Proc. of International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA 2025)
7-9 August 2025, Antalya-Türkiye

Using Natural Language Processing to Correlate University Curricula with Required Job Skills

Vidya Ramnarine
The University of the West Indies
St. Augustine, Trinidad
vidyaramnarine.vr.vr@gmail.com

Nicholas Chamansingh
The University of the West Indies
St. Augustine, Trinidad
nicholas.chamansingh@gmail.com

Patrick Hosein
The University of the West Indies
St. Augustine, Trinidad
patrick.hosein@uwi.edu

Abstract—**Aligning academic curricula with industry demands is essential for preparing graduates with relevant job-market skills. This study examines the alignment between our University's Department of Electrical and Computer Engineering (DECE) curriculum and industry job requirements across five thematic areas: Communication Systems, Computer Systems Engineering, Control Systems, Electronic Systems, and Energy Systems. Using Natural Language Processing (NLP), we extract and analyze key skills from course outlines and job postings. Web scraping with Selenium and BeautifulSoup gathers job descriptions, while SkillNER identifies required competencies. We then used Latent Dirichlet Allocation (LDA) topic modeling to assess thematic alignment. Findings reveal that, while core technical concepts are well-represented, emerging technologies and software tools are not. Alignment varies across thematic areas, hence the need for curriculum updates to better reflect industry expectations.**

Keywords—Curriculum and Job Market Alignment, Latent Dirichlet Allocation (LDA), Natural Language Processing (NLP), Web Scraping

I. Introduction

Electrical and Computer Engineering (ECE) is undergoing a significant transformation due to the evolution of technologies like artificial intelligence (AI), the Internet of Things (IoT), and data science. These advancements are reshaping industries and creating new opportunities for innovation and automation [1]. On the other hand, many educational institutions have been slow to adapt their curriculum. In particular, ECE curricula fail to adequately reflect the current technical skills and competencies [2].

This shift has exposed a widening gap between the skills taught in academia and those required by industry [3]. It is crucial to revise engineering curricula to ensure graduates are equipped with the competencies needed in today's technology-driven workplace [3]. To bridge this skills gap, data-driven methods offer a promising solution. [4] emphasizes the role of data-driven approaches in identifying the skills gap between university curricula and industry requirements.

Although some studies have examined the correlation between curricula and industry needs, these have been conducted in different regions and fields often neglecting ECE. This research aims to bridge this gap by employing text analysis techniques to assess the alignment of the Department of Electrical and Computer Engineering (DECE) curriculum with current ECE job market needs. Web scraping, Natural Language Processing (NLP), and topic modeling will be used in this study will systematically identify key areas where the curriculum meets or diverges from industry expectations. Industry data across five regions, the US, UK, Canada and the Caribbean, will be collected from leading job boards in those regions using web scraping. NLP techniques will be used to preprocess and analyze the text and topic modeling will uncover latent patterns in job descriptions and syllabi. This project will focus on analyzing both hard skills and soft skills.

This study aims to provide an assessment of the alignment between the DECE undergraduate curriculum and ECE industry needs. This study will be limited to job postings from specific regions and industries, which may not fully represent global trends. Our study provides valuable insights and recommendations that could guide future curriculum revisions, ensuring that DECE graduates are better prepared to meet the demands of the workforce. This paper is outlines as followed: a comprehensive literature review on curriculum alignment, methodology, analysis and results, discussion of key findings and finally the study by summarizing key insights and offering recommendations for curriculum enhancements.

II. Literature Review

Research typically uses web scraping and Natural Language Processing (NLP) to analyze educational content and identify alignment with industry needs. [5] employed web scraping tools such as BeautifulSoup and Scrapy to collect publicly available data, including course syllabi. They used NLTK and spaCy for text preprocessing, followed by standard text mining and statistical methods to identify recurring skills and themes in academic content. Similarly, [6] addressed the mismatch between the skills demanded by employers and the competencies fostered by university curricula.

