IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Analyzing Natural Language Processing Techniques to Extract Meaningful Information on Skills Acquisition from Textual Content

LUIS JOSE GONZALEZ-GOMEZ[1], SOFIA MARGARITA HERNANDEZ-MUNOZ[2], ABIEL BORJA[2], JOSE DANIEL AZOFEIFA[1], JULIETA NOGUEZ[2], and Patricia Caratozzolo[1,2]

[1]Institute for the Future of Education, Tecnologico de Monterrey, Monterrey, Mexico (e-mail: ljgonzal@tec.mx, jd.azofeifa@tec.mx, pcaratozzolo@tec.mx)

[2]School of Engineering and Sciences, Tecnologico de Monterrey, Mexico City, Mexico (e-mail: A01655084@tec.mx, A01654937@tec.mx, jnoguez@tec.mx)

Corresponding author: Patricia Caratozzolo (e-mail: pcaratozzolo@tec.mx).

The authors would like to acknowledge the financial support of Writing Lab, and the Challenge-Based Research Funding Program, Grant no. I030-IFE002-C2-T1-E, both of the Institute for the Future of Education, Tecnologico de Monterrey, Mexico, in the production of this work. We also thank the Vicerrectoría de Investigación y Posgrado, the Research Group of Advanced Artificial Intelligence, and the Cyber Learning and Data Science Laboratory of Tecnologico de Monterrey.

ABSTRACT Natural Language Processing (NLP) combines linguistics, computer science, and AI to enable computers to understand and interpret human language, making it crucial for analyzing large amounts of language data. This technology -paired with predictive models- has significant potential to forecast the relevance and evolution of skills needed in the industry, enhancing skills acquisition and alignment with job market demands and playing a key role in workforce development and educational planning. This paper comprehensively analyzes skills acquisition using NLP and predictive models. This analysis highlights significant advancements in NLP, showcasing its transformational impact on extracting and interpreting data from textual content. We conducted an extensive literature search under the systematic review guidelines, from which we selected the most relevant works for this analysis. This work examined how NLP techniques are used and adapted to extract meaningful insights from the textual content and identified which NLP models are employed to create taxonomies or classifications of skills. It explored how these models predict behaviors or outcomes in specific areas. The obtained findings show that NLP has constantly evolved in recent years, encompassing techniques that reinforce textual information extraction and underline the adaptability of NLP to various disciplines and contexts. Creating taxonomies and structured knowledge organization was a significant focus, highlighting its applicability in multiple fields. Finally, we discuss the ongoing evolution and adaptability of the NLP models to different disciplines and their integration with predictive models, which offer valuable insights that enrich the interactions between textual data and AI.

INDEX TERMS Artificial Intelligence, Engineering Education, Future Skills, Natural Language Processing, Predictive Models.

I. INTRODUCTION

NATURAL language processing (NLP) is a dynamic interdisciplinary field that integrates linguistics, computer science, and artificial intelligence to facilitate interactions between computers and humans through natural language. NLP aims to develop algorithms and models that can efficiently process and analyze large volumes of natural language data, enabling computers to understand and interpret the content of various documents, including their contextual subtleties [1]. Currently, NLP technology can extract valuable information and insights from documents in multiple languages, with algorithms developed to classify and organize documents, improving their accessibility and usability.

A significant aspect of the evolution of NLP is its application to text analysis and classification using machine learning techniques for clustering. Text analysis allows extracting meaningful information from textual data, including sentiment analysis processes, topic extraction, entity recognition, and document classification [2]. NLP techniques employ machine learning algorithms to process and understand texts on a large scale. Clustering algorithms provide a significant advantage in grouping similar documents or text fragments based on their content, structure, or context.

As the labor market evolves, there is an increasing need to identify the future needs of professional skills and competencies [3]. This presents an opportunity to adapt and enhance current trends in NLP and apply them to the vast world of skills and competency classification through predictive models of future skills. The creation of Knowledge, Skills, and Abilities (KSA) taxonomies responds to the pressing need to align educational and occupational profiles with the changing demands of Industry 4.0 [4]. These taxonomies redefine how we conceptualize and prepare for occupations, incorporating a visual dimension to offer interactive frameworks for exploring future workforce dynamics [5]. Leveraging machine learning tools and diverse data sources such as ESCO, O*NET Frameworks, FutureSkills of the government of Singapore, and the World Economic Forum, dynamic taxonomies forecast and address changing labor market requirements [6]. Through these efforts, the goal is to unlock the future of the workforce, encouraging a proactive approach to skills development and education in line with industry demands. Leveraging the power of NLP text analysis, clustering, and classification, a model could be developed to feed a dynamic taxonomy of competencies that adapt to the changing landscape of occupational capabilities.

The main objective of this analysis is to provide a comprehensive summary of the current state of NLP and predictive models to define future skills. This includes evaluating various research papers, journals, and other studies in this field to understand in a better way the advances and challenges in these fields. This work offers a current perspective on the advances in these two disciplines and facilitates the identification of critical areas of opportunity for future research.

While existing studies highlight the predictive ability of NLP models across various contexts, our analysis points to the potential for developing more refined models specifically tailored to predict the relevance and evolution of skills and competencies required in the industry. As industries evolve, some competencies become obsolete while others gain importance. NLP models capable of predicting these changes from large volumes of textual data would be invaluable to educators, employers, and practitioners, opening up a new frontier in skills acquisition.

The paper is organized as follows: subsection I-A provides a brief definition of key concepts used in this paper; section II outlines the research methodology employed in designing and conducting this analysis, including the Research Questions, **RQs**; section III presents the results and analysis, followed by a discussion where a summary of findings, the identification of opportunity areas, and a discussion on the limitations of this work; finally, section IV presents the conclusions and suggests directions for future work.

A. DEFINITION OF TERMS

This subsection presents the definitions of the relevant terms related to this work, where we identify the following:

- **Future skills**: competencies that enable individuals to effectively address intricate problems in rapidly evolving contexts while demonstrating self-organizational capabilities. These qualifications are rooted in cognitive, motivational, volitional, and social resources and are imbued with core values. Furthermore, they can be acquired through a deliberate process of learning. In this context, knowledge encompasses a comprehensive array of facts, theories, and principles pertinent to a specific field of work or study; skills denote the practical capabilities essential for executing particular tasks with proficiency; and abilities to possess the sensory, physical, psychomotor, or cognitive means necessary for effectively undertaking a given study or task [7]. Notably, these Future Skills are intricately intertwined with the knowledge of emerging technologies [8].
- **Natural Language Processing (NLP)**: NLP refers to the branch of artificial intelligence, specifically artificial intelligence, that works with text processing capabilities to understand texts like humans do [9]. The NLP combines computational linguistics and rule-based modeling of human language with statistical, machine learning, and deep learning models [10].
- **Neural Network Models**: This is a computational structure inspired by the human nervous system, designed to process information by simulating interconnected units called neurons. These neurons are organized into layers, specifically an input layer, one or more hidden layers, and an output layer. The strength or intensity of connections between these neurons, known as weights, are initially set to random values. The model learns by adjusting these weights based on the difference between its predictions and known outcomes through a repeated iterative process [11]. As it undergoes training with known data examples, the network refines its weights, improving its predictive accuracy. Once trained, the neural network can predict new, previously unseen data [12].
- **POS Tagging**: Commonly referred to as grammatical tagging, it is a foundational pre-processing task in NLP. It involves assigning appropriate grammatical categories, such as verbs, nouns, adjectives, adverbs, and determiners, to a text's words and punctuation based on their inherent definition and the context in which they appear [13]. The significance of POS tagging extends to understanding text's morphological and syntactical structure, enabling more advanced NLP tasks [13].
- **Pre-trained language model**: Pre-trained language model refers to large amounts of trained textual data with machine learning models. They mainly benefit specific NLP tasks by acquiring general linguistic features of linguistics such as syntax, grammar, and semantics. Typically, these models are used in several NLP applications, such as named entity identification, text summarization, and sentiment analysis [14]. However, there are several areas of opportunity where pre-trained linguistic models can be used. [15].

