synonyms list for the subject area. Some study domains may also demand developing software pieces to treat special classes of terms, for example, chemical compound names. Besides that, another transformation routine that might be implemented, according to the requirements for the domain, relates to information to be stored in other dimension tables – for instance, country names normalization and date range conversion. This step also includes collecting and preparing the data to be loaded in Date, Geographic, and Relation Dimensions, as they are not obtained during the DSS execution but are previously available.

The fourth activity is the Load Module implementation, which depends on the solutions chosen for the DW and the OLAP tool. This step consists of performing the necessary setup so the tool will correctly load the data provided by the Transformation Module into the DW on top of which the OLAP tool will run. To finalize system assembly, the settings determined in the Analysis Environment Setting step are implemented in the OLAP tool, resulting in the OLAP Interface.

It is worth mentioning that, despite demanding customization, the modules exhibit relatively similar functionality across different domains. Indeed, once the development of a DSS reaches its conclusion, the System Assembly steps of the subsequent efforts (development for other TF processes) are facilitated through the potential reuse of already developed modules. This reusability severely optimizes the development process, while underscoring the scalability and efficiency of the methodology. It also promotes consistency in constructing the TF processes, allowing for a more agile and economical approach that easily adapts the methodology to different TF analytical needs and domains. Therefore, the cumulative experience earned from previous DSSs contributes to a more robust and versatile TF development.

V. EXPERIMENT AND RESULTS

To demonstrate the applicability of TForMIX, we present its application to produce a DSS tailored for a specific TF need of a specific subject. We tested the generated system on an experiment that consisted of four specialists using it to support the conduction of a TF process on the selected subject, followed by completing an evaluative questionnaire.

A. IMPLEMENTATION OF THE METHOD

To carry out the experiment, we executed the proposed method to build a DSS aimed at supporting TF processes concerning the subject “Machine Learning for Natural Language Processing” based on scientific articles about the topic.

TABLE 2. Analysis questions for the DSS.

ID	Question	Example
Q1	What are the mentioned subtypes, parts, and characteristics of a particular concept?	"What are the mentioned parts of `Information Extraction`?"
Q2	Which documents mention a particular concept or any concept of a particular category?	"Which documents mention `ELMo`?"
Q3	Which documents mention the use of a particular core concept or any core concept of a particular category, to perform a particular action or any action?	"Which documents mention the use of any core concept to perform `Topic Modeling`?"
Q4	How did the number of references to a particular concept or any concept in a particular category evolve yearly?	"How did the number of references to evaluation metrics evolve yearly?"
Q5	How did the number of references to an action as the object of use of a central concept evolve annually? (It must be possible to select a particular action, any action of a particular category, or any action. Similarly, for the central concept.)	"How did the use of the `bilingual dictionary` material for the `lexical approach` method evolve yearly"
Q6	How did the number of references to a particular concept applied to evaluate another particular concept evolve annually?	"How did the number of references to `accuracy` applied to evaluate by `LSTM Model` evolve annually?"
Q7	For a particular country and period, what are the most mentioned concepts of a particular category?	"What are the most mentioned tasks in Brazil between 2015 and 2019?"
Q8	For a particular country and period, what core concepts were employed to perform a particular action (or any action of a particular category or any action), and in which documents do they occur?	"In 2022, which core concepts are most often mentioned as being used for the `text classification` task?"
Q9	For a particular country and period, what actions are mentioned as objects of use of a particular core concept (or any core concept of a particular category or any core concept), and in which documents do they occur?	"In the last year, which actions involved the use of any ML method?"
Q10	For a particular country and period, which entities are mentioned as subtypes, parts, and characteristics of a particular entity or any entity of a certain category?	"Between 2018 and 2021, in the US, which are the five entities most mentioned as subtypes of a method?"

This topic was chosen due to the convenient availability of a dataset for training the NER and RE models: the SCIERC dataset[1] [14], that contains NLP and Computer Vision paper abstracts, with annotations of NE and relations. This allowed the implementation of the method without manually annotating examples.

In the **Domain Analysis** step, we detected that the DSS should be able to support the identification of trends on: objectives of applying ML to NLP; ML methods employed; materials and evaluation metrics used. It should treat concepts that deal with central objects of study of the domain (core concepts) and concepts that express activities that result in an effect in the context of the domain (actions) distinctly. It should also support the identification of subtypes, subdivisions, and characteristics of the concepts. To allow for temporal and spatial examinations, the system should utilize the documents' year and country of publication, and, to facilitate the visualization of the results, the OLAP Interface should additionally provide the title and abstract of the documents. Based on these needs, we prepared the analysis questions for the system listed in Table 2.

In the **Extraction Requirements Definition** step, the instantiation of the Conceptual Meta-schema, considering the enumerated analysis questions, originated the Categories and Relations Conceptual Schema presented in the dotted frames (b) and (c) of Figure 4. The Schema contains Method, Metric, and Material as Personality classes, and Task as an Action class. The Schema is complemented by the Other Scientific Term class, whose instances are meaningful scientific terms other than methods, metrics, materials, and tasks. The Relations `Used-for` and `Feature-of` are Syntagmatic, while `Part-of` and `Hyponym-of` are Paradigmatic, meaning they can only occur between entities of the same category.

As there were no suitable NER and RE models available for the needs imposed by the analysis questions, it was necessary to train both models during the **NER and RE Models Instantiation** step. To train the NER and RE models, we employed the aforementioned SCIECR dataset, which provides the essential categories of entities (material, method, metric, task, other scientific terms) and relations (used-for, feature-of, part-of, hyponym-of).

For the NER model, we opted to fine-tune a BERT model using a transformers-based NER from the open-source Python library spaCy.[2] The uncased version of SciBERT [19], a BERT model specifically trained on scientific texts, was used as the language model. The training process employed the hyperparameters from the original library configuration.

For the RE model, a custom relation extraction component was also built using spaCy to enable the use of a single language processing pipeline with customized NER and RE components to process the texts, in the subsequent assembly of the Information Extraction component. We customized a spaCy project template named rel_component,[3] assessing all

[1] http://nlp.cs.washington.edu/sciIE/data/sciERC_raw.tar.gz

[2] https://spacy.io

[3] https://github.com/explosion/projects/tree/v3/tutorials/rel_component

TABLE 3. Precision, recall, and F1-score values obtained by the RE model for some threshold values.

Threshold	Precision	Recall	F1-score
0.00	0.70	100.0	1.40
0.05	19.50	36.88	25.37
0.10	25.36	32.90	28.52
0.20	31.87	28.52	29.99
0.30	36.61	25.68	30.10
0.40	40.71	23.35	29.60
0.50	44.89	21.39	28.88
0.60	48.99	18.98	27.26
0.70	53.08	16.04	24.52
0.80	58.76	12.96	21.09
0.90	67.01	8.08	14.30
0.99	49.91	0.31	0.60

pairs of entities within a range of 100 tokens from each other, taking into account the restrictions imposed by the relations categorized as paradigmatic according to the Conceptual Schema.

The chosen approaches required the use of a single dataset, containing the NER and RE examples, to train both models. Also, as the SCIERC dataset contains NE overlaps, which the spaCy NER model utilized does not accept, the overlapped entities were discarded. The resulting dataset contained 446 paper abstracts, annotated with 5.820 NE and 3.391 relations of interest.

