Harnessing Job Advertisement for Higher Education Curriculum Development using Text Mining Approach

Zurina Saaya and Nor Hafeizah Hassan,

Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Malaysia

Abstract
In the recent rapidly evolving job market, aligning educational curricula with industry needs is crucial to prepare graduates with relevant skills and ensuring their employability. This paper explores the potential of utilizing job advertisements as a rich and dynamic source of information for data-driven curriculum development. Traditional methods, such as industry surveys and expert consultations, often lack the immediacy and granularity to keep pace with evolving workforce demands. Job advertisements, on the other hand, provide a real-time snapshot of the specific skills, knowledge, and qualifications employers actively seek. By leveraging text mining techniques such as keywords extraction, term frequency, topic modeling and text clustering, the large corpora of job advertisements can be analyzed. The results can be used to help curriculum developers to identify the emerging trends, in-demand technical competencies and soft skills needed in their curricula. This approach enables the development of curricula tailored to the job market's requirements effectively, while bridging the gap between theoretical knowledge and practical application to give competitive advantage to all students, curricula provider and industry.

Keywords:
text mining, job advertisement, higher education, curriculum development, computer network

1. Introduction

The curriculum is the main service outcome of an education provider. An education provider is responsible for the development of a good curriculum and then serving it to the students through an appropriate learning and teaching process. Specifically for higher education providers (HEPs), developing a curriculum for tertiary education involves a multi-stage process to ensures student able to gain the knowledge and skills they need [1]. HEPs should ensure that they produce graduates that meet the current and future needs of the industry and at the same time, fulfil their obligations to society.

A formal curriculum is a structured sequence of planned educational activities designed to achieve specific learning outcomes. It outlines the order in which content within a particular subject or field [1]. Typically, a curriculum development cycle has four stages namely Plan, Implement, Monitor and Review, and Improve as illustrated in Fig. 1. Each stage involves a list of specific activities. The first step in the curriculum development process is to plan. This involves setting clear goals and objectives for the curriculum identifying the resources required for implementation considering the needs of the learners and the industry. By understanding the needs and requirements of industry, HEPs can ensure that graduates are equipped with the necessary skills and knowledge to succeed in their chosen fields.

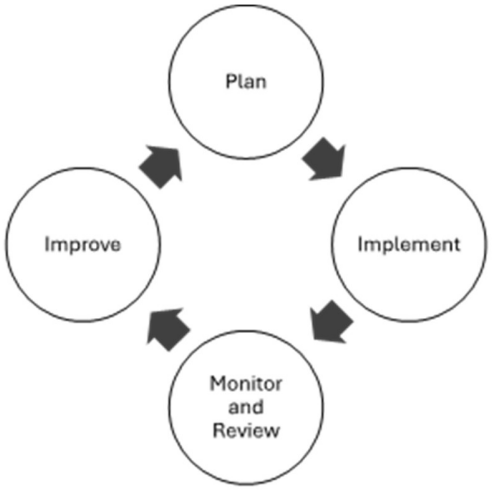

Fig. 1 Curriculum development cycle

After the planning phase, the next step is to implement the curriculum. The implementation phase involves creating the teaching materials, delivering the curriculum through various instructional methods, ensuring that the learning experiences are aligned with the learning objectives and conducting the course assessment. After the curriculum has been implemented, it is important to review its effectiveness. This involves getting feedback from students, educators, and industry stakeholders, assessing the achievement of learning outcomes, and evaluating the overall impact of the curriculum on student learning and development.

Based on the feedback and review findings, the final step is to improve the curriculum. This phase may involve making modifications to the content, instructional methods, or assessment strategies to better meet the needs and expectations of the students and the industry. It is an

Manuscript received November 5, 2025
Manuscript revised November 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.11.1

iterative process aimed at continuous improvement and ensuring that the curriculum remains relevant and effective. In the development of a curriculum, it is essential to ensure that it aligns with the needs and demands of various industries [2]. One way to achieve this is by utilizing job advertisements as a valuable source of information [3] especially during curriculum planning and review process. Job advertisements are one of mechanisms for employers to attract, recruit, and hire suitable candidates for job vacancies. Employers use job advertisements to provide comprehensive descriptions of the job role, including specific tasks and responsibilities. This helps potential candidates understand the requirements of a position and determine if their skills and experience align well with the vacancy [4]. As for the curriculum developer, they can analyze job advertisements to gain valuable insights into the skills and qualifications employers seek. This information can then be strategically incorporated into the curriculum content, ensuring its relevance and alignment with industry demands. [5].