- **Syntactic patterns**: Refer to the established arrangements of words within sentences and clauses in the English language, ensuring coherence and meaningful expression. These patterns dictate the acceptable word orders, especially when various linguistic elements, like indirect objects or prepositional phrases, are involved [16].
- **TF-IDF**: Standing for term frequency-inverse document frequency is a statistical measure used to determine the significance of a word in a document relative to a corpus. It adjusts for words that naturally appear more frequently, thus providing a more accurate representation of word importance [17]. TF-IDF scores words based on their frequency in a specific document instead of their general frequency in the entire corpus, leading to more relevant search and recommendation results [18].
- **Transformer models**: The Transformer model is a neural network that learns from context, meaning, and making relationships in sequential data like the words in this sentence. Transformer models apply an evolving set of mathematical techniques (called attention or self-attention) to detect a variety of subtle ways in which even distant data elements influence and depend on each other [19].
- **Word Embedding-Based Models**: these models learn the meaning of words by considering the context of words, and thus, they consider the semantic relations between words. To do so, they transform words into numerical vectors that can be represented in a multidimensional space. Within this space, words with similar meanings are positioned closer, and words with dissimilar meanings are positioned further apart. As a result, distances between word vectors become informative about the meaning of words [20].

II. RESEARCH METHODOLOGY

This work was conducted following systematic review guidelines outlined by Page et al. [21], Kitchenham and Charters [22], Xiao and Watson [23], and Torres-Carriét al. [24] aiming to identify work related to predictive models and NLP for skills from the last four years. By employing this kind of analysis in the research, we ensure a methodologically rigorous, robust, and transparent approach, enhancing the credibility and reproducibility of the findings.

The methodology developed for this work included the definition of **RQs** to guide the study, the search phase in well-known and reputable databases, and finally, the discussion and findings. This final step involved answering each **RQ** by contrasting different approaches and works from various authors in the databases. This process aimed to provide a comprehensive understanding of the current state of the art in the areas of interest.

The following subsections detail the objectives of the analysis, the **RQs**, and the process of searching for relevant information from formal sources to answer these questions. Subsequently, we analyze and discuss the findings, contrasting the achievements of different researchers and examining how their work fits into the research and ongoing investigation activities. Finally, we draw conclusions based on the gathered data and propose future research directions.

A. OBJECTIVE OF THE ANALYSIS

The primary objective of this analysis is to recognize and summarize the current state of the predictive models and NLP in relation to skills acquisition. This includes a comprehensive review of various research papers, journals, and other studies in the field. By extensively examining diverse literature sources, we aim to gain a holistic understanding of the advancements and challenges in predictive modeling and NLP, particularly related to the creation of skills taxonomies. This review provides a present perspective on the depth and breadth of current knowledge in the field.

Additionally, this work seeks to identify opportunities to enhance NLP techniques for predicting future skills. Such advancements could benefit workers and employers when developing strategies for aligning employees' current knowledge and skills with the industry's emerging needs.

B. RESEARCH QUESTIONS

To guide this study toward the goals of identifying relevant work related to predictive models and NLP for skills acquisition and determining future research directions, we identified the following three **RQs**:

- **RQ1.** How are NLP techniques used and adapted to extract meaningful insights from textual content?
- **RQ2.** Which and how are NLP models employed to create taxonomies or classifications of skills?
- **RQ3.** How are NLP models being applied to predicting behaviors or outcomes in specific areas?

C. INFORMATION SOURCES

To ensure access to up-to-date and credible information pertinent to the (**RQs**), we identified several databases renowned for their extensive content and esteemed reputation:

- **Web Of Science** [25]: An all-encompassing research database esteemed for its extensive coverage of scientific literature. Its design boasts a user-friendly interface complemented by advanced tools for tailored search experiences.
- **Scopus** [26]: An expansive abstract and citation database encompassing research literature from diverse disciplines, with features that facilitate refined search capabilities.

The search engines include other specialized databases, such as Science Direct, IEEE Xplore, and Springer Link. The search keyword combinations should retrieve articles related to the research topic.

D. SEARCH STRATEGY

Considering the **RQs**, we meticulously identified key phrases employed in relevant databases. These precise terms were chosen to maximize the retrieval of pertinent data related to the critical terms derived from our RQs.

TABLE 1: **Keywords used in queries**

Database	Keywords search combinations
Web of Science	(NLP AND model AND skill) OR (skill AND Natural Language Processing AND (predictive model OR predictive models))
Scopus	(NLP AND model AND skills) OR (skill AND Natural Language Processing AND (predictive model OR predictive models))

Table 1 displays the keyword combinations used as queries that resulted from more than one result in the searches across multiple databases. This table provides the specific combinations of terms and the logical connectors used to filter items.

E. ELIGIBILITY CRITERIA

The selection of literature for this study followed a set of eligibility criteria to ensure that the most relevant studies were included to cover the objectives of this work.

Considered articles were published between 2019 and 2023, focusing on the most recent advances in NLP and predictive models for skills acquisition. Studies were included if they directly addressed the **RQs** related to the use of NLP techniques to extract information from textual content, the development of skill taxonomies, or the application of predictive models in skills acquisition. Only articles from artificial intelligence, computer science, software engineering, and related disciplines were considered to ensure contextual relevance. The search was limited to works published in peer-reviewed journals and conferences to ensure the credibility of the research, and only works in English were included to maintain consistency. Bibliographic searches were conducted in the Web of Science and Scopus databases, selected for their broad coverage of quality academic articles. Specific keyword combinations, as mentioned in Table 1, were used to capture the most relevant studies for this work.

Articles published before 2019 were excluded to maintain the focus on recent research. Sources not present in Web of Science or Scopus were excluded to ensure the scientific rigor of the findings. We omitted studies unrelated to the core disciplines of artificial intelligence, computer science, or software engineering and articles focused on unrelated topics such as theoretical linguistics. Duplicate publications identified in the databases were removed to avoid redundancy. Articles that did not specifically address the use of NLP in skills acquisition or the development of predictive models for skills-related taxonomies were excluded, along with studies lacking a clear methodology or relevant results. Non-English articles were excluded to ensure consistency and accuracy.

In the first phase of identification, 118 records were screened from Scopus (n = 16) and WoS (n = 102) using the keyword combinations defined in Table 1. During this process, 38 articles published before 2019 were excluded from the initial screening criteria. Additionally, 26 duplicate entries were identified and removed.

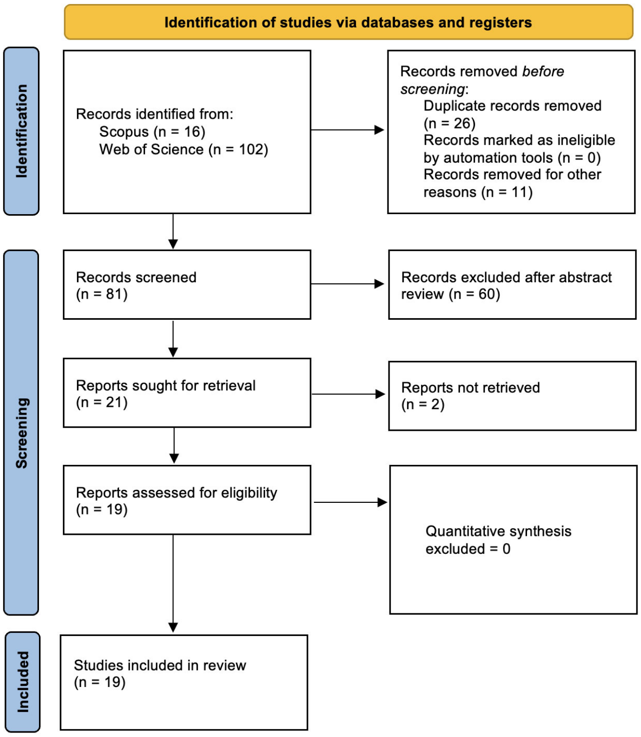

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Scopus (n = 16)
Web of Science (n = 102)
Records removed *before screening*:
Duplicate records removed (n = 26)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 11)
Records screened (n = 81)
Records excluded after abstract review (n = 60)
Reports sought for retrieval (n = 21)
Reports not retrieved (n = 2)
Reports assessed for eligibility (n = 19)
Quantitative synthesis excluded = 0
Studies included in review (n = 19)

FIGURE 1: **PRISMA flow diagram**

To conduct an assessment of the included studies, we employed eligibility criteria considering the key aspects directly addressing the **RQs** (see Figure 1). We filtered the initial set of 118 documents described in Table 2, resulting in 81 articles for further analysis.

TABLE 2: **Included and Excluded Studies Based on Evaluation Criteria**

Exclude reason	Number of articles excluded
Not related categories	4
Not related subject area	7
Duplicates	26
Total number of articles excluded	**37**
Record after excluded	**81**

The 81 articles were randomly split among the researchers, as shown in Table 3. These articles were screened by reading the title and the abstract, looking for information valuable to the **RQs**. As a result of the reading screening, 60 articles were excluded, leaving 21 articles. The exclusion criteria were based on factors such as the programming language employed, predictive models, NLP techniques, and the application area. In addition, the related category criteria focused primarily on artificial intelligence, computer science, and software engineering computer science.