The study used automated data extraction and analysis to compare job market requirements with university curricula. Job descriptions were scraped from tech-focused job sites, while curricula were obtained from university websites using web crawling. TF-IDF and cosine similarity were used to

9979-8-3315-3562-9/25/$31.00 ©2025 IEEE

measure the textual alignment between job postings and course content. This provided a quantitative measure of overlap in language usage between the two domains.

They found a prevalent disparity in the technology sector, where rapid advancements often outpace academic program updates. However, the study did not apply topic modeling techniques such as Latent Dirichlet Allocation (LDA), which will be used in this study. Additionally, our study also incorporates closeness metric to move beyond word similarity, enabling a more nuanced evaluation of how well the curriculum aligns with the demands of the job market.

Gurcan and Cagiltay [7] investigated the growing demand for Big Data Software Engineering (BDSE) skills by identifying the knowledge domains and technical competencies prioritized in the job market. They collected job advertisements from Indeed, focusing on roles containing BDSE-related keywords, and applied LDA-based topic modeling to uncover latent themes. Their methodology was structured into four main stages: data collection, data preprocessing, topic modeling, and competency mapping. Preprocessing included tokenization and stopword removal. Topic modeling was done using LDA with MALLET, using Gibbs sampling to identify underlying topics.

While their study provides a comprehensive analysis of industry skill requirements, it does not evaluate the alignment of these needs with academic curricula. Our study builds on this by using LDA to analyze job descriptions and comparing them directly with university syllabi to quantify curriculum alignment through a closeness metric; thereby bridging the gap between market demands and academic training.

Walker [8] explored the use of NLP to map academic curricula to relevant occupations by extracting skills from course descriptions. The study used the SkillNER Python library to identify specialized skills from data science programs at two universities. These skills were compared to occupation descriptions from the O*NET database using TF-IDF weighting and cosine similarity to measure alignment. [8] showed that SkillNER could successfully extract a significant number of specialized skills. Precision scores for occupation matching demonstrated that course descriptions contain enough information to support skill extraction.

The study also demonstrated the feasibility of using automated tools for curriculum analysis. However, the study was limited to data science programs and soft skills were underrepresented. Despite these limitations, the study provides valuable insights into the potential of automated NLP tools like SkillNER to support curriculum analysis and better align educational programs with industry requirements.

III. Methodology

This section outlines the systematic approach used to evaluate the alignment between the DECE undergraduate curriculum and ECE job requirements. The methodology integrates data collection through web scraping, and preprocessing techniques using NLP methods as well as LDA for topic modeling. The methodology is divided into several key stages.

First, web scraping tools were used to collect job descriptions from job search websites and course outlines from the department. To ensure relevance and alignment of the job postings with the DECE curriculum, job postings were filtered using search terms such as ”Computer Systems Engineer, Electronics Engineer, Power Engineer, Controls Engineer, and Communication Systems Engineer” to obtain relevant descriptions for each area. Next, the data was preprocessed using NLP preprocessing techniques to ensure consistency and relevance. Then, the skills were extracted from the data using the SkillNER [8] Python library.

The course outlines and the job descriptions were grouped according to their respective thematic areas. Topic modeling was applied to each thematic area using LDA to identify latent themes within both datasets. This approach allowed for a comparative analysis between the skills emphasized in the job market and those taught in the curriculum. It ensures the extraction of insights from unstructured textual data, enabling a data-driven evaluation of the curriculum's relevance to industry demands. The full codebase and sample data used in this study are available on GitHub[1].