In this paper, we will explore the significance of exploiting job advertisement for curriculum development using text mining methods. We will discuss how text mining offers a practical approach for extracting relevant information from job advertisements and how this information can be used to identify gaps and overlaps between industry need and curriculum content. In this research, the gaps are the areas where the curriculum lacks content that is crucial for job performance. These gaps highlight the need for curriculum updates or additional training programs. The overlaps are the areas where the curriculum and job descriptions align well. These findings confirm that the educational program is effectively preparing students for the job market

2. Research Background

Based on the curriculum development guidelines by Malaysian Qualifications Agency (MQA), the curriculum development process should involve stakeholders' needs analysis [1]. One of the purposes of this analysis is to ensure the curriculum is aligned with market supply and demand. These stakeholders can provide valuable information on the current job market demands and the specific skills and competencies needed in various industries [6]. Stakeholders may include employers, professional bodies, agencies that represent national interest, students, alumni and academic members.

Traditional approaches for understanding the dynamic nature of the job market often rely on surveys and consultations with industry experts, which can be time-consuming and may not be fully effective [[7]. Therefore, many studies have used job advertisements to get information about industry requirements and trends. These advertisements serve as an alternative resource for researchers to analyze the skills, qualifications, and competencies that employers seek in potential candidates. For instance, Maghsoudi [8] investigates the skill demands in computer science through content in job postings, revealing key industry trends. Similarly, Gardinera et al. [9] used job advertisement data to identify the skills and knowledge that are needed by industry for jobs within big data. Such research highlights the importance of job advertisements in understanding and addressing the evolving needs of the job market.

The application of text mining techniques to analyze job advertisements has emerged as a significant area of research. This approach leverages the power of natural language processing (NLP) to extract valuable insights from large volumes of unstructured textual data contained in job postings. One of the purposes of text mining in job advertisement analysis is to identify the skills and competencies demanded by employers. Pejić-Bach et al. [10] utilized text mining techniques specifically word frequencies and cluster analysis to identify relevant information about the needed knowledge and skills, benefiting educational institutions, human resource professionals, and individuals aiming to enter or excel in Industry 4.0 organizations.

Beside job advertisement, text mining also has been extensively applied in healthcare research to improve patient care, support clinical decision-making, and streamline medical research [11], [12]. Other than in the legal field, text mining contributes to legal research, document classification, and the analysis of legal opinions and contracts [13], [14]. In the context of curriculum development, text mining offers unique opportunities for educators and instructional designers to gain deeper insights into educational resources [15], learning materials [16] and student feedback [17] [18].

3. Methodology

Our methodology consists of three main steps: data collection, data preprocessing, and text analysis. For text analysis, we utilize text mining techniques such as keyword extraction, topic analysis, and cluster analysis. We employ Python's text mining library to calculate the term frequency-inverse document frequency (TF-IDF) score for each term in a document. Additionally, we use Voyant Tools [19] for term frequency and topic analysis, and Carrot2 [20] for cluster analysis. Figure 2 illustrates the methodology that we used in this study.