TABLE 3: **Number of articles assigned by researcher**

Researcher 1	Researcher 2
n = 39	n = 42
Total number of articles excluded	**60**
Record after excluded	**21**

Subsequently, this screening phase was centered on aligning with the primary focus of the study, which revolves around applying NLP techniques and predictive models to skills. Articles were omitted if they failed to provide insights into NLP models for information retrieval from documents or if they did not offer methods for predicting skills.

For the eligibility round, as seen in Table 4, three investigators were assigned seven articles to analyze, and each analyzed article was reviewed in detail to ensure that the articles answered the **RQs**, with a total of 21 articles analyzed in the eligibility round. Finally, after careful review, we decided to exclude 2 of them, resulting in a final collection of 19 eligible articles that matched the research criteria.

TABLE 4: **Number of articles assigned by researcher**

Researcher 1	Researcher 2	Researcher 3
n = 7	n = 7	n = 7
	Total number of articles excluded	2
	Record after excluded	19

By applying these detailed eligibility criteria, we aimed to ensure a comprehensive and focused analysis of the most relevant literature in NLP and predictive models for skills acquisition.

F. QUALITY ASSESSMENT

After acquiring the 19 selected articles, three distinct parameters were discerned, aligning with specific research inquiries. The eligibility criteria comprised three subcategories, delineated by the extent to which the articles satisfied these predetermined conditions.

QA1. The study proposes using an NLP technique to extract meaningful insights from textual content.

1.1 The study does not identify or propose any specific NLP technique, or it does, but without giving enough details to understand how the method was used to extract the insights.

1.2 The study presents an NLP technique for extracting information from textual corpora. The method is explained, but the insights it yields are not specified.

1.3 The study describes a specific NLP technique and presents how it's being used to extract information and meaningful insights from text corpora.

QA2. The article discusses which and how NLP models are employed to create taxonomies or classifications of skills.

2.1 The article does not mention using NLP to build taxonomies or to organize or classify skills.

2.2 The study identifies one or more NLP models for building taxonomies or organized data but fails to relate these classifications with job skills.

2.3 The study presents an NLP model that outputs some classification or hierarchization of related data. The data has been extracted from a textual corpus and converted to some word embeddings for further analysis and for finding similarities.

QA3. The study proposes using NLP models to make predictions or outcomes in a specific area.

3.1 The study does not mention using an NLP model to forecast future values based on the training data.

3.2 The study proposes using an NLP model to foresee changes in the semantic or syntactic representation and relations between words in a corpus.

3.3 The study uses an NLP model to analyze different relations of the vocabulary within a corpus, intending to find changes in those representations that could envision some pattern or tendency.

G. DATA EXTRACTION

During the data extraction phase, the primary focus was to gauge the depth with which the selected papers addressed each of the three proposed **RQs**. Each paper was assigned a subjective percentage, determined by the researcher, indicating the strength of its association with each **RQ**, based on the paper's approach to the subject matter.

The results of this evaluation are visualized in Figure 2, which presents the percentage of each subgroup in the quality assessment.

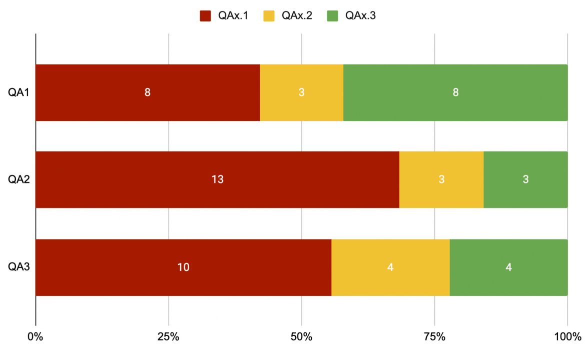


FIGURE 2: **Percentage of each subgroup in the quality assessment.**

In Figure 2, the color red represents papers that relate to a given **RQ** at a level of 33% or less. Yellow indicates a relationship ranging from 34% to 66%. Lastly, the green bars signify papers with a strong connection, relating more than 66%

From the previous chart, we can see that **RQ1** was more discussed in the papers, followed by **RQ3** and finally **RQ2**. Interestingly, the limited attention given to **RQ3** might indicate that applying NLP models in predictive analytics for specific areas remains an emergent field, and perhaps the complexities of prediction, especially in dynamic scenarios, challenge the existing NLP methodologies.

This distribution also unveils potential directions for future research endeavors. The predominant red bars for **RQ2** in Figure 2 suggest a vast space for novel studies, aiming to delve deeper into applying NLP models for crafting skill taxonomies or classifications. Yellow bars signal an intermediary exploration level, emphasizing a need to investigate further and refine NLP techniques' adaptation to extract meaningful insights from diverse textual data. On the other hand, the green bars, particularly in the context of **RQ3**, hint at a burgeoning space. There's an opportunity for more comprehensive research focusing on the nuanced applications of NLP models in predicting behaviors or outcomes in specific sectors, underlining the ever-evolving nature of NLP in contemporary analytics.

III. RESULTS AND DISCUSSION

The evaluation of the 19 selected articles was conducted meticulously in alignment with the three **RQs**. We crafted a detailed table to facilitate a comprehensive understanding and comparative analysis (Figure 2). This tabular representation is a valuable tool for identifying common themes, detecting patterns, and highlighting potential avenues for future research. It effectively organizes the diverse contributions across different research teams and individual studies.

An overview of the findings, as outlined in Table 5, is provided below, presenting an analytical commentary on their relevance to the **RQs**. Although several papers addressed themes relevant to all three **RQs**, certain studies resonated more profoundly with specific queries. The subsequent discussions delve into the relationship of each paper with the research objectives.

The compilation and assessment of these papers were meticulously guided by their alignment with and contributions to the **RQs**. Within this segment, we unpack the varied strategies and outcomes of the reviewed papers, underscoring their significance and limitations in the broader landscape of NLP research.

Based on the information in the Table 5, we can perform a data analysis that will allow us to evaluate the types of data in the papers we reviewed. This analytical exercise is important so we can emphasize the characteristics of the information collected and identify the possible patterns, trends, and aspects that may influence the study context.

In Figure 3, we can identify the percentage relative to the scope of application of the papers analyzed. This visual evaluation exercise effectively identifies and understands the proportional distribution of the different domains or thematic areas addressed in the research.

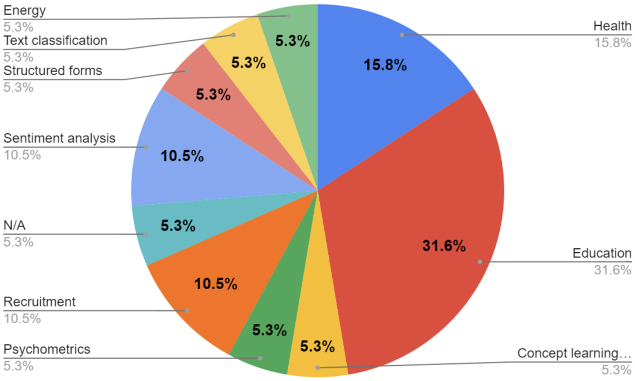


FIGURE 3: **Percentage of each Application Area.**

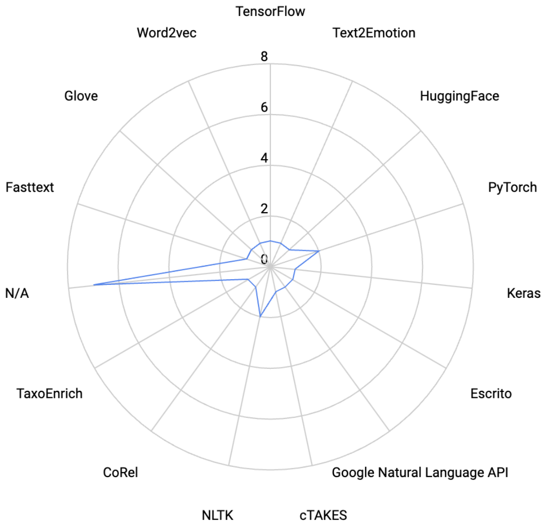


FIGURE 4: **Library or framework used on the reviewed works.**

Figure 4 shows a count of the libraries or frameworks used in the papers analyzed. This analysis allows us to quantitatively visualize the prevalence of technological tools in all analyzed studies. The visual representation of the data facilitates the identification of the most recurrent libraries or frameworks, highlighting those that have been most frequently adopted.

In Figure 5, we can see the various NLP techniques in the analyzed papers, giving an insight into the preferences and frequency of use of the most recurrent techniques within the field of study. This visual approach facilitates the identification of relevant trends and patterns.