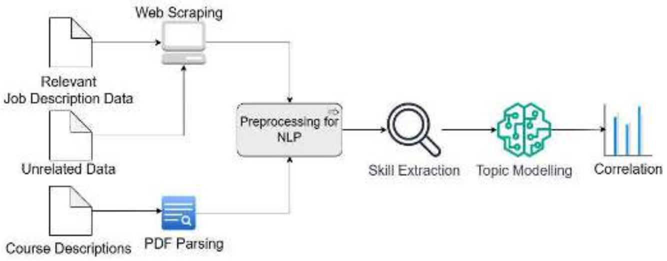

Fig. 1. Overview of Methodology

A. Data Collection

The data collection process for this project was designed to compile robust datasets of job descriptions and syllabus. These datasets served as the foundation for evaluating the alignment between the DECE curriculum and the ECE jobs. Job descriptions were sourced from five major online platforms, LinkedIn, Indeed, Reed.uk, CaribbeanJobs and Workopolis, each selected for its relevance to a specific geographical region. The selected geographic regions, Caribbean, USA, Canada, and the UK, were chosen to reflect both local employment landscapes and international migration trends among UWI graduates, ensuring the dataset captured relevant industry demands from key destinations where students often seek career opportunities.

Job postings were collected between October 2024 and March 2025 from publicly available job search websites, and course outlines were provided by the Department of Electrical and Computer Engineering at the UWI. Job postings were selected based on relevance to five thematic areas (Communication Systems, Computer Systems Engineering, Control Systems, Electronic Systems, and Energy Systems). To ensure fair and unbiased sampling, job postings were retrieved

[1] https://github.com/VidyaRamnarine/ECE-Curriculum-NLP-Analysis.git

by entering standardized, role-specific search terms such as "Power Engineer," "Controls Engineer," and "Communication Systems Engineer" were used as search terms. The results were manually verified to ensure alignment with ECE roles.

Web scraping was used to extract job descriptions from these platforms. A two-stage scraping process was developed for each platform. In the first stage, selenium was used to extract the Job posting URLs. Secondly relevant sections of the job description were extracted. BeautifulSoup [9] was used to parse the HTML content and specific sections of interest, such as Responsibilities, Qualifications and Experience were located to capture relevant sections. The objective of this study was to collect and analyze publicly available data, therefore, no ethical clearance or formal consent procedures were required. All identifiable information from the job descriptions, such as employer names, were removed to ensure privacy. The work was conducted solely for academic and research purposes, with no external funding involved and no conflicts of interest to declare.

Course outlines were obtained in PDF format from our University's Quality and Assurance Office. The PyMuPDF [10] python library was used to parse the PDF files, because it displayed strong performance in preserving word order and paragraph structure during text extraction , ensuring that downstream analysis operates on semantically accurate text. A dataset of unrelated job descriptions and course outlines formed the lower bound for correlation. This structured approach enabled consistent extraction and effective alignment analysis between industry demands and academic content.

B. Data Preprocessing

Preprocessing is a critical step in Natural Language Processing (NLP) pipelines, ensuring raw text is standardized and optimized for downstream analysis [11]. One benefit of preprocessing is the reduction of noise in the dataset [12]. This is valuable for information extraction tasks, where reducing noise can improve model performance [13]. Furthermore, specialized preprocessing methods can handle unique challenges presented by different types of text corpora [14].

The preprocessing pipeline for this study addresses the unique requirements of the data and preserves domain-specific terminology in the field of Electrical and Computer Engineering (ECE). SpaCy [15] was seleccted for implementing the preprocessing steps due to its higher tokenization ability and faster speed compared to other libraries. The job descriptions were collected from dynamic web pages hence the text contained HTML markup [13] which was removed in the web scraping stage before further processing.

The first step is non-alphabetic character removal, which filters out any unnecessary characters such as punctuation marks, numbers and special characters. [16] noted that non-alphabetic characters are commonly removed when focusing on words and phrases in text. Since natural language comprises of words that often appear in different cases, all text is converted to lowercase ensuring that the same word is consistently recognized, reducing redundancy in tokenization [11].

Then tokenization decomposes the text into a sequence of individual words [13]. Each token is filtered based on the stopword and alphabetic conditions. Stop-word removal is applied to eliminate commonly used words such as 'the,' and 'is', which has negligible impact on improving topic inference [11]. One critical decision in the preprocessing pipeline is the exclusion of stemming. This project follows the approach used by Gurcan and Cagiltay [7], where stemming was avoided in the preprocessing pipeline to preserve the domain-specific terms.