The final datasets were converted from the original file format to spaCy binary files required by the chosen spaCy tools. Each spaCy binary file is composed of spaCy Docs containing the texts and the position and category information of the entities (in the `ents` attribute) and relations (in the `_.rel` attribute).

To estimate the performance of the trained models over unseen data, we performed a 10-fold cross-validation and evaluated the results considering the F1-score measure. The total dataset was split into training (284 documents), validation (72 documents), and test (90 documents) sets. The mean precision, recall, and F1-score obtained by the NER model were 0.66, 0.67, and 0.67, respectively. The performance of the RE model depends on an additional threshold, chosen to lower bound the confidence score of the relations to be considered. Table 3 shows the mean F1-score for the considered threshold values. To train the NER and RE models to be used in the DSS, we split the total dataset into training (356) and validation (90) document sets.

In the **Analysis Environment Setting** step, the analysis questions were rephrased to unveil the filtering, grouping, and displaying needs of the analysis environment in light of the conceptual metaschema and attributes of the dimensional schema. For example, to support Q9, the interface should allow selecting one or more instances of Personality or one or more categories of Personality and list the instances of Action that are objects of a `used-for` relation whose objects are the Personality instance(s) selected and the documents (title and abstract) in which they occur. In addition, it must allow filtering by country and year of the documents in which the mentioned relation occurs.

Based on the joint analysis of the rephrased questions, we established that the analysis environment of the produced system should have four different panels: Top Occurrences, that shows the NEs (or Entity instances) that occur most frequently (Q7); Entities Information, for analyses focused on the set of Entity instances (Q2, Q4); Relations Information, for analyses focused on the set of Relation instances (Q3, Q5, Q6, Q8, Q9, Q10); and Entity Exploration, for analyses focused on a specific Entity instance, exploring its hyponyms, parts, and characteristics (Q1).

To exemplify the functionality of the panels, Figure 7 shows the Top Occurrences panel. The left boxes contain tools to filter the NE occurrences by country and year and to enable drilling the Entity Dimension to explore each entity category or supercategory separately. From the TF perspective, this functionality allows analysts to look for elements in texts not by the mere occurrence of text strings but by the role the terms play in the contexts they occur. For example, it is possible to explore elements that are mentioned as materials in the documents. The graph on the right side shows the NEs ordered by the number of documents in which they occur. The remaining panels are illustrated in Figures 8 to 11 in Appendix A.

In the **System Assembly** step, we developed a pre-extraction module properly set to obtain from the Scopus database's export files: the required bibliometric data (document identifier, title, citation count, publication type, year, and authors' country) and the set of abstracts that will serve as a basis for the textual information extraction. The former is stored in a staging area, and the latter is passed to the IE Module.

We developed the IE Module, which builds a language processing pipeline using the trained NER and RE models. The module passes each abstract text through the pipeline, getting the information about the NE and relations present in the text. The relation information consists of matrices of pairs of entities containing a value between 0 and 1 for each existing relation, denoting the probability of that pair of entities being related by each type of relation. The name and category of the identified entities are stored in a staging area, together with the document identifier. Also, for each pair of entities found, it is checked whether, for each type of relation, the assigned probabilities of existing relationships between the entities are higher than a specific threshold. If this is true, information about the occurrence of the relations in the document (with the head and tail entities, the relation type, and the document identifier) is also stored in the staging area.

The choice of the threshold considered that, for vast domains such as that of the DSS in question (use of ML for NLP), the analyses occur over numerous documents. Thus, to avoid too many false positives being displayed in the analyses, it is necessary to prioritize the precision of the extraction. We analyzed the performance of the RE model in

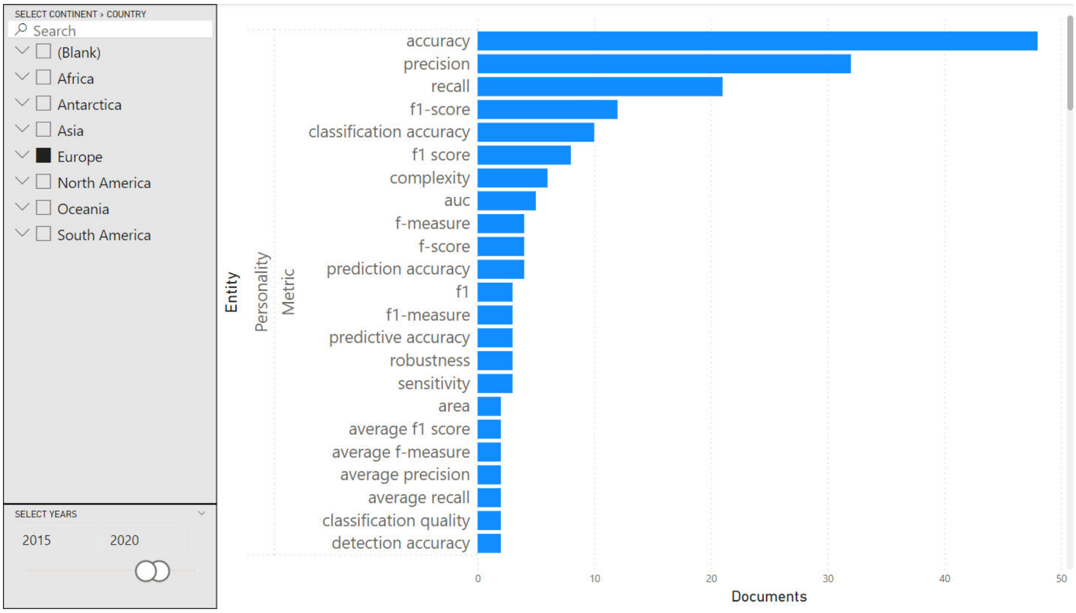


FIGURE 7. Top Occurrences panel of the OLAP Interface.

the cross-validation conducted in the **NER and RE Models Instantiation step**. For each threshold value, we calculated the F-beta score between precision and recall using beta = 0.25. Following that, the threshold that maximized this measure was 0.8.

This step also included the development of the Transformation Module, which comprises routines to: reduce entity terms to singular; assign identifiers to NE and relations; organize their occurrence as entries to fact tables and their attributes as entries to the respective dimension tables; retrieve documents' publication year, title, and abstract, and organize them to store in date and document dimension tables; and retrieve the country of each author of the documents, preparing them for storage in the geographic dimension table and the mapping data between the geographic dimension and fact tables. Finally, it stores all data in separate data frames, which are exported to files in Comma Separated Values (CSV) format to serve as inputs for the Load Module.

For this experiment, we chose Power BI as the OLAP tool. Therefore, the Load Module, DW, and OLAP Interface components of the DSS work within Power BI, in which we implemented the settings determined in the Analysis Environment Settings step.