Figure 6 shows the visualization of the predictive models mostly used in the analyzed papers. This visualization allows us to understand the trend of functional predictive models,

TABLE 5: **Summary of the article's features.**

#	Reference	Programming language / package used	Library or framework used	NLP Technique	Predictive model used	Application Area
1	(Afshar et al., 2019) [27]	Python	cTAKES	Embeddings	SciBERT, Transformer, BERT	Health
2	(Blandin et al., 2020) [28]	Python	TensorFlow	TF-IDF	Logistic regression model	Education
3	(Huang et al., 2020) [29]	N/A	CoRel	Sentiment analysis	N/A	Concept learning & relation transferring
4	(Dehbozorgi & Mohandoss, 2021) [30]	Python	Text2Emotion	Neural network models	N/A	Education
5	(Laverghetta et al., 2021) [31]	N/A	N/A	N/A	BERT	Health
6	(Saeed et al., 2021) [32]	Python	Keras	Transformer-based, non-transformer-based	BERT	Education
7	(Laverghetta et al., 2022) [33]	R	HuggingFace	Tokenization, Sentiment Analysis, POS Tagging	K-Nearest Neighbor	Psychometrics
8	(Pansare et al., 2022) [34]	N/A	N/A	Tokenization, stemming	N/A	Recruitment
9	(Deepak et al., 2022) [35]	N/A	N/A	CountVectorizer, TF-IDF transformation	N/A	Education
10	(Bamman, et al., 2023) [36]	Python	PyTorch	Stemming, Summarization, and Polarity Analysis	N/A	Health

#	Reference	Programming language / package used	Library or framework used	NLP Technique	Predictive model used	Application Area
11	(Delaforge et al., 2022) [37]	N/A	N/A	Binary text classification	N/A	N/A
12	(Cao et al., 2022) [38]	N/A	N/A	Embeddings	BERT	Sentiment analysis
13	(Elteto et al., 2022) [39]	N/A	N/A	Pretrained language models	BERT	Education
14	(Jiang et al., 2022) [40]	N/A	TaxoEnrich	N/A	BERT	Structured forms
15	(Lin et al., 2023) [41]	Python	PyTorch	Transformer-based language models	N/A	Recruitment
16	(Chi et al., 2023) [42]	N/A	N/A	Tokenization, POS Tagging, Parsing	BERT	Text classification
17	(Gombert et al., 2023) [43]	Java	Escrito	Clustering, Logistic regression, Neural network	N/A	Energy
18	(Admeur et al., 2023) [44]	N/A	Google Natural Language API	N/A	Bayesian sequence	Education
19	(Vianna et al., 2023) [45]	Python	NLTK	Logistic regression, Neural network models	N/A	Sentiment analysis

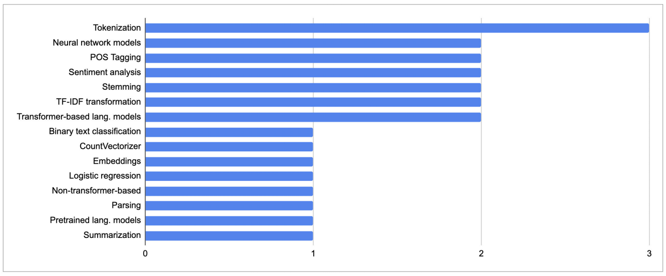


FIGURE 5: **NLP techniques used on the reviewed works.**

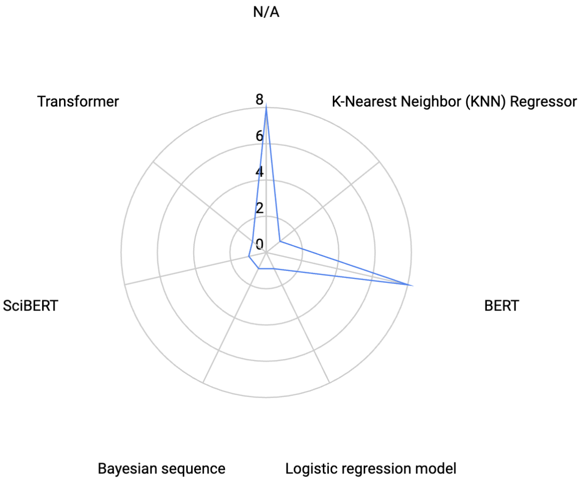


FIGURE 6: **Predictive model used on the reviewed works.**

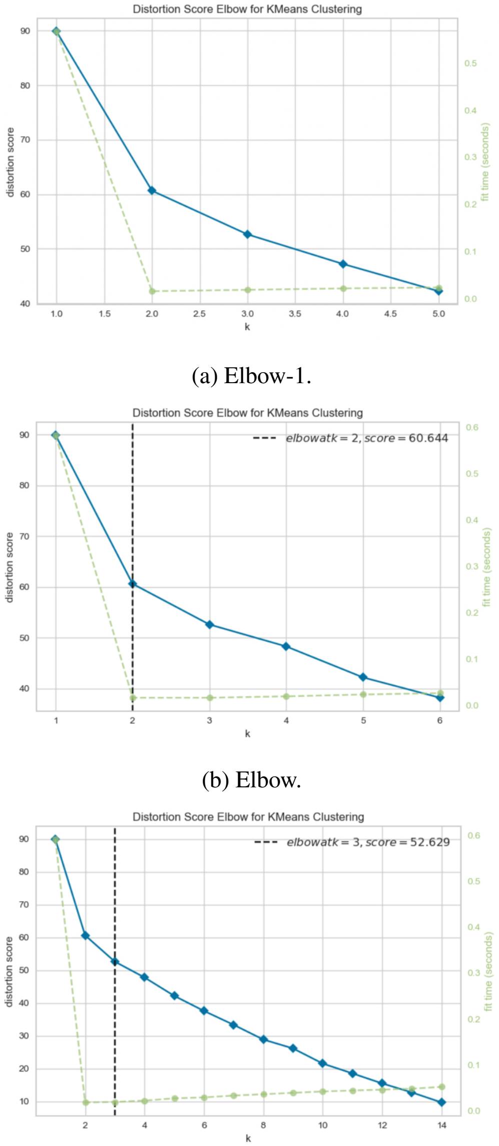
offering a strategic vision for those interested in the field of study.

Along with the discoveries of the individual analysis of each data obtained, a clustering of all these data together was carried out (Table 5) based on the work of Azofeifa et al [46], which seeks to glimpse in a defined space how related or similar they are the works from a general point of view of the whole. Remember that we have the following data: Programming language/package used, Library or framework used, NLP Technique, Predictive model used, Application Area, and whether it is a systematic literature review. Considering that each job can use zero, one or more of the options each of these selections has available. To compare the items, we define a metric that assigns each item a value depending on the options used. In particular, we assign a number 1 if the article uses the characteristic or a 0; otherwise, establishing the value of the paper based on whether or not it possesses each of all the available characteristics and the distance between documents in the n-dimensional space will be calculated. using the L2 Euclidean norm [47].

We can perform a cluster analysis in an n-dimensional space by having a defined distance between items. First, we define the optimal number of groups by applying the elbow method, as shown in Figure 7.

Subsequently, by applying the k-means algorithm, we group the data into a predefined number of groups. The number of clusters is at the “elbow” of the graph, which shows a trade-off between the number of items in each cluster (the higher, the better) and the compactness of each cluster (the more compact, the better). Thanks to the application of this method, several clusters k = 2 were found. Figure 7 also shows the choice of clusters for k = 1 and k = 3, which shows that k + 2 is a good choice. so we proceeded to make a view of the subdivision of the group as a dendrogram in Figure 8.

The k-means results in group items according to their distance; it is not apparent what each group includes and why, where initially in space is five dimensions, but using the T-distributed Stochastic Neighbor Embedding (T-SNE) algorithm can be projected in 2D, Figure 9, where the algorithm



(a) Elbow-1.

(b) Elbow.

(c) Elbow+1.

FIGURE 7: **Elbow method applied. (a, b, c).**

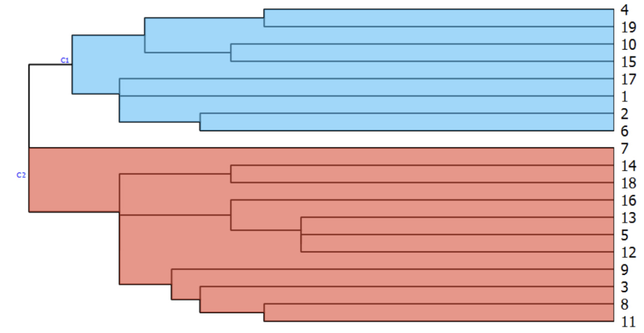


FIGURE 8: **Dendrogram of the cluster subdivision.**

seeks to maintain nearby points in the higher dimensional space also in a 2D space. The main parameters of the T-SNE algorithm are the following: perplexity = 100, exaggeration = 1, and 6 components PCA, and the results obtained from this are shown in Figure 9, where two groups are identified, and each point with its respective number corresponds to an article and its reference for Table 5.