C. Skill Extraction

Technical and soft skills in all datasets were extracted. Although TF-IDF was initially considered, it overemphasized general terms due to frequency-based weighting, overlooking less frequent but highly relevant technical skills. To address this, SkillNER was used. It is a spaCy-based NER framework trained on the Lightcast Open Skills Taxonomy which contains over 32,000 skills sourced from job postings, resumes, and professional profiles [17]. It uses a multi-stage pipeline that includes pattern-based and database-driven matching, POS-based filtering, lemmatization, and n-gram detection to extract normalized skill entities [18].

Each document was then manually mapped to one or more of the DECE's five thematic areas: Communication Systems, Computer Systems, Control Systems, Electronic Systems, and Energy Systems. Where relevant, job descriptions were duplicated across categories. Additionally, the dataset comprising unrelated job descriptions and syllabi was processed in parallel to serve as a lower bound in the correlation analysis.

To assess the accuracy of the SkillNER tool, a manual validation was performed by reviewing the extracted skills from approximately 50 job descriptions. Each set of extracted skills was compared against the original job text to evaluate whether the identified terms accurately reflected the required competencies. The majority of the extractions were found to be contextually appropriate and relevant.

D. Coherence Score

An essential step in Latent Dirichlet Allocation (LDA) is determining the optimal number of topics [19]. The number of topics, commonly referred to as the k-parameter, must be carefully selected to ensure that the resulting topics are both distinct and semantically meaningful [20].To guide this selection, topic coherence was used as a diagnostic metric. Coherence measures the semantic similarity between the most significant words within a topic, providing insight into how interpretable or meaningful a topic is to humans.

High coherence scores typically correspond to more consistent and intelligible topics, while lower scores may indicate noise or overfitting. In this study the C_v, coherence metric was used. This approach combines several desirable properties of earlier coherence measures making it suitable for evaluating topic quality in moderately sized corpora. To determine the optimal number of topics, an Elbow Curve was plotted based on coherence scores to identify an inflection point indicating diminishing returns. This preliminary estimate was subsequently refined by evaluating closeness scores across candidate topic numbers to select the most suitable k.

E. Closeness Metric

The closeness metric was computed using LDA to generate topic clusters from the Job Descriptions (JD) corpus. The optimal number of topics was determined by maximizing the coherence value. The number of topics that yielded the highest coherence score was selected as the topic distribution for JD. These selected topics are referred to as the base topics (T). Next, the documents from three datasets Course Syllabi (CS), Random Unrelated (RU), and Job Descriptions (JD) were compared against the base topics, T, to assess their closeness. All associated words and its probabilities were identified for each topic in the Base Topics (T).

Then for each document in a dataset, words from the topic that appeared in the document were extracted and their probabilities were added. This process was repeated for all topics and documents. The topic with the highest summed probability was selected. The final closeness metric was computed as the average of the highest probabilities across all documents in the dataset. The topic distributions for each document were retrieved, and the associated probabilities for words in each topic were extracted. If a word in the topic appeared in the document, its frequency was multiplied by its probability, and the sum of these values were computed for each topic. The highest topic probabilities were averaged across all documents to obtain the final closeness score.

IV. Results

This section presents the results of the analysis done to evaluate the alignment between the DECE syllabus and the skills emphasized in ECE job descriptions. The key results presented are coherence scores, document to topic alignment, the closeness factor and the difference in topics for each thematic area, providing a comprehensive view of the alignment between the curriculum and job market.