B. EXPERIMENT EXECUTION

We submitted the produced system for testing by four TF experts from the Brazilian Army's Technological Information and Management Agency (AGITEC). Each expert used the system to analyze the knowledge domain of ML applied to NLP to answer the analysis questions listed in Table 2. The analyses were based on titles and abstracts of 4714 scientific papers extracted from the Scopus database using the search string: (TITLE-ABS-KEY ("natural language processing") AND TITLE-ABS-KEY ("machine learning")) AND PUBYEAR > 2013 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA, "ENGI")) AND (LIMIT-TO (LANGUAGE, "English")). From these documents, the system's IE Module extracted the quantities of entities and relations shown in Table 4. Since the RE model was not able to identify feature-of relations in the texts (which can be explained by the low number of examples provided in the training step), we removed this relationship from the Entity Exploration panel.

TABLE 4. Quantities of entities and relations extracted by the IE Module.

Type	Name	Amount
Entities	Method	39573
	OtherScientificTerm	22508
	Generic	15401
	Material	15036
	Task	14004
	Metric	4175
Relations	Used-for	6357
	Evaluate-for	1705
	Hyponym-of	3180
	Part-of	31
	Feature-of	0

We also asked each expert to answer an evaluation questionnaire available in the supplementary material. We developed the questionnaire following proposals from Rogers et al. [28], which established a list of goals for the usability of interactive systems: effectiveness, efficiency, safety, utility, learnability, and memorability. The authors also proposed practical techniques for determining how well a system is designed according to those goals. One of the techniques presented is the application of questionnaires to users, with the recommendation to present statements to be scored according to rating scales containing no more than five points. The resulting questionnaire is available in Appendix B.

C. ANALYSIS OF THE RESULTS

The group of evaluators was composed of four participants who work with general TF (i.e., not focused on a specific technology domain). Conducting this experiment, at first hand, poses a unique set of challenges, primarily from the requirement of highly specialized individuals. Regarding their length of experience with TF processes, one participant had accumulated over seven years of expertise in the area, another one had between three and seven years, a third had between one and three years of experience, and the last one had been working with TF for less than one year. In this manner, the group was composed of individuals possessing a diverse range of experience levels.

Before offering their opinions about the DSS itself, the analysts evaluated the representativeness of the proposed analyses concerning the tasks carried out during a TF process based on scientific articles. This aspect is highly relevant to the research because, since the analysis questions were imposed on the evaluators as a roadmap for exploring the DSS, it was necessary to ensure that these questions actually reflect the usual analytical requirements of TF processes. Considering this, half of the evaluators fully agreed, while the others partially agreed that the proposed analysis questions represented the tasks conducted in TF processes. Thus, it is reasonable to conclude that the proposed analysis achieved good representativeness.

One of the evaluators who partially agreed with this representativeness provided an additional comment that clarified the reason for not fully agreeing. This evaluator stated that a full TF process includes tasks for which the DSSs in question may not be used and that this arises from the inherent nature of these tasks and their specific positions within the overall process. On the other hand, the evaluator stated that a series of analyses, which are not present in the analysis questions, can be easily carried out using this system. As complemented by the evaluator, "The process is much more than 'data', and the system is applicable in only one facet of the process, by merging 'manual' techniques, such as literature review, with automated techniques such as data processing, for which the system applies perfectly."

The objective evaluation of the system included the following aspects: effectiveness, efficiency, reliability, usefulness, and learnability. Regarding the system's general effectiveness, the evaluators assessed its ability to support them in answering the analysis questions. These questions focused on exploring relevant concepts located in the text and meaningful semantic relationships among their entities, complemented by the bibliometric data. Two evaluators agreed, and two partially agreed that the system allowed for carrying out the proposed analyses. As these analyses were considered representative of usual TF analytical demands, it means that applying NER and RE to recognize relevant concepts and how they are semantically connected in texts to allow exploring this information in a multidimensional manner can be considered effective in supporting analysts in conducting their tasks.

TABLE 5. Opinions about the estimated time to perform the proposed analyses.

Evaluator		Estimated time for analyses	
ID	Experience	with the DSS	without the DSS
1	>7 years	3 to 7 days	>7 days
2	<1 year	1 to 3 days	3 to 7 days
3	1 to 3 years	1 to 3 days	3 to 7 days
4	3 to 7 years	>7 days	>7 days

The evaluators also specifically assessed the effectiveness of using categories and metacategories. Among them, three fully agreed while one expressed partial agreement regarding the categories (Method, Material, Metric, and Task) that contributed to the expansion of analytical capacity in the tasks. Regarding the contribution of using metacategories (Personality and Action), all evaluators unanimously expressed that they fully agreed with their contribution to the expansion. This result highlights the relevance of applying the NER technique in the TF context. Furthermore, it confirms that the use of the metaschemas contributes positively to the DSS, expanding the possibilities of analysis offered by the NER categories.

As for the efficiency of the system, three evaluators partially agreed, while one fully concurred that, once users have learned how to use the system, it speeds up the fulfillment of the proposed analyses. In addition, two evaluators fully endorsed and two partially agreed that the system allows for carrying out the proposed analyses. To complement the evaluation of the system's efficiency, we established a comparison between the informed estimated time to perform the proposed analyses with and without the use of the system, based on the evaluators' experience in conducting TF processes. Table 5 presents the result, which shows that three of the four evaluators estimate that using the system can reduce time spent on the analyses. The proposed time intervals chosen for the answers did not allow this aspect to be analyzed for the fourth evaluator, since the provided estimations were "more than seven days" both with and without the system.

In addition to this positive finding for efficiency, it is noteworthy that, in general, the DSS was considered useful by the evaluators, since it offers the correct type of functionality to allow users to accomplish their objectives. Three evaluators totally and one partially agreed that the system provided a set of functions that facilitated the fulfillment of the proposed analyses. When faced with the statement "You would consider replacing the manual analysis of documents with the use of the system,'' one evaluator totally agreed, two partially agreed, and the last one totally disagreed. The additional comments provided by this last evaluator helped clarify the reason for the opposing opinion since they stated that the TF process includes tasks to which the system in question is not applicable.

The system's learnability aspect was also considered positive. Two evaluators partially agreed, and two totally agreed that learning how to use the system is not costly in terms of time or effort. Thus, once the system is ready, the cost of learning to operate it can be offset by the agility it provides to the analysis execution. However, this advantage is contrasted by one of the additional comments provided. One evaluator raised a concern, highlighting the effort required to create the datasets for training the NER and RE models as a potential obstacle to fully replacing the manual analyses with the usage of the systems since this step can potentially reduce the time required for the initial study to understand the target theme of the TF. Whether this reduction is compensated by the time spent on the annotation effort to create the dataset, however, can not be determined at this time.

Another evaluator reported in the additional observations that there are some improvements to be implemented in the presentation of the analyses. It is important to note that this observation is more related to implementation problems than to the system's idea and the method itself. However, although the focus of this research is to validate the proposed method and not the quality of the present implementation, this report is valid to point out that problems in the implementation can compromise the system's performance – a relevant consideration to most computer systems.

Finally, to evaluate the reliability aspects of the system, the evaluators were confronted with the statement ''The system generates consistent results for the proposed analyses.'' Three evaluators totally agreed, and one partially agreed with this statement. Thus, we can assume that the system was considered reliable for the intended usage, meaning the user trusts the information it provides as a result of the performed query operations.