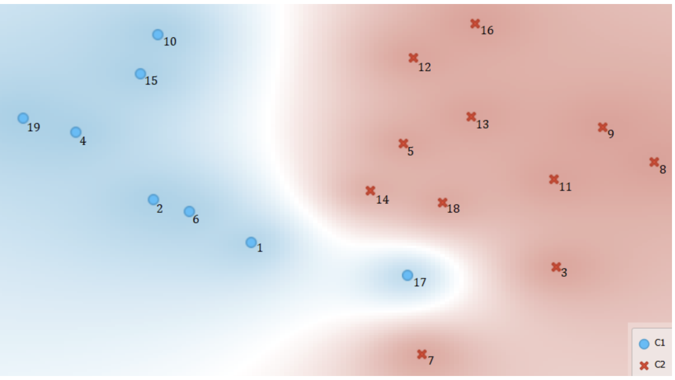


FIGURE 9: **T-distributed Stochastic Neighbor Embedding (T-SNE) has been used for visualization of the relationship among papers. The image shows two clusters identified by the dendrogram and their correlation distribution.**

The preceding charts and figures illustrate the multiple approaches adopted in NLP to overcome the challenge of deciphering unstructured text and extracting valuable information from a spectrum of distinct sources. Each visual representation encapsulates a unique facet of NLP methodologies, reflecting the diversity researchers employ in tackling the complexities inherent in textual data.

Below, we revisit the initial **RQs** that set the groundwork for this comprehensive exploration. With a discerning lens, we embark on a journey to address and unravel these questions, leveraging the rich tapestry of insights gleaned from the various contributions of selected researchers. The objective is to not only contrast the diverse methodologies employed but also analyze the approaches that each researcher brings to the forefront.

By comparing these perspectives, we aim to comprehensively understand the multifaceted landscape of NLP applications in extracting meaning and valuable information from unstructured textual content to construct taxonomies or classifications. Through the analysis, we seek to unveil patterns, trends, and novel insights that collectively contribute to the evolving narrative of NLP research. Now, we will answer and discuss each of the **RQs** thanks to the findings and analysis carried out in this work.

RQ1.- How are NLP techniques used and adapted to extract meaningful insights from textual content?

The ever-evolving field of NLP has been pivotal in transforming how we extract and interpret data from textual content. This section explores the diverse methodologies and perspectives that have sought to answer this central query, illuminating NLP's expansive reach and potential in extracting and interpreting information from textual data.

Afshar et al [27] employs a robust architecture to convert clinical documents into standardized vocabularies. This translation from diverse clinical narratives into a unified lexicon illustrates the power of NLP in the medical realm, showcasing its ability to standardize and unify complex and varied textual data. The study compared CUIs and n-grams as inputs to a regularized logistic regression classifier using tf-idf transformation, with hyperparameters (L1 vs L2 regularization and coefficient C) optimized through grid search for the highest AUC ROC, while evaluating recall, specificity, NPV, and PPV using Python and RStudio.

Diving into deeper semantic understanding, Blandin et al. [28] employ machine learning classifiers to discern text appropriateness, ensuring age-aligned content. This method highlights the capacity of NLP to evaluate and categorize content based on semantic appropriateness, which is crucial for tailoring information to specific audiences. The study involves three models: a standard regression model with a first layer of 606 dimensions, a multi-task model with age prediction and binary classification (adult vs. children), and a sequence model combining classification and regression for children's texts, with hyperparameters (hidden layers, activation function, dropout) shared across models. Similarly, Huang et al. [29] explore seed-guided topical taxonomy construction with their CoRel framework. This framework crafts intricate knowledge structures from text, reinforcing that NLP techniques are evolving beyond simple text processing to understanding and creating complex relationships within data. Their method expands user-given seed taxonomies by using a relation transferring module for discovering root topics and subtopics and a concept learning module for generating topical clusters, leveraging a pre-trained BERT model to train a relation classifier on minimal seed data with data augmentation and masked token techniques to infer hierarchical relationships.

Dehbozorgi's work [30] integrates the analysis of emotions with educational outcomes, employing NLP techniques to translate speech-based emotions into predictors of academic success. Aspect-based emotion analysis is performed with POS tagging, and the KNN algorithm is applied to the combined aspects and emotions as feature vectors to predict student performance. This application underscores NLP's potential in interdisciplinary research, particularly in understanding the emotional dimensions of communication and their impact on learning.

Similarly, Cao et al. [38] enhance sentiment analysis by injecting user identity into pre-trained models, emphasizing the personalization dimension of NLP techniques. For the sentiment analysis, they use U-PLMs, incorporating user identity through embedding-based personalization in the embedding module and attention-based personalization in BERT's self-attention module, with a two-stage training process. User embeddings are added to token representations for personalized document-level bias, and BERT's transformer encoder processes the token sequence with multi-head self-attention and feed-forward networks across L layers to predict sentiment categories. Adapting pre-existing models for specific contexts exemplifies NLP's shift towards more tailored and context-aware insights.

Éltető et al. [39] highlight an intersection of Bayesian modeling and NLP. Although primarily targeting human skill acquisition, their hierarchical Bayesian sequence model demonstrates NLP methodologies' flexibility and interdisciplinary nature, with potential applications in robust text sequence analysis in NLP.

Laverghetta et al. [33] epitomize deep semantic understanding by exploring the intricate psychological layers embedded in linguistic expressions. This research delves into the nuances of language, moving beyond surface-level analysis to uncover deeper meanings and implications.

Lin et al. [2] focus on closing the gap between structured and unstructured data. Their work demonstrates a push to harmonize structured representations, such as taxonomies and graphs, with the intricacies of unstructured textual data. Their approach combines a seed-guided method and relation transferring to adapt existing taxonomies and employs advanced semantic understanding to derive structured graphs from free-form text.

Collectively, these articles address the **RQ1**, showcasing the vast and dynamic reach of NLP. From fundamental sentiment analysis to intricate taxonomy construction, the literature paints a picture of an ever-evolving NLP landscape. Techniques range from pre-trained models to Bayesian analytics, machine learning classifiers, and semantic parsing. These methods not only bolster the extraction of textual insights but also underscore NLP's adaptability across diverse disciplines and contexts.

RQ2.- Which and how are NLP models employed to create taxonomies or classifications of skills?

In exploring the literature on NLP models for establishing taxonomies or categorizations of competencies, we identified several significant contributions aligned with the research focus. These contributions provide valuable insights into the intersection of NLP and skills categorization, enhancing our understanding of the topic. Below, we examine the most salient contributions and discuss how their methods and findings contribute to our overall understanding of the topic.

Huang et al. [29] presents an innovative approach to creating skills taxonomies. Their method focuses on constructing thematic taxonomies guided by conceptual seeds. This approach is notable for its ability to expand the taxonomy structure in amplitude and depth. Including conceptual names as seeds enriches the corpus and lays the foundation for building a robust taxonomy. Furthermore, the transfer of relations along multiple paths, driven by a relation classifier, not only reveals root nodes and uncovers new topics and subtopics but also provides a hierarchical structure essential for skill classification. Their innovative techniques and approach offer valuable insights that effectively contribute to the construction of taxonomies.

Jiang et al [40] provides an original perspective on organizing knowledge into structured forms, highlighting its applicability across various domains. Their work, particularly the TaxoEnrich framework, addresses a vital issue in an automated manner to incorporate new concepts without starting from scratch. The framework consists of four modules: taxonomy-contextualized embedding generation, sequential feature encoder, query-aware sibling encoder, and query-position matching. Taxonomy-contextualized embeddings are generated using pseudo sentences derived from taxonomic relationships, with separate handling of ancestor and descendant paths. These embeddings feed into a sequential feature encoder that models structural taxonomy information in a vertical view. The query-aware sibling encoder captures horizontal structural information by selecting relevant siblings based on query-relatedness, and the final query-position matching model computes relatedness using both fine- and coarse-grained relationships between the query node and the candidate parent, child, and siblings for accurate taxonomy completion. TaxoEnrich has demonstrated efficacy by significantly outperforming several existing methods, showcasing its potential applicability in skill categorization. This framework emphasizes the importance of evolving and enhancing existing taxonomies directly relevant to our goals.