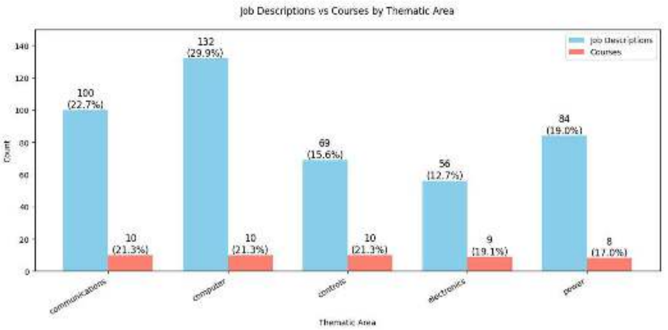

Fig. 2. Bar Chart showing Distribution of Job Descriptions and Courses by Thematic Area

As shown in Figure 2 there is a clear disparity in the number of job descriptions across thematic areas. The Computer Systems Engineering category had the most job descriptions, more than double the job descriptions collected for Electronic Systems, which had the lowest representation. In contrast, the number of courses across thematic areas remains uniform, ranging from 8 to 10 per area. The variation in the number of job descriptions collected may be attributed to how job postings were categorized on job boards. The terminology used in job descriptions may have influenced the number of postings retrieved for each thematic area. Despite using specific search terms there was a noticeable overlap with other thematic-related job postings. For example, despite searching for control systems jobs, many of the results included positions for computer systems jobs.

A. Coherence Measure

Coherence scores, particularly the C_v metric, are commonly used to assess the semantic consistency of topics. Prior studies show that coherence scores between 0.45 and 0.50 are typical when standard preprocessing is applied. For instance, [23] and [24] both reported peak coherence scores around 0.48–0.50 across different domains.

This study achieved a notably higher coherence score of 0.65 for 41 topics in the Electronics domain. However, qualitative analysis revealed that topic interpretability decreased with higher topic counts, due to redundancy and overlap. Additionally, given the dataset size, a reduced range of 1–20 topics was selected.

Within this range, the optimal number of topics for Electronics was 18, yielding a coherence score of 0.56. Similar analyses were conducted across other thematic areas, with optimal topic counts as follows: Communications (19), Computer (19), Controls (20), and Power (17). These values were chosen based on peak coherence within the practical topic range.

B. Closeness Metric

To evaluate the alignment between the ECE curriculum and industry job market requirements, a closeness metric (ρ) was developed. It was adapted from the method described in [21], where topic modeling was applied to evaluate the similarity between different sets of documents. This approach utilizes LDA to extract thematic topic probabilities from textual data, allowing a quantitative comparison between ECE job descriptions (JD), the DECE syllabus (CS) and random unrelated jobs and courses (RU).

Since JD abstracts were used to generate T, there should be significant overlap between the words in JD and words in the identified topics, making C_{JD} the upper bound. In contrast, RU documents are expected to have minimal overlap with T, making $C_{RU}$the lower bound. The CS dataset was analyzed to determine how closely it aligned with JD, as measured by the alignment metric.

$$\rho = \frac{C_{\mathrm{CS}} - C_{\mathrm{RU}}}{C_{\mathrm{JD}} - C_{\mathrm{RU}}} \tag{1}$$

where C_{JD}, C_{CS}, and C_{RU} represent the average topic closeness scores for the JD, CS, and RU datasets, respectively. A ρ value close to 1 indicates strong alignment between the curriculum and job market demands and lower values suggest weaker alignment.

To obtain a general measure of alignment across all thematic areas, a weighted ρ score was computed. This was done to prevent thematic areas with a smaller number of job descriptions but high alignment results from skewing the results. This value was determined by weighting individual ρ scores by the number of job descriptions within each thematic area.

$$\rho_w - \frac{\sum(\rho_{\text{thematic}} \times N_{\text{thematic}})}{N_{\text{total}}} \tag{2}$$

Where ρ_{thematic} is the alignment score of a thematic area, N_{thematic} is the number of job descriptions in that thematic area and N_{total} is the total number of job descriptions. The final weighted alignment score ρ_w was calculated to be 0.8038, indicating an overall alignment of approximately 80% between the ECE curriculum and industry job expectations. This suggests that around 20% of the skills demanded by the industry are not currently covered by the curriculum.

V. Discussion

Using topic modeling techniques, the study analyzed course outlines and job descriptions to identify key areas of convergence and divergence. The findings reveal varying degrees of alignment between what is taught in the DECE and the skills demanded in the ECE industry.