In summary, the experiment confirmed the applicability in TF processes of DSSs built according to the proposed method, demonstrating that the combined use of NER and RE, together with adequate dimensional modeling, effectively supports part of the tasks involved in these processes. In particular, categorizing entities into classes relevant to the target domain showed its potential to assist analyses by reducing the time required to process data manually. The experiment also proved that having more than one level of hierarchy for entities is useful. This finding justifies using the additional layer of specialization in the metaschema through the Personality, Action, and Characteristic metaclasses, which were designed to expand the possibilities of grouping and filtering when querying entities located in texts, bringing new analysis perspectives.

While the analysts' opinions make clear the impossibility of the system completely replacing a human analyst in the TF process and point out the need for improvements in the implementation of the DSS sample used in the experiment, in general, the proposed method is capable of producing a DSS with good usability and that contributes positively to the conduct of TF analytical processes.

VI. CONCLUSION

TF processes demand a great effort from analysts, as they usually involve analyzing a large set of documents for each domain under investigation. The related literature shows that adopting DW based systems in combination with IE techniques improves corpora analysis [7], [8], [9]. However, these works report on an experience of building a domain-specific DW and do not provide a domain-independent approach to guide developers on the construction of a DW based system for corpora analysis. On the other hand, some works [10], [11], [12] have concentrated on using IE techniques in TF processes, but have not combined them with a DW based system, thus failing to offer an analytical view of the corpus.

To fill this gap, this work presents TForMIX, a novel method to produce tailor-made DSS systems. It applies NER and RE, in conjunction with dimensional modeling techniques, to support TF analytical processes based on technical documents such as scientific articles and patents. Also, TForMIX is based on a conceptual metaschema that supports both collecting the requirements for extraction and exploring the information extracted. This metaschema works as an important component of the method, facilitating both the initial setup and the subsequent data analysis. In addition, this work provides a set of typical analytical queries (Q1-Q10) based on the metaschema, which may be useful for most applications of the method, i.e., for the execution of other TF processes. Finally, the practical experiment with TF analysts has yielded promising results, showing that the method produces DSSs that contribute positively to TF analytical processes.

Future research challenges include applying the complete methodological process to experimentation to ascertain whether the advantages gained from the analyses mitigate the time and resources invested in the system assembly. Moreover, it is also worth investigating alternative RE techniques to obtain a more comprehensive set of extracted relations, which would allow for a better quality in the analysis since RE is a key component in the method. Furthermore, additional case studies can be extended using patent documents, as they provide another perspective on research and development activities in several areas, enabling a more comprehensive analysis of the relationships between different concepts and entities.

APPENDIX A
OLAP INTERFACE

Figures 8 to 11 show the four panels in the OLAP Interface of the DSS produced for the experiment. The gray areas in the panels provide filtering tools for the analyses, while the remaining spaces contain the results.

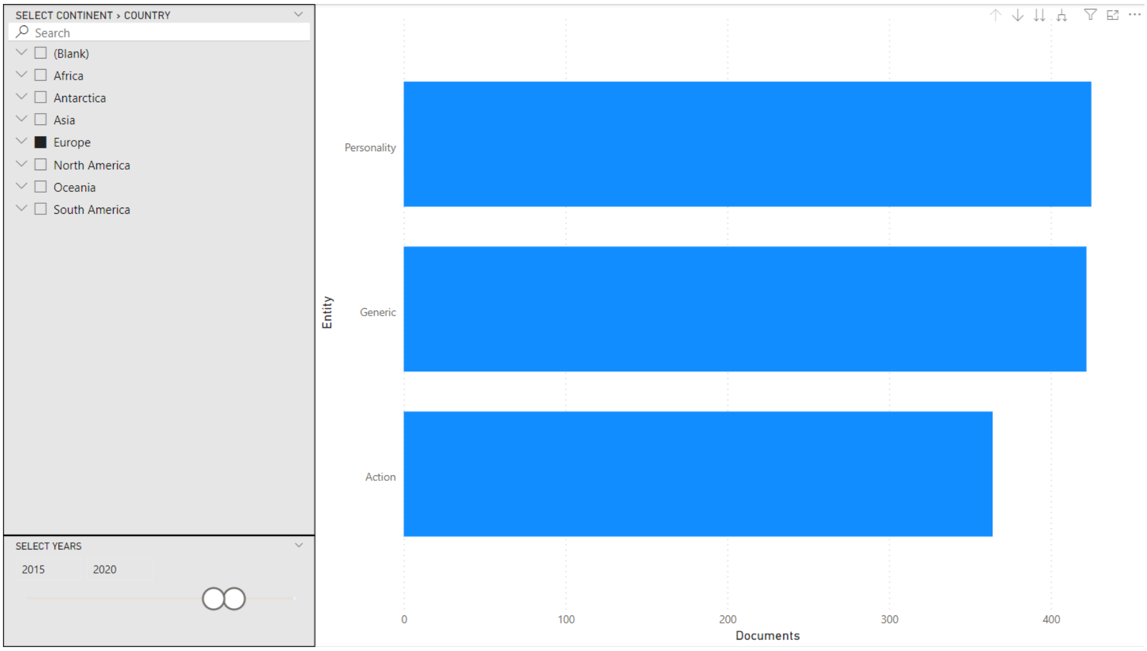
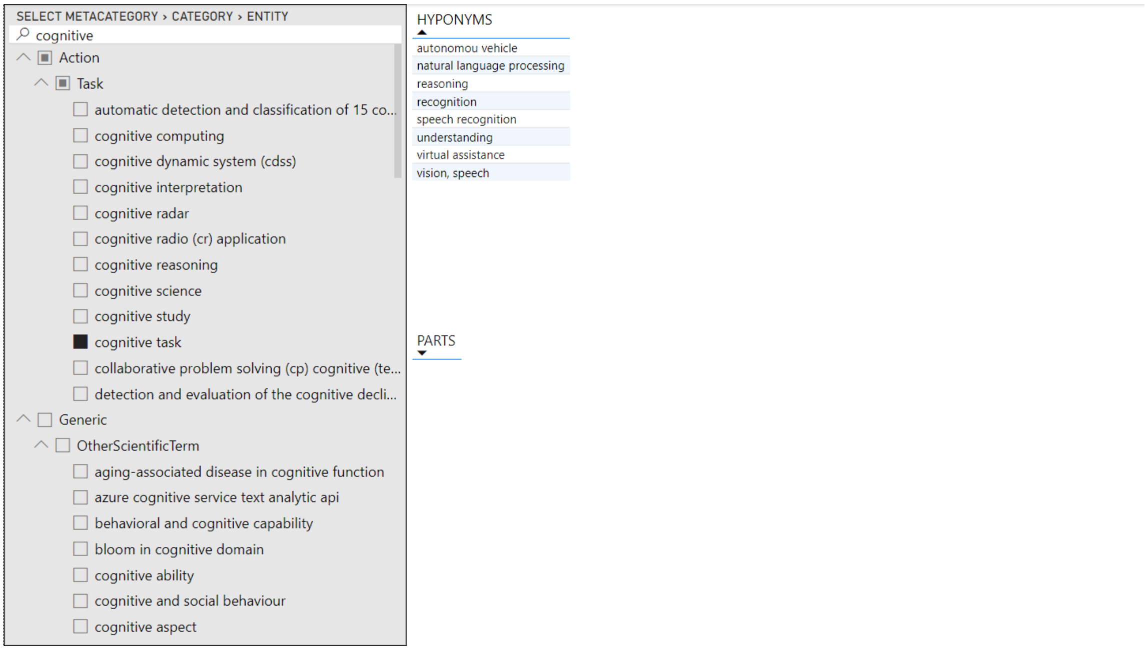


FIGURE 8. Top Occurrences panel of the OLAP Interface.