Lin et al. [2] significantly contribute to our understanding of applying NLP models in constructing skill taxonomies, specifically within the context of the LinkedIn talent marketplace. Their completion task is a supervised classification problem using BERT-based embeddings and LinkedIn's proprietary data to predict relationships between skills by leveraging synthetic training sentences and a carefully designed labeling process. Their innovative approach to building skill graphs highlights the practical application of NLP models. Lin et al. underscore the importance of quality text corpus, demonstrating the effectiveness of NLP techniques and providing solutions to common problems, such as irrelevant information in real-world text corpora. This study is crucial for understanding how NLP can be effectively applied in categorizing skills in professional environments, offering practical insights into the construction and refinement of skill taxonomies.

RQ3.- How are NLP models applied to predicting behaviors or outcomes in specific areas?

NLP models are increasingly employed to predict behaviors and outcomes across various domains, demonstrating their versatility and effectiveness.

In the clinical domain, Afshar et al. [27] utilize NLP to process clinical documents by extracting unique concept identifiers for large-scale clinical research. By applying machine learning techniques, this study highlights the effectiveness of NLP in analyzing unstructured clinical data, offering potential applications in healthcare research and data interoperability.

NLP models also play a crucial role in knowledge structuring and taxonomy expansion. Huang et al. [29] developed a seed-guided topical taxonomy construction method, while Jiang et al. [40] introduced the TaxoEnrich framework. These models automate the incorporation of new concepts into existing taxonomies by leveraging semantic and structural information, enhancing knowledge organization with significant implications for domains such as e-commerce and named entity recognition.

NLP's application in psychometrics and language assessment is underscored by Laverghetta et al. [33]; they employ transformer-based language models to predict linguistic competencies and psychometric properties, highlighting the capabilities and limitations of current models. The potential for NLP to predict human performance on reasoning tasks introduces efficiencies in psychometrics. Similarly, Pansare et al. [34] combine the Myers-Briggs Type Indicator with machine learning algorithms to predict personality traits based on linguistic patterns. This approach aids in selecting candidates with suitable personalities for specific roles, showcasing the applicability of NLP in job profiling.

In educational contexts, Blandin et al. [28] propose using innovative machine-learning approaches to determine the age-appropriateness of textual content, offering scalable and objective solutions for educational platforms and libraries. Additionally, Dehbozorgi & Mohandoss [30] explore Aspect-Based Emotion Analysis using NLP to correlate students' sentiments with academic performance. This predictive capability enhances collaborative learning environments by providing insights into student emotions and their impact on educational outcomes.

In the field of talent management, Lin et al. [41] develop approaches to constructing skill graphs for platforms like LinkedIn. These approaches leverage NLP models to predict relationships between skills, facilitating the efficient matching of job descriptions with user profiles. This contributes to personalized career advancement recommendations and better-aligning skills with job market demands.

Applying NLP models to predict behaviors and outcomes spans diverse fields such as education, talent management, psychometrics, personality assessment, knowledge structuring, and healthcare. As NLP models evolve and adapt, their interdisciplinary impact on predictive modeling becomes increasingly apparent, providing valuable insights and efficiencies across various domains.

Among the authors who addressed **RQ1**, there was a clear trend towards the diverse and innovative use of NLP techniques to extract meaningful insights from textual content. Afshar et al. [27] demonstrated NLP's adaptability in the medical sector by standardizing clinical narratives. Dehbozorgi & Mohandoss [30] highlighted the intersection of NLP with emotion analysis and educational prediction. A prominent trend is the shift from fundamental text analysis to deeper semantic understanding, as seen in Cao et al. [38], who infused sentiment analysis with user identity, and Blandin et al. [28] who focused on discerning age-appropriate content. Huang et al. [29] and Lin et al. [41] emphasized harmonizing structured taxonomies with raw textual data, bridging the gap between structured and unstructured data. Meanwhile, Éltető et al. [39] illustrated the field's interdisciplinary nature by combining Bayesian modeling with NLP, enhancing analytical capacities. Laverghetta et al. [31], [33] delve into the semantic intricacies and psychological layers within linguistic expressions. Collectively, these authors highlight the expansive scope, adaptability, and evolving nature of NLP across various disciplines and contexts.

Several key insights emerged in exploring how NLP models are employed to create taxonomies or classifications of skills, addressing **RQ2**. Lin et al. [41] highlighted the significance of a robust text corpus in constructing skill graphs for platforms such as LinkedIn. Their work not only illustrated the complexities of developing such a system but also addressed the challenges of filtering out irrelevant information in real-world datasets. Meanwhile, Huang et al. [29] introduced an innovative technique for crafting seed-guided thematic taxonomies, utilizing conceptual names as foundational seeds to enrich the taxonomy's content and deploying relation classifiers to generate hierarchical structures, thereby streamlining the taxonomic framework. Additionally, Jiang et al.'s [40] TaxoEnrich emerged as a promising solution for expanding existing taxonomies, automating the incorporation of new concepts, and demonstrating a level of efficiency superior to many contemporary methods.

In addressing **RQ3** on the application of NLP models for predicting behaviors or outcomes, the study found that these models are versatile across several domains. Blandin et al. [28] harnessed machine learning to assess the age-appropriateness of textual content in educational platforms, while Dehbozorgi & Mohandoss [30] employed Aspect-Based Emotion Analysis to correlate student sentiments with academic outcomes. In talent management, skill graphs for platforms like LinkedIn predict skill interrelations, streamlining the matching of job profiles with user profiles. Laverghetta et al. [31], [33] utilized transformer-based models to predict linguistic capabilities in psychometrics, and Pansare et al. [34] combined the Myers-Briggs Type Indicator with machine learning to foresee personality traits from language patterns, optimizing job role fitment. Additionally, Huang et al. [29] and Jiang et al. [40] demonstrated the evolution of knowledge taxonomies through NLP. In contrast, Afshar et al. [27] showcased NLP's prowess in extracting insights from clinical data. These findings underscore NLP's potential in diverse predictive modeling applications, ranging from education to healthcare.

The research areas identified in the revised studies highlight a distinct opportunity for refining NLP methodologies explicitly tailored for skill extraction and categorization. For instance, the contributions of Lin et al. [41], and Huang et al. [29] provide foundational steps; however, there remains ample room for continued innovation in constructing dynamic skill graphs. This underscores the need for further exploration in automatic skill extraction and categorization from diverse textual sources, as well as how NLP can be leveraged to identify emerging or niche skills that may not yet be categorized in established taxonomies.

While the studies under review emphasize the predictive capabilities of NLP models in various contexts, there remains ample room for exploring more refined models that precisely predict skill relevancy and evolution. As industries undergo continuous transformation, specific skills become obsolete while others rise in importance. NLP models that forecast such shifts based on extensive textual data, such as Word Economic Forum reports detailing the skills needed to perform specific jobs, would offer immense value to educators, employers, and professionals.

Multilingual NLP can be crucial in defining the future skills for a global and dynamic workforce [48]. However, we found no relevant multilingual NLP studies related to skills acquisition. We attributed the fact to two main reasons: Firstly, the significant amount of available English-language materials, including academic research, industry reports, and educational content on skills acquisition, predominantly available in English. Secondly, not using a multilingual NLP approach increases accuracy and performance when dealing with a single language, typically English. By concentrating on the English language, NLP models can be fine-tuned to understand the context more effectively, leading to more accurate extraction and analysis of skills-related information.

The limitations of this study are underscored by the diverse array of approaches evident in the data extraction process, as depicted in Figure 4, which illustrates the number of libraries or frameworks utilized in the analyzed articles. The heterogeneity in the selection of technological tools introduces variability in the results and may impact the comparability between different studies. Furthermore, the visualization presented in Figure 5, detailing the NLP techniques employed, highlights the necessity to address the complexity inherent in the diverse methodological approaches. Another significant limitation relates to the domains covered by the research, as depicted in Figure 3 where application areas are identified; the omission of specific relevant disciplines may leave areas of interest unexplored. Additionally, the limitation in the number of authors examined in the review, which was narrowed to focus specifically on works relevant to this study, may impact the breadth and depth of the analysis conducted.

Another inherent limitation of this study manifests in the restricted number of papers included in the review. While this restriction was necessary to maintain a selective approach, it may hinder the representativeness of the overall findings. Nonetheless, this limitation presents an opportunity for future research endeavors. Expanding the sample to include a more significant number of papers could enhance the richness and validity of the study. Although we recognize the current restriction in terms of the number of documents analyzed, we also recognize it as a clear opportunity for the development and expansion of future research.