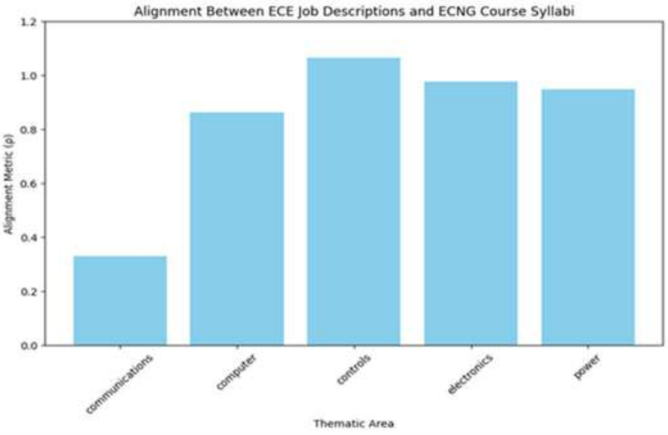

Fig. 3. Closeness Scores by Thematic Area

A significant gap was observed in the Communications thematic area, whereas the Electronics and Power areas demonstrated strong alignment. The Computer Systems area showed relatively high alignment, though several industry-standard tools were missing. In contrast, the Controls thematic area exhibited an anomaly with an unusually high closeness factor, warranting further investigation (Figure 3).

In the Communications area, only 33% of industry-required skills were covered in the curriculum, indicating misalignment. Courses focused on academic concepts such as "Fourier Transform" and "Learning Outcomes," while job postings emphasized hands-on competencies like "installation" and "troubleshooting". The observed misalignment reflects a curriculum that prioritizes theoretical foundations over practical applications—consistent with [25], who highlighted similar gaps in Jordanian universities. However, there was an overlap for soft skills such as "technical writing" and "communication skills" in the course outlines and job descriptions.

The Computer Systems thematic area achieved 86% alignment with industry skills. However, low absolute values for both C_{JD} and C_{CS} suggest dispersed topic coverage. Essential industry tools such as Git, SQL, and Linux were missing. However, "version control" was one of the skills that overlapped and "GitHub" appeared in the extracted skills for the computer thematic. Additionally, words related to Cloud Computing and DevOps such as Docker, Kubernetes and cloud platforms, were highly prevalent in job descriptions but not in courses.

For the Electronics and Power thematic areas, the alignment was very strong, at 98% and 95% respectively. Prominent overlapping terms such as, "communication skills," and "electronic circuits," in Electronics reflect a focus on the hardware and collaborative aspects of electronic design. This alignment is consistent with findings by [26] where the integration of hands-on skills with theoretical knowledge was said increases industry relevant skills. In the Power Thematic area, some of the top overlapping words like, "industry standards," "power distribution," "planning" and "leadership" suggests the curriculum covers power engineering concepts and soft skills needed in the industry.

The Controls area had a closeness factor of 1.0638, which exceeded the maximum of 1.0. This result indicated that the course syllabi contained a higher degree of alignment with the base topics derived from job descriptions than the job descriptions themselves. Given that the job descriptions served as the basis for generating the base topics, such a result was unexpected. If the course outlines include terms that overlap with the JD topics, this could inflate their closeness score.

VI. Conclusion and Recommendations

This study assessed the extent to which the DECE undergraduate curriculum aligns with the evolving demands of the ECE job market. Using Natural Language Processing (NLP) techniques, technical and soft skills were analyzed across job descriptions and curriculum. This study focused on job postings relevant to the undergraduate students at our university, aiming to highlight general skill gaps. Foundational engineering concepts were preserved, while the analysis identifies missing tools and competencies that can enhance job readiness.

The findings reveal that while the curriculum provides a solid foundation in core engineering principles, it falls short in some emerging areas. Electronics and Power Systems show strong alignment with industry expectations, whereas Communications and Computer Systems exhibit notable gaps. Control Systems showed full alignment but may reflect an overemphasis on theoretical content. Overall, the curriculum covers approximately 80% of industry-required competencies, indicating a strong foundation with room for targeted improvements.