SELECT METACATEGORY › CATEGORY › ENTITY

cognitive

- ▣ Action
  - ▣ Task
    - ☐ automatic detection and classification of 15 co...
    - ☐ cognitive computing
    - ☐ cognitive dynamic system (cdss)
    - ☐ cognitive interpretation
    - ☐ cognitive radar
    - ☐ cognitive radio (cr) application
    - ☐ cognitive reasoning
    - ☐ cognitive science
    - ☐ cognitive study
    - ■ cognitive task
    - ☐ collaborative problem solving (cp) cognitive (te...
    - ☐ detection and evaluation of the cognitive decli...
- ☐ Generic
  - ☐ OtherScientificTerm
    - ☐ aging-associated disease in cognitive function
    - ☐ azure cognitive service text analytic api
    - ☐ behavioral and cognitive capability
    - ☐ bloom in cognitive domain
    - ☐ cognitive ability
    - ☐ cognitive and social behaviour
    - ☐ cognitive aspect

HYPONYMS

autonomou vehicle
natural language processing
reasoning
recognition
speech recognition
understanding
virtual assistance
vision, speech

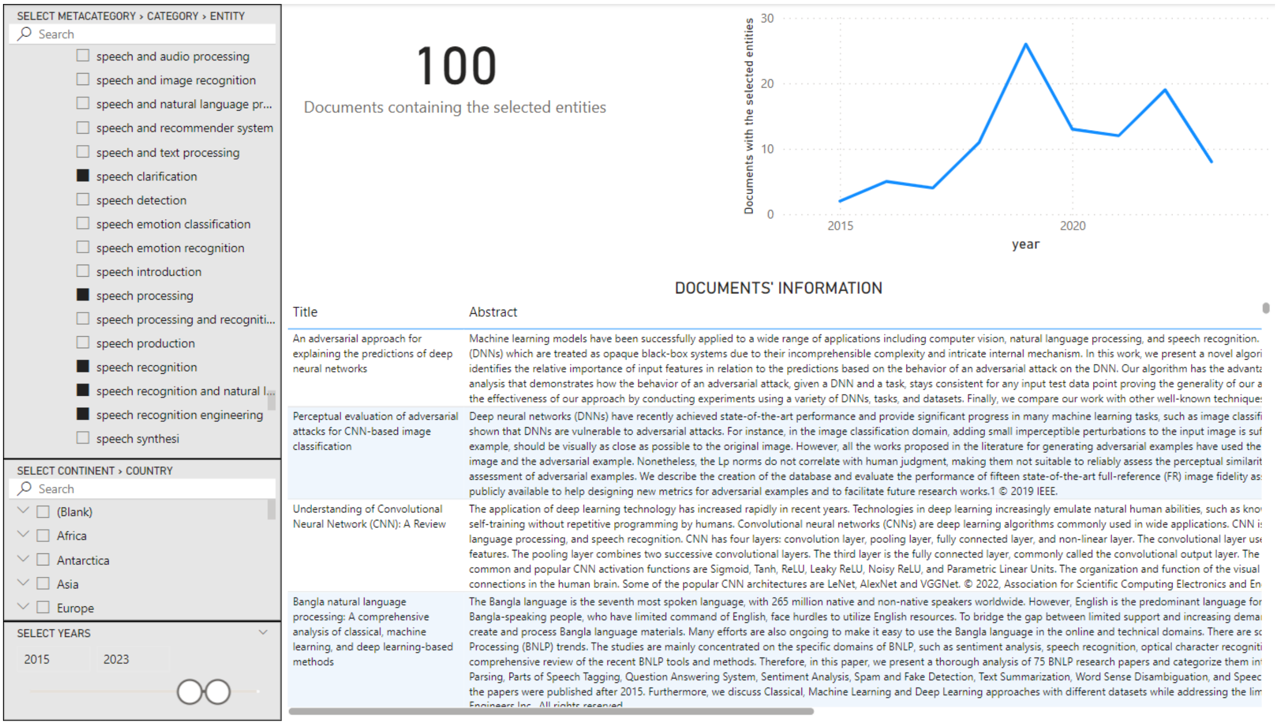
PARTS

FIGURE 9. Entity Exploration panel of the OLAP Interface.

APPENDIX B EVALUATION QUESTIONNAIRE

The questionnaire applied in the experiment was divided into three main sections: (i) questions about the interviewee; (ii) evaluation of the proposed analyses; and (iii) evaluation of the system. In (ii), the interviewees were asked to evaluate the validity of the proposed analyses through a list of analysis questions. In (iii), they were asked to evaluate the system based on their experience of using it to answer the analysis questions.

For questions that inquired about agreement with a statement, we provided four options: I totally agree, I partially agree, I partially disagree, and I totally disagree. The interviewees were instructed to select the option "I partially agree" when they consider that they agree with the statement in a proportion greater than 50% and less than 100%;



SELECT METACATEGORY › CATEGORY › ENTITY

Search

□ speech and audio processing
□ speech and image recognition
□ speech and natural language pr...
□ speech and recommender system
□ speech and text processing
■ speech clarification
□ speech detection
□ speech emotion classification
□ speech emotion recognition
□ speech introduction
■ speech processing
□ speech processing and recogniti...
□ speech production
■ speech recognition
■ speech recognition and natural l...
■ speech recognition engineering
□ speech synthesi