IV. CONCLUSION AND FUTURE WORK

This analysis presents an encompassing overview of the current landscape of NLP in skill extraction, prediction, and taxonomy construction. The findings collectively underscore the adaptability and depth of NLP across various disciplines, from clinical narratives to talent management and education. Authors have demonstrated a paradigm shift from basic text analytics to profound semantic understanding, bridging the gap between raw data and structured taxonomies. The interdisciplinary nature of NLP, its expansive scope, and its evolving capabilities stand as a testament to the field's potential to shape the future of knowledge management and information extraction.

Throughout this investigation, the potential of NLP in crafting dynamic taxonomies of skills becomes evident. However, the real novelty lies in the opportunities for further refinement and expansion of these techniques, particularly in predicting skill evolution and relevancy within a rapidly changing job market. NLP's capability to forecast emerging or niche skills from vast textual datasets hints at its future potential in shaping educational strategies, informing hiring practices, and professional development pathways.

Future research should prioritize refining NLP methodologies specifically designed for skill extraction and categorization to address the evolving needs of a dynamic workforce. For example, one key direction is the development of more robust models for Named Entity Recognition (NER) to accurately identify and extract emerging skills and occupations from diverse textual sources, including unstructured data like job descriptions, academic publications, and industry reports. This would require advancing NER models to capture contextual nuances, particularly for niche skills that may not be captured in existing taxonomies.

Building on this, another avenue of future work is constructing dynamic skill graphs that model the evolution of skills and occupations over time. This can be achieved by integrating NLP with longitudinal datasets to detect patterns in skill development, transitions across roles, and shifts in industry demands. In particular, predictive modeling techniques can be applied to forecast skill evolution, offering insights into which skills will grow in demand or become obsolete. This is crucial for answering the third **RQ**, which concerns the ability of NLP models to predict and adapt to emerging trends in skills and occupations.

In our initial work, we have begun designing an architecture incorporating both NER and relation extraction models.

The NER component focuses on identifying unique entities, such as skills and occupations, in unstructured text, while the relation extraction model infers relationships between these entities, helping to discover connections that might not be explicitly stated. Future work will build on this foundation by exploring methods for automating the discovery of new skill-occupation relationships. For instance, relation extraction models could be enhanced to identify and predict associations between emerging skills and the industries most likely to require them, contributing to the forecasting of industry trends.

Moreover, the predictive potential of these models should be evaluated using real-world datasets, such as the World Economic Forum's reports on future job skills, to ensure that the models can deliver actionable insights. This line of research would benefit educational institutions and organizations looking to stay ahead of workforce trends by anticipating skills that will be in high demand and adjusting curricula and training programs accordingly. Businesses could also leverage such models to better inform workforce planning, helping them align their recruitment strategies with emerging skill requirements.

The need for this architecture is underscored by the increasing volume of unstructured text in reports, articles, and other media, providing a rich source of information for uncovering new insights about occupations and skills. As industries evolve and new skills emerge, the ability of NLP models to predict these changes will become increasingly important, making this an exciting and necessary area of continued research.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support of Writing Lab and the Challenge-Based Research Funding Program, Grant no. I030-IFE002-C2-T1-E, both of the Institute for the Future of Education, Tecnologico de Monterrey, Mexico, in the production of this work. We also thank the Vicerrectoría de Investigación y Posgrado, the Research Group of Advanced Artificial Intelligence, and the Cyber Learning and Data Science Laboratory of Tecnologico de Monterrey.

REFERENCES

[1] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From traditional to deep learning," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 2, pp. 1–41, 2022.

[2] X. Li, X. Wu, Z. Luo, Z. Du, Z. Wang, and C. Gao, "Integration of global and local information for text classification," *Neural Computing and Applications*, vol. 35, no. 3, pp. 2471–2486, 2023.

[3] P. Caratozzolo, U. Cukierman, B. Nørgaard, K. Schrey-Niemenmaa, J. D. Azofeifa, and V. Rueda-Castro, "Future skills forecasting: Ensuring quality learning for every segment of the workforce," in *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2024, pp. 1–5.

[4] J. D. Azofeifa, V. Rueda-Castro, L. J. Gonzalez-Gomez, S. M. Gómez-Puente, J. Noguez, and P. Caratozzolo, "Unlocking the future of infocomm workforce: A visual ksa matrix taxonomy approach to education and occupational profiles," in *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2024, pp. 01–09.

[5] Shaping skills. Accessed: 2023-11-29. [Online]. Available: https://shapingskills.mx

[6] P. Caratozzolo, J. D. Azofeifa, L. A. Mejía-Manzano, V. Rueda-Castro, J. Noguez, A. J. Magana, and B. Benes, "A matrix taxonomy of knowledge, skills, and abilities (ksa) shaping 2030 labor market," in *2023 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2023, pp. 1–8.

[7] World Economic Forum, "Building a common language for skills at work: A global taxonomy." World Economic Forum, Geneva, Switzerland, 2021.

[8] U.-D. Ehlers, *Future skills: The future of learning and higher education*. BoD, Books on Demand, 2020.

[9] E. Cambria and B. White, "Jumping nlp curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[10] IBM. What is natural language processing? Accessed: 2024-08-13. [Online]. Available: https://www.ibm.com/topics/natural-language-processing

[11] Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, 2016.

[12] IBM, "Neural networks in spss modeler," accessed: 2024-08-13. [Online]. Available: https://www.ibm.com/docs/en/spss-modeler/18.0.0?topic=networks-neural-model

[13] D. Kumawat and V. Jain, "Pos tagging approaches: A comparison," *International Journal of Computer Applications*, vol. 118, no. 6, 2015.

[14] P. Marcelino, "Transfer learning from pre-trained models," *Towards data science*, vol. 10, no. 330, p. 23, 2018.

[15] C.-H. Chiang, Y.-S. Chuang, and H.-y. Lee, "Recent advances in pre-trained language models: Why do they work and how do they work," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, 2022, pp. 8–15.

[16] Grammarly, "What is syntax? definition, examples of syntax in literature," accessed: 2024-08-13. [Online]. Available: https://www.grammarly.com/blog/syntax/

[17] A. Mishra and S. Vishwakarma, "Analysis of tf-idf model and its variant for document retrieval," in *2015 international conference on computational intelligence and communication networks (cicn)*. IEEE, 2015, pp. 772–776.

[18] F. Karabiber, "Tf-idf (term frequency-inverse document frequency)," accessed: 2024-08-13. [Online]. Available: https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/

[19] R. Merritt, "What is a transformer model?" 2024, accessed: 2024-08-16. [Online]. Available: https://blogs.nvidia.com/blog/what-is-a-transformer-model/

[20] T. Widmann and M. Wich, "Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text," *Political Analysis*, vol. 31, no. 4, pp. 626–641, 2023.

[21] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *bmj*, vol. 372, 2021.

[22] B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, and S. Linkman, "Systematic literature reviews in software engineering–a tertiary study," *Information and software technology*, vol. 52, no. 8, pp. 792–805, 2010.

[23] Y. Xiao and M. Watson, "Guidance on conducting a systematic literature review," *Journal of planning education and research*, vol. 39, no. 1, pp. 93–112, 2019.

[24] P. V. Torres-Carrión, C. S. González-González, S. Aciar, and G. Rodríguez-Morales, "Methodology for systematic literature review applied to engineering and education," in *2018 IEEE Global engineering education conference (EDUCON)*. IEEE, 2018, pp. 1364–1373.

[25] Clarivate, "Web of science platform," 2024, accessed: 2024-08-16. [Online]. Available: https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/

[26] Elsevier, "Scopus: Comprehensive, multidisciplinary, trusted abstract and citation database," 2024, accessed: 2024-08-16. [Online]. Available: https://www.elsevier.com/products/scopus

[27] M. Afshar, D. Dligach, B. Sharma, X. Cai, J. Boyda, S. Birch, D. Valdez, S. Zelisko, C. Joyce, F. Modave *et al.*, "Development and application of a high throughput natural language processing architecture to convert all

clinical documents in a clinical data warehouse into standardized medical vocabularies," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1364–1369, 2019.

[28] A. Blandin, G. Lecorvé, D. Battistelli, and A. Etienne, "Age recommendation for texts," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 1431–1439.

[29] J. Huang, Y. Xie, Y. Meng, Y. Zhang, and J. Han, "Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1928–1936.

[30] N. Dehbozorgi and D. P. Mohandoss, "Aspect-based emotion analysis on speech for predicting performance in collaborative learning," in *2021 IEEE frontiers in education conference (FIE)*. IEEE, 2021, pp. 1–7.