To enhance graduate readiness, this study recommends several curriculum refinements. “On-the-move” learning through field visits and VR simulations can strengthen theory-to-practice connections. Gamification can improve engagement and analytical skills through challenge-based learning and digital rewards. In Communications Systems, where curriculum alignment is lowest, practical and simulation-based courses should be introduced alongside industry internships [3]. For Computer Systems, project-based learning should incorporate tools like Git, Docker, and Kubernetes to reflect real-world development practices. Finally, expanding industry partnerships with organizations like Cisco and Google can provide certification-based training and exposure to current technologies.

Despite the insights gained, this study has some limitations. This study focused on undergraduate curricula, yet some job postings may require graduate-level qualifications, potentially overstating skill gaps. Future work should stratify job postings by required education level and assess graduate curricula to determine whether these gaps are addressed in higher-level programs.

Lastly, the number of postings collected from the Caribbean was significantly smaller that those collected from the North American and UK market. This introduces regional bias, as the skill demands in larger economies may not fully reflect the needs of the local job market. Additionally, certain thematic areas, such as Electronics and Controls, were underrepresented in the job data, which may have skewed topic modeling outcomes and alignment scores.

References

[1] R. Chandra, A. Rajput, P. Deshpande, and J. Gain, "Unleashing innovation in electrical and electronic engineering: The role of new technology in driving progress", *J. Electr. Syst.*, vol. 20, pp. 453–460, May 2024, doi: https://doi.org/10.52783/jes.3340.

[2] G. Leandro, G. Fernandes, and C. Ferreira, "Complementary training program in electrical engineering and computer engineering undergraduate courses", *IEEE Trans. Educ.*, vol. 65, pp. 670–683, Nov. 2022, doi: https://doi.org/10.1109/te.2022.3161866.

[3] D. Gope and A. Gope, "Students and academicians views on the engineering curriculum and industrial skills requirement for a successful job career", *Open Educ. Stud.*, vol. 4, pp. 173–186, Jan. 2022, doi: https://doi.org/10.1515/edu-2022-0011.

[4] A. N. Radi, A. Aslam, Khadidos, A. O. Alaa, and S.-U. Hassan, "Bridging the skill gap between the acquired university curriculum and the requirements of the job market: A data-driven analysis of scientific literature", *J. Innov. Knowl.*, vol. 7, p. 100190, Jul. 2022, doi: https://doi.org/10.1016/j.jik.2022.100190.

[5] S. Lunn, J. Zhu, and M. Ross, "Utilizing web scraping and natural language processing to better inform pedagogical practice", in *Proc. IEEE Front. Educ. Conf.*, Oct. 2020, pp. 1–9, doi: https://doi.org/10.1109/FIE44824.2020.9274270.

[6] Y. A. Januzaj, D. Sylqa, A. Luma, and L. Gashi, "A textual content analysis model for aligning job market demands and university curricula through data mining techniques", *Int. J. Interact. Mobile Technol.(iJIM)*, vol. 18, pp. 164–176, Aug. 2024, doi: https://doi.org/10.3991/ijim.v18i14.47901.

[7] F. Gurcan and N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling", *IEEE Access*, vol. 7, pp. 82541–82552, 2019, doi: https://doi.org/10.1109/access.2019.2924075.

[8] R. Walker, "Mapping curricula to skills and occupations using course descriptions", in *Proc. IEEE World Eng. Educ. Conf. (EDUNINE)*, Mar. 2024, doi: https://doi.org/10.1109/edunine60625.2024.10500452.

[9] M. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application", *Int. J. Adv. Soft Comput. Appl.*, vol. 13, pp. 145–168, Nov. 2021, doi: https://doi.org/10.15849/ijasca.211128.11.

[10] N. S. Adhikari and S. Agarwal, "A comparative study of PDF parsing tools across diverse document categories", 2024.