SELECT CONTINENT › COUNTRY

Search

□ (Blank)
□ Africa
□ Antarctica
□ Asia
□ Europe

SELECT YEARS

2015 2023

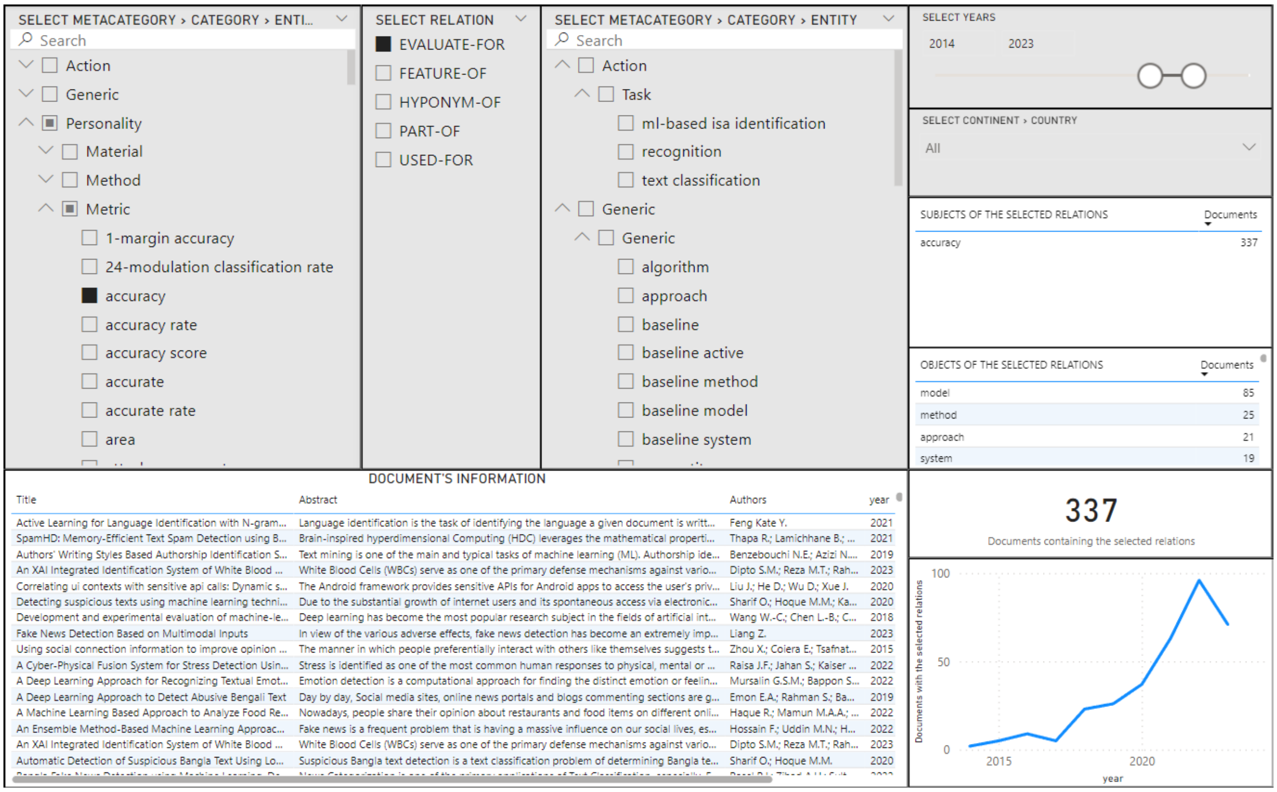
100

Documents containing the selected entities

DOCUMENTS' INFORMATION

Title	Abstract
An adversarial approach for explaining the predictions of deep neural networks	Machine learning models have been successfully applied to a wide range of applications including computer vision, natural language processing, and speech recognition. (DNNs) which are treated as opaque black-box systems due to their incomprehensible complexity and intricate internal mechanism. In this work, we present a novel algori identifies the relative importance of input features in relation to the predictions based on the behavior of an adversarial attack on the DNN. Our algorithm has the advanta analysis that demonstrates how the behavior of an adversarial attack, given a DNN and a task, stays consistent for any input test data point proving the generality of our a the effectiveness of our approach by conducting experiments using a variety of DNNs, tasks, and datasets. Finally, we compare our work with other well-known technique
Perceptual evaluation of adversarial attacks for CNN-based image classification	Deep neural networks (DNNs) have recently achieved state-of-the-art performance and provide significant progress in many machine learning tasks, such as image classifi shown that DNNs are vulnerable to adversarial attacks. For instance, in the image classification domain, adding small imperceptible perturbations to the input image is suf example, should be visually as close as possible to the original image. However, all the works proposed in the literature for generating adversarial examples have used the image and the adversarial example. Nonetheless, the Lp norms do not correlate with human judgment, making them not suitable to reliably assess the perceptual similarit assessment of adversarial examples. We describe the creation of the database and evaluate the performance of fifteen state-of-the-art full-reference (FR) image fidelity as publicly available to help designing new metrics for adversarial examples and to facilitate future research works.1 © 2019 IEEE.
Understanding of Convolutional Neural Network (CNN): A Review	The application of deep learning technology has increased rapidly in recent years. Technologies in deep learning increasingly emulate natural human abilities, such as kno self-training without repetitive programming by humans. Convolutional neural networks (CNNs) are deep learning algorithms commonly used in wide applications. CNN i language processing, and speech recognition. CNN has four layers: convolution layer, pooling layer, fully connected layer, and non-linear layer. The convolutional layer use features. The pooling layer combines two successive convolutional layers. The third layer is the fully connected layer, commonly called the convolutional output layer. The common and popular CNN activation functions are Sigmoid, Tanh, ReLU, Leaky ReLU, Noisy ReLU, and Parametric Linear Units. The organization and function of the visual connections in the human brain. Some of the popular CNN architectures are LeNet, AlexNet and VGGNet. © 2022, Association for Scientific Computing Electronics and En
Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods	The Bangla language is the seventh most spoken language, with 265 million native and non-native speakers worldwide. However, English is the predominant language for Bangla-speaking people, who have limited command of English, face hurdles to utilize English resources. To bridge the gap between limited support and increasing dema create and process Bangla language materials. Many efforts are also ongoing to make it easy to use the Bangla language in the online and technical domains. There are so Processing (BNLP) trends. The studies are mainly concentrated on the specific domains of BNLP, such as sentiment analysis, speech recognition, optical character recogniti comprehensive review of the recent BNLP tools and methods. Therefore, in this paper, we present a thorough analysis of 75 BNLP research papers and categorize them in Parsing, Parts of Speech Tagging, Question Answering System, Sentiment Analysis, Spam and Fake Detection, Text Summarization, Word Sense Disambiguation, and Speec the papers were published after 2015. Furthermore, we discuss Classical, Machine Learning and Deep Learning approaches with different datasets while addressing the lim Engineers Inc. All rights reserved.

FIGURE 10. Entities Information panel of the OLAP Interface.



SELECT METACATEGORY › CATEGORY › ENTI...

Search

□ Action
□ Generic
▣ Personality
□ Material
□ Method
▣ Metric
□ 1-margin accuracy
□ 24-modulation classification rate
■ accuracy
□ accuracy rate
□ accuracy score
□ accurate
□ accurate rate
□ area

SELECT RELATION

■ EVALUATE-FOR
□ FEATURE-OF
□ HYPONYM-OF
□ PART-OF
□ USED-FOR

SELECT METACATEGORY › CATEGORY › ENTITY

Search

□ Action
□ Task
□ ml-based isa identification
□ recognition
□ text classification
□ Generic
□ Generic
□ algorithm
□ approach
□ baseline
□ baseline active
□ baseline method
□ baseline model
□ baseline system

SELECT YEARS

2014 2023

SELECT CONTINENT › COUNTRY

All

SUBJECTS OF THE SELECTED RELATIONS	Documents
accuracy	337

OBJECTS OF THE SELECTED RELATIONS	Documents
model	85
method	25
approach	21
system	19