[31] A. Laverghetta Jr, A. Nighojkar, J. Mirzakhalov, and J. Licato, "Can transformer language models predict psychometric properties?" 2021.

[32] T. Saeed, M. Sufian, M. Ali, and A. U. Rehman, "Convolutional neural network based career recommender system for pakistani engineering students," in *2021 International Conference on Innovative Computing (ICIC)*. IEEE, 2021, pp. 1–10.

[33] A. Laverghetta Jr, A. Nighojkar, J. Mirzakhalov, and J. Licato, "Predicting human psychometric properties using computational language models," in *The Annual Meeting of the Psychometric Society*. Springer, 2021, pp. 151–169.

[34] A. Pansare, P. Panwar, and P. Kosamkar, "Personality prediction with natural language processing using questionnaire responses," in *2022 IEEE Pune Section International Conference (PuneCon)*. IEEE, 2022, pp. 1–6.

[35] G. Deepak, V. Adithya, and A. Santhanavijayan, "Ontoblogdis: a knowledge-centric ontology driven socially aware framework for influential blogger discovery," in *Proceedings of Emerging Trends and Technologies on Intelligent Systems: ETTIS 2021*. Springer, 2022, pp. 37–47.

[36] D. Bamman, D. Hovy, D. Jurgens, K. Keith, B. O'Connor, and S. Volkova, "Proceedings of the fifth workshop on natural language processing and computational social science (nlp+ css)," in *Proceedings of the (NLP+ CSS)*, 2022.

[37] A. Delaforge, J. Azé, S. Bringay, C. Mollevi, A. Sallaberry, and M. Servajean, "Ebbe-text: Explaining neural networks by exploring text classification decision boundaries," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 10, pp. 4154–4171, 2022.

[38] X. Cao, J. Yu, and Y. Zhuang, "Injecting User Identity Into Pretrained Language Models for Document-Level Sentiment Classification," *IEEE ACCESS*, vol. 10, pp. 30 157–30 167, 2022.

[39] N. Elteto, D. Nemeth, K. Janacsek, and P. Dayan, "Tracking human skill learning with a hierarchical Bayesian sequence model," *PLOS Computational Biology*, vol. 18, no. 11, Nov. 2022.

[40] M. Jiang, X. Song, J. Zhang, and J. Han, "Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 925–934.

[41] S. Lin, Y. Yuan, C. Jin, and Y. Pan, "Skill graph construction from semantic understanding," *ACM Web Conference 2023 - Companion of the World Wide Web Conference 2023*, p. 978 – 982, 2023.

[42] Z. Chi, H. Huang, L. Liu, Y. Bai, X. Gao, and X.-L. Mao, "Can pretrained english language models benefit non-english nlp systems in low-resource scenarios?" *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 32, p. 1061 – 1074, 2024.

[43] S. Gombert, D. Di Mitri, O. Karademir, M. Kubsch, H. Kolbe, S. Tautz, A. Grimm, I. Bohm, K. Neumann, and H. Drachsler, "Coding energy knowledge in constructed responses with explainable nlp models," *Journal of Computer Assisted Learning*, vol. 39, no. 3, p. 767 – 786, 2023.

[44] D. Vianna, F. Carneiro, J. Carvalho, A. Plastino, and A. Paes, *Sentiment analysis in Portuguese tweets: an evaluation of diverse word representation models*. Springer, 2024, vol. 58, no. 1, pp. 223–272.

[45] ——, "Sentiment analysis in Portuguese tweets: an evaluation of diverse word representation models," *Language Resources and Evaluation*, 2023.

[46] J. D. Azofeifa, J. Noguez, S. Ruiz, J. M. Molina-Espinosa, A. J. Magana, and B. Benes, "Systematic review of multimodal human–computer interaction," vol. 9, no. 1, p. 13, 2022.

[47] M. Barni, V. Cappellini, and A. Mecocci, "Fast vector median filter based on euclidean norm approximation," *IEEE Signal Processing Letters*, vol. 1, no. 6, pp. 92–94, 1994.

[48] O. Majewska, I. Vulić, and A. Korhonen, "Linguistically guided multilingual nlp: Current approaches, challenges, and future perspectives," *Algebraic Structures in Natural Language*, pp. 163–188, 2022.

LUIS JOSE GONZALEZ-GOMEZ holds a Ph.D. in Computer Science from Tecnológico de Monterrey, where his research focused on using Natural Language Processing (NLP) and Machine Learning (ML) to create a framework for analyzing job posts to identify future skills. He also holds a Master's degree in Educational Technology from Tecnológico de Monterrey, with research on using mobile devices in education. Additionally, he is a Computer Systems Engineer from Tecnológico de Monterrey. With over 18 years of experience teaching university-level courses related to Object-Oriented Programming (OOP), web development, and UI/UX, Luis has extensive experience in his field. His research interests lie in artificial intelligence, particularly applying and adapting cutting-edge NLP techniques.

SOFIA MARGARITA HERNANDEZ-MUNOZ Graduated from a Computer Technology Engineering degree, she has participated in the Beautiful Patterns Camp. This project seeks to encourage middle and high school girls to study a STEAM career. Currently, she has one year of experience and continues working at IBM. Additionally, she has completed a research stay focused on the field of Artificial Intelligence.

ABIEL BORJA graduated in Engineering in Computer Science from Tecnológico de Monterrey, Mexico. Passionate about emerging technologies, data science, and artificial intelligence, Abiel has participated in various projects that have significantly enhanced their expertise. One of the notable projects involved developing a dynamic taxonomy using natural language processing (NLP) and fine-tuning techniques. Abiel Borja has professional experience as a data analyst, applying technical skills to solve real-world challenges and deepen their understanding of data analysis. These roles have provided a solid foundation in computer science principles and programming. Abiel is dedicated to continuous learning and professional growth, with a keen interest in contributing to the future of technology and its potential to drive innovation. Their research interests include data science, artificial intelligence, and technological advancement.

JOSE DANIEL AZOFEIFA received the Bachelor's degree in Computer Engineering from the Instituto Tecnologico de Costa Rica in 2016 and the Doctorate in Engineering Sciences from the Tecnologico de Monterrey, Mexico City Campus, in December 2021. His research focuses on virtual reality, exploring multisensory environments to enhance actor immersion in scenario management for decision-making processes. From 2022 to 2023, he served as a professor in the Computer Engineering Department at the Tecnologico de Monterrey, Mexico City Campus, demonstrating his commitment to academia and practical application. Dr. Azofeifa is a postdoctoral fellow in the ShapingSkills project at the Institute for the Future of Education at Tecnologico de Monterrey. His expertise includes 2D/3D games, virtual reality, haptics, and dynamic KSA taxonomies. He has extensive experience in development, teaching, and project management. Dr. Azofeifa is a versatile professional with a strong track record in various areas of computer engineering. His contributions to the field are reflected in his numerous projects and collaborations.

JULIETA NOGUEZ was a professor-researcher at the Computer Science Department of the Tecnológico de Monterrey, Mexico City Campus. She was co-leader of the Advanced Artificial Intelligence research group. She was responsible for the Cyber-Learning & Data Sciences Lab. She belongs to the National Research System of Mexico (SNI level II), the IEEE Computer Society, the IEEE Education Society, the Mexican Society of Artificial Intelligence, and the Mexican Academy of Computing. She got three awards (2 Gold winners and one silver winner) for her participation in the Project "Open Innovation Laboratory for Rapid Realization for Sensing, Smart, and Sustainable Products". QS Stars Reimagine Education. She obtained seven first-place awards for Educational Innovation from Tecnologico de Monterrey. She has published more than 150 research articles in international journals and conferences. She has supervised 13 doctoral theses and 14 master's theses. Her research interests are data science, artificial intelligence, and educational innovation.

PATRICIA CARATOZZOLO received her Ph.D. from the Universitat Politécnica de Catalunya, Barcelona, Spain. She is a Full Researcher at the Institute for the Future of Education and an Assistant Professor at the School of Engineering and Sciences, Tecnologico de Monterrey, Mexico. Her areas of expertise are educational innovation, socially oriented interdisciplinary STEAM, critical and creative thinking, continuing education, upskilling and reskilling of the workforce, lifelong learning culture, and future skills. She has supervised six master's theses and published over 50 research articles in international journals and conferences. Dr. Caratozzolo is a Senior Member of IEEE, a member of Women in Engineering (WIE), the European Society for Engineering Education (SEFI), the American Society for Engineering Education (ASEE), the Executive Committee of the International Federation of Engineering Education Societies (IFEES), the National System of Researchers (SNI) of the Mexican Council of Educational Research (CONACYT), and the Executive Committee of the International Association of Continuing Engineering Education (IACEE).