[11] L. M. Chandrapati and C. K. Rao, "Descriptive answers evaluation using natural language processing approaches", *IEEE Access*, vol. 12, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3417706.

[12] A. Sharou, Z. Li, and L. Specia, "Towards a better understanding of noise in natural language processing", in *Proc. Recent Adv. Nat. Lang. Process. Conf.*, 2021, doi: https://doi.org/10.26615/978-954-452-072-4007.

[13] C. P. Chai, "Comparison of text preprocessing methods", *Nat. Lang. Eng.*, vol. 29, pp. 1–45, Jun. 2022, doi: https://doi.org/10.1017/s1351324922000213.

[14] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data", *J. Big Data*, vol. 6, Oct. 2019, doi: https://doi.org/10.1186/s40537-019-0254-8.

[15] A. Omran and C. Treude, "Choosing an NLP library for analyzing software documentation: A systematic literature review and a series of experiments", in *Proc. IEEE/ACM 14th Int. Conf. Mining Softw. Repositories (MSR)*, Singapore, Jan. 2017, pp. 187–197.

[16] G. C. Banks, H. M. Woznyj, R. S. Wesslen, and R. L. Ross, "A reviewof best practice recommendations for text analysis in R (and a user-friendly app)", *J. Bus. Psychol.*, vol. 33, pp. 445–459, Jan. 2018, doi: https://doi.org/10.1007/s10869-017-9528-3.

[17] S. Fareri, N. Melluso, F. Chiarello, and G. Fantoni, "SkillNER: Mining and mapping soft skills from any text", *Expert Syst. Appl.*, vol. 184, p. 115544, Dec. 2021, doi: https://doi.org/10.1016/j.eswa.2021.115544.

[18] SkillNer Inc., "SkillNer: Skills extractor and more", 2021. [Online]. Available: https://skillner.vercel.app/

[19] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, and A. Niekler, "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology", *Commun. Methods Meas.*, vol. 12, pp. 93–118, Feb. 2018, doi: https://doi.org/10.1080/19312458.2018.1430754.

[20] S. Sbalchiero and M. Eder, "Topic modeling, long texts and the best number of topics. Some problems and solutions", *Qual. Quant.*, vol. 54, pp. 1095–1108, Feb. 2020, doi: https://doi.org/10.1007/s11135-020-00976-w.

[21] N. Chamansingh and P. Hosein, "Using topic modelling to correlate a research institution’s outputs with its goals", in *Adv. Intell. Syst. Comput.*, pp. 147–156, Jan. 2020, doi: https://doi.org/10.1007/978-3-030-39442-4-13.

[22] S. Sahoo, J. Mati, and K. Tewari, "Multivariate Gaussian Topic Modelling: A novel approach to discover topics with greater semantic coherence", Mar. 2025.

[23] S. Bellaour, M. M. Bellaour, and I. E. Ghada, "Topic modeling: Comparison of LSA and LDA on scientific publications", in *Proc. 2021 4th Int. Conf. Data Storage Data Eng.*, Feb. 2021, doi: https://doi.org/10.1145/3456146.3456156.

[24] M. Hanafi, I. N. Nugraha, and S. Adi, "Adoption of various topic modelling algorithm to analysis Indonesian tourism customer feedback", in *Proc. 2022 4th Int. Conf. Cybern. Intell. Syst. (ICORIS)*, Nov. 2024, pp. 1–5, doi: https://doi.org/10.1109/icoris63540.2024.10903770.

[25] S. S. Alja’Afreh, A. Alabadleh, and A. Aljaafreh, "An industry-related course development for mobile communications systems in undergraduate curriculum of communications engineering", in *Proc. Int. e-Eng. Educ. Serv. Conf.*, 2021, pp. 29–35.

[26] L. S. Admuthe, A. V. Shah, and S. J. Patil, "Curriculum design and implementation of project-based learning for electronics engineering graduates", *J. Eng. Educ. Transform.*, vol. 33, Nov. 2019.