DOCUMENT'S INFORMATION

Title	Abstract	Authors	year
Active Learning for Language Identification with N-gram...	Language identification is the task of identifying the language a given document is writt...	Feng Kate Y.	2021
SpamHD: Memory-Efficient Text Spam Detection using B...	Brain-inspired hyperdimensional Computing (HDC) leverages the mathematical properti...	Thapa R.; Lamichhane B.; ...	2021
Authors' Writing Styles Based Authorship Identification S...	Text mining is one of the main and typical tasks of machine learning (ML). Authorship ide...	Benzebouchi N.E.; Azizi N....	2019
An XAI Integrated Identification System of White Blood ...	White Blood Cells (WBCs) serve as one of the primary defense mechanisms against vario...	Dipto S.M.; Reza M.T.; Rah...	2023
Correlating ui contexts with sensitive api calls: Dynamic s...	The Android framework provides sensitive APIs for Android apps to access the user's priv...	Liu J.; He D.; Wu D.; Xue J.	2020
Detecting suspicious texts using machine learning techni...	Due to the substantial growth of internet users and its spontaneous access via electronic...	Sharif O.; Hoque M.M.; Ka...	2020
Development and experimental evaluation of machine-le...	Deep learning has become the most popular research subject in the fields of artificial int...	Wang W.-C.; Chen L.-B.; C...	2018
Fake News Detection Based on Multimodal Inputs	In view of the various adverse effects, fake news detection has become an extremely imp...	Liang Z.	2023
Using social connection information to improve opinion ...	The manner in which people preferentially interact with others like themselves suggests t...	Zhou X.; Coiera E.; Tsafnat...	2015
A Cyber-Physical Fusion System for Stress Detection Usin...	Stress is identified as one of the most common human responses to physical, mental or ...	Raisa J.F.; Jahan S.; Kaiser ...	2022
A Deep Learning Approach for Recognizing Textual Emot...	Emotion detection is a computational approach for finding the distinct emotion or feelin...	Mursalin G.S.M.; Bappon S...	2022
A Deep Learning Approach to Detect Abusive Bengali Text	Day by day, Social media sites, online news portals and blogs commenting sections are g...	Emon E.A.; Rahman S.; Ba...	2019
A Machine Learning Based Approach to Analyze Food Re...	Nowadays, people share their opinion about restaurants and food items on different onli...	Haque R.; Mamun M.A.A.; ...	2022
An Ensemble Method-Based Machine Learning Approac...	Fake news is a frequent problem that is having a massive influence on our social lives, es...	Hossain F.; Uddin M.N.; H...	2022
An XAI Integrated Identification System of White Blood ...	White Blood Cells (WBCs) serve as one of the primary defense mechanisms against vario...	Dipto S.M.; Reza M.T.; Rah...	2023
Automatic Detection of Suspicious Bangla Text Using Lo...	Suspicious Bangla text detection is a text classification problem of determining Bangla te...	Sharif O.; Hoque M.M.	2020

337

Documents containing the selected relations

FIGURE 11. Relations Information panel of the OLAP Interface.

and, conversely, to select the "I partially disagree" option when they consider that they agree with the statement in a proportion lower than 50% and higher than 0%.

The questions of each section are listed as follows.

- Section I – Information about the interviewee
  - 1 Name
  - Do you work with general Technological Foresight or focused on a specific S&T domain? (Provided

options: General/ Specific S&T domain/ None of the above)
  - Technological Foresight experience (Provided options: less than one year/ one to three years/ three to seven years / more than seven years)
- Section II – Evaluation of the proposed analyses
  - [–] Do you agree with the statement "The proposed analyses are representative of the processes conducted in Technological Foresight tasks on scientific articles"?
- Section III – Evaluation of the system
  - ◇ About effectiveness:
    - Do you agree with the statement "The system allows the execution of the proposed analyses"?
    - Do you agree with the statement "The classification of entities into categories (Method/ Material/ Metric/ Task) contributes positively to the expansion of analytical capacity"?
    - Do you agree with the statement "The classification of entities in metacategories (Personality/Action) contributes positively to the expansion of analytical capacity"?
  - ◇ About efficiency:
    - Do you agree with the statement "Once the use of the system has been learned, it streamlines the performance of the proposed analyses"?
    - Based on your experience in conducting prospecting analytical processes, how long do you estimate it would take to carry out the proposed analyses without using the system? (Provided options: less than one day/ one to three days/ three to seven days / more than seven days)
    - Based on your experience in conducting prospecting analytical processes, how long do you estimate it would take to carry out the proposed analyses using the system? (Provided options: less than one day/ one to three days/ three to seven days / more than seven days)
    - Do you agree with the statement "Once the use of the system has been learned, it can be said that it allows the proposed analyses to be better carried out"?
  - ◇ About reliability:
    - ∗ Do you agree with the statement "The system generates consistent results for the proposed analyses"?
  - ◇ About utility:
    - Do you agree with the statement "The system provides a set of functions that indeed facilitates the realization of the proposed analyses"?
    - Do you agree with the statement "Would you consider replacing manual document review with using the system?"
  - ◇ About learnability:
    - Do you agree with the statement "It is costly to learn to use the system"?
  - ◇ Open question:
    - Please provide additional comments about the system. If possible, provide justifications for your previous answers, explaining why you agree/disagree with the statements.

REFERENCES

[1] Ö. Saritas, P. Bakhtin, I. Kuzminov, and E. Khabirova, "Big data augmentated business trend identification: The case of mobile commerce," *Scientometrics*, vol. 126, no. 2, pp. 1553–1579, Feb. 2021.

[2] K. Kim, K. Park, and S. Lee, "Investigating technology opportunities: The use of SAOx analysis," *Scientometrics*, vol. 118, no. 1, pp. 45–70, Jan. 2019.

[3] J. M. Vicente-Gomila, M. A. Artacho-Ramírez, M. Ting, and A. L. Porter, "Combining tech mining and semantic TRIZ for technology assessment: Dye-sensitized solar cell as a case," *Technol. Forecasting Social Change*, vol. 169, Aug. 2021, Art. no. 120826.

[4] R. Grishman, "Information extraction: Techniques and challenges," in *Proc. Int. Summer School Inf. Extraction*, Jan. 1997, pp. 10–27.

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

[6] P. Sawadogo and J. Darmont, "On data lake architectures and metadata management," *J. Intell. Inf. Syst.*, vol. 56, no. 1, pp. 97–120, Feb. 2021.

[7] L. Chiudinelli, M. Gabetta, G. Centorrino, N. Viani, C. Taşcă, A. Zambelli, M. Bucalo, A. Ghirardi, N. Barbarini, E. Sfreddo, C. Tondini, R. Bellazzi, and L. Sacchi, "Ontology-driven real world evidence extraction from clinical narratives," in *Proc. 17th World Congr. Med. Health Informat.*, Aug. 2019, pp. 1441–1442.

[8] G. Dietrich, J. Krebs, G. Fette, M. Ertl, M. Kaspar, S. Störk, and F. Puppe, "Ad hoc information extraction for clinical data warehouses," *Methods Inf. Med.*, vol. 57, no. 1, pp. e22–e29, May 2018.

[9] B. T. Nguyen, T. T. N. Doan, S. Thanh, K. Q. Tran, A. T. Nguyen, A. T. Le, A. M. Tran, N. Ho, T. T. Nguyen, and D. T. Huynh, "Ad hoc information extraction for clinical data warehouses," *IEEE Access*, vol. 10, pp. 87681–87697, 2022.

[10] H. Miao, Y. Wang, X. Li, and F. Wu, "Integrating technology-relationship-technology semantic analysis and technology roadmapping method: A case of elderly smart wear technology," *IEEE Trans. Eng. Manag.*, vol. 69, no. 1, pp. 262–278, Feb. 2022.

[11] G. Puccetti, V. Giordano, I. Spada, F. Chiarello, and G. Fantoni, "Deriving technology intelligence from patents: Preposition-based semantic analysis," *J. Informetrics*, vol. 186, no. 1, Jan. 2023, Art. no. 122160.

[12] V. Giordano, F. Chiarello, N. Melluso, G. Fantoni, and A. Bonaccorsi, "Text and dynamic network analysis for measuring technological convergence: A case study on defense patent data," *IEEE Trans. Eng. Manag.*, vol. 70, no. 4, pp. 1490–1503, Apr. 2023.

[13] G. M. Coelho, "Technological foresight methodologies and experiences in portuguese: 'Prospeccao tecnologica metodologias e experiencias," Technol. Trends Project, Rio de Janeiro, Tech. Rep. 12, 2003.

[14] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3219–3232, doi: 10.18653/v1/d18-1360.

[15] S. Kim, I. Park, and B. Yoon, "SAO2Vec: Development of an algorithm for embedding the subject-action-object (SAO) structure using Doc2Vec," *PLoS ONE*, vol. 15, no. 2, Apr. 2020, Art. no. e0227930.

[16] R. Shiyali and G. Malur, "Prolegomena to library classification," in *Polymers of Hexadromicon*, vol. 3. London, U.K.: Asia Publishing House, 1967.

[17] C. Nannan, C. Dehua, and L. Jiajin, "Entity recognition approach of clinical documents based on self-training framework," in *Proc. Adv. Int. Sys. Comp.*, Aug. 2019, pp. 259–265.

[18] F. Chiarello, A. Cimino, G. Fantoni, and F. Dell'Orletta, "Automatic users extraction from patents," *World Pat. Inf.*, vol. 54, no. 1, pp. 28–38, Sep. 2018.

[19] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. 9th Int. Joint Conf. Natural Lang. Process. Conf. Empirical Methods Natural Lang. Process. (EMNLP-IJCNLP)*, Nov. 2019, pp. 3615–3620.

[20] I. Dahlberg, "A referent-oriented, analytical concept theory for interconcept," *Knowl. Org.*, vol. 5, no. 3, pp. 142–151, Sep. 1978.

[21] J. M. Robert, *Database Design for Smarties: Using UML for Data Modeling.* San Mateo, CA, USA: Morgan Kaufmann, 1999.

[22] P. Daniel, *Decision Support Systems: Concepts and Resources for Managers.* Westport, CN, USA: Quorum Books, Mar. 2002.

[23] A. L. Antunes, E. Cardoso, and J. Barateiro, "Incorporation of ontologies in data warehouse/business intelligence systems—A systematic literature review," *Int. J. Inf. Manage. Data Insights*, vol. 2, no. 2, Nov. 2022, Art. no. 100131.

[24] D. Moody and M. Kortink, "From ER models to dimensional models, part II: Advanced design issues," *Business Intell. J.*, vol. 8, pp. 20–29, Sep. 2003.

[25] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling.*, 3rd ed., Toronto NJ, ON, Canada: Wiley, 2013.

[26] W. Inmon, *Building the Operational Data Store*, 2nd ed., Hoboken, NJ, USA: Wiley, 1999.

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

[28] Y. Rogers, H. Sharp, and J. Preece, *Interact. Design: Beyond Human–Computer Interaction*. Hoboken, NJ, USA: Wiley, 2015.

[29] P. F. Vieira and L. Gerncina, "A organizao do conhecimento em ambientes digitais: Aplicao da teoria da classificao facetada," *Perspectivas em Cincia da Informa*, vol. 17, no. 4, pp. 18–40, Oct. 2012.

[30] V. Broughton, "The need for a faceted classification as the basis of all methods of information retrieval," *Aslib Proc., New Inf. Perspect.*, vol. 58, no. 1, pp. 49–72, Jan. 2006.

[31] R. Popper, "Foresight methodology," in *The Handbook of Technology Foresight: Concepts and Practice*, 1st ed., Cheltenham, U.K.: Edward Elgar, 2008, pp. 44–88.

[32] K. Halicka, "Innovative classification of methods of the future-oriented technology analysis," *Technol. Econ. Develop. Economy*, vol. 22, no. 4, pp. 574–597, Jun. 2016, doi: 10.3846/20294913.2016.1197164.

[33] K. M. Coelho e S. Borschiver, "'Technological roadmap of levulinic acid produced from lignocellulosic biomass', in portuguese: Roadmap tecnológico do ácido levulínico produzido a partir de biomassa lignocelósica," *Proc. Cadernos de Prospecção*, vol. 9, no. 4, p. 481, Dec. 2016.

[34] S. Mishra, S. G. Deshmukh, and P. Vrat, "Matching of technological forecasting technique to a technology," in *Proc. Technol. Forecasting Social Change*, 2002, vol. 69, no. 1, pp. 1–27, doi: 10.1016/s0040-1625(01)00123-8.

GISELLE F. ROSA received the bachelor's degree in computer engineering from the Military Engineering Institute, in 2008, the degree a specialization in innovation management from Linköping University, Sweden, in 2014, and the master's degree in systems and computing from the Military Engineering Institute, in 2024. She is currently an Adjunct Professor with the Technological Information Section, Brazilian Army's Technological Innovation and Management Agency, working mainly on the following topics: innovation management and technological foresight.

JONES O. AVELINO received the Tecg. degree in data processing from Universidade, Brazil, in 2001, and the M.Sc. degree in systems and computing from Instituto Militar de Engenharia (IME), in 2021, where he is currently pursuing the D.Sc. degree in defense engineering, focusing on artificial intelligence, in particular, large language models and natural language processing, to represent knowledge in the military domain. His research interests include knowledge graphs and conceptual data modeling.

MARIA CLAUDIA CAVALCANTI received the D.Sc. degree in computer engineering from the Federal University of Rio de Janeiro, Brazil, in 2003. She is currently a Full Professor with the Computer Engineering Department, Instituto Militar de Engenharia (IME), Brazil. Previously, she was a System Analyst with the Federal University of Rio de Janeiro (UFRJ), from 1985 to 2004. After receiving her D.Sc. degree, since 2004, she has been with IME, a Professor and a Researcher involved with the Postgraduate Program on Systems and Computing (PGSC) and the Postgraduate Program on Defense Engineering (PGED). Also, she has been participating in research projects on those topics, with funding from several Brazilian government agencies, including CNPq, CAPES, FINEP, and FAPERJ. Her current research interests include metadata and ontologies for data interlinking, conceptual modeling, semantic web (of data), data modeling for NoSQL DBMS, more recently, cybersecurity, and command and control systems (C2).

JULIO CESAR DUARTE received the degree from Instituto Militar de Engenharia (IME), in 1998, and the master's degree in computer science and the Ph.D. degree from the Pontifical Catholic University of Rio de Janeiro (PUC-Rio), in 2003 and 2009, respectively. He complemented his education by completing a Postdoctoral Internship with PUC-Rio, in 2021. He is currently a Professor with the Postgraduate Program in Systems and Computing and the Pro-Rector of Teaching and Research, IME. He is a professional with academic training and experience in computer engineering. His academic and professional performance is marked by a multidisciplinary approach, with a significant emphasis on developing computer systems. His experience spans several areas of computing, including artificial intelligence, machine learning, deep learning, and Portuguese natural language processing. In addition, he has been conducting research on multimodal media processing, large-scale language models, and malware analysis.