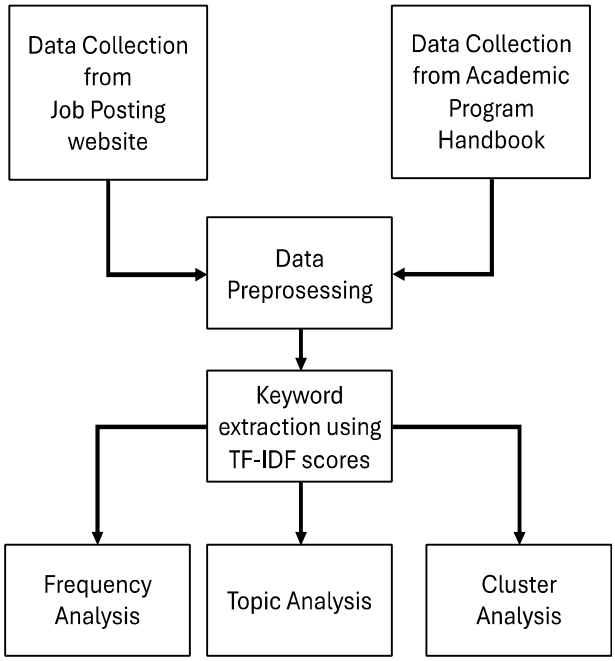

Fig. 2 Research methodology

3.1 Data Collection

For curriculum development, we have collected textual information from academic curricula in computer science specialized in computer networking from one university in Malaysia. This information is available in the Academic Program Handbook which can be accessed from the faculty website. Specifically, textual data from course summaries are gathered and it will be used for finding the overlaps and gaps between curriculum content and industry needs. As for job advertisement, the data is obtained from JobStreet, a job-posting website. JobStreet is a major online platform in Southeast Asia specifically designed for job searching and recruitment. It was founded in 1997 and has become the leading employment website in the region, operating in Malaysia, Singapore, Indonesia, and the Philippines. Job listing in Jobstreet is organised into categories and subcategories. For example, in the Information and Communication Technology category there are sub-categories such as Database Development and Administration, Testing and Quality Assurance, Business/System Analyst, Network and System Administration. To extract relevant advertisements for our target academic curricula i.e. computer networking, we focused on job advertisements that fall under categories Network and System Administration. Data collection was done in three days, from 26th to 28th April 2025. A total of 581 job advertisements were collected during this period.

3.2 Data Preprocessing

Text content is extracted from job-posting website by using web scraping method. This technique allowed us to gather a large volume of job advertisements efficiently and systematically. The initial preprocessing involved removing common English-language stop words, which are frequently occurring words such as "and," "the," and "is," that do not carry significant meaning and can skew the analysis. In addition to common English-language stop words, we also added a custom list of frequent words specific to job advertisements, such as "job," "career," "company," and other similar terms. These words exist in almost all job postings and can interfere with the extraction of meaningful keywords if not excluded. By filtering out these additional terms, we aimed to refine the dataset further and enhance the quality of our text analysis.

After the stop words and additional frequent terms were removed, the remaining text underwent further cleaning to remove any other noise such as punctuation, numbers, and special characters. This preprocessing step was crucial to ensure that the analysis focused only on the significant terms that contribute to understanding the job market and requirements. The resulting dataset, after all preprocessing steps are completed, contained 406,678 total words and 9,712 unique word forms. This pre-processed corpus provided a robust foundation for subsequent text mining and analysis, enabling us to uncover patterns and important information that reflect the underlying trends and demands in job advertisements.

3.3 Keywords Extraction

In the first part of our analysis, we use term frequency-inverse document frequency (TF-IDF) scores to extract important keywords within a document relative to a collection of documents. TF-IDF is a commonly used technique in text mining [21], [22]. It essentially counts how often a particular word appears within a document. This seemingly simple idea holds significant weight, as it allows us to understand what a document is truly about. By identifying the important keywords in the document, we can get a glimpse into the main topics and themes the document explores. For this analysis we treat each job advertisement as a document. Then we use Scikit-learn, an open-source machine learning library in Python to calculate TF-IDF scores for each term in the document (job advertisement) [23].

TF-IDF consists of three components. The first component is Term Frequency (TF). This component measures how often a term *t* appears in document *d*, refer to Eq. (1). It's calculated by dividing the number of times a term appears in a document by the total number of terms in that document. Essentially, it quantifies the importance of a term within a document. However, it does not consider the frequency of the term in the entire corpus.

$$TF(t,d) = \frac{Frequency\ term\ t\ in\ document\ d}{Total\ terms\ in\ document\ d} \quad (1)$$

The second component is Inverse Document Frequency (IDF). This component measures the importance of a term across the entire corpus. It's calculated by taking the logarithm of the total number of documents in the corpus divided by the number of documents containing the term, see Eq. (2). The IDF is higher for terms that appear in fewer documents and lower for terms that appear in many documents.

$$IDF\ (t, D) = log \frac{Total\ number\ of\ document\ in\ D}{Number\ of\ document\ containing\ term\ t} \quad (2)$$

The last component is TF-IDF score. The TF-IDF score of a term in a document combines the TF and IDF values. It is computed by multiplying the term frequency (TF) by the IDF for each term in the document, refer Eq. (3).

$$TF.IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad (3)$$

For each document we choose the top 50 terms with the highest TF-IDF scores, see Eq. (4). These scores represent the terms that are considered most significant or important within the context of the document, based on their frequency within the document and their rarity across the entire corpus, see Figure 4. To identify gaps and overlaps between job demands and the curriculum we are using the collection of course content summaries to match with the corpus in job advertisement data set. We also need to extract the terms from course content summaries based on its TD-IDF scores.

$$Top\ 50\ Terms\ = \{t_i \in T \mid rank(TF.IDF_{t_i}) \leq 50\} \quad (4)$$

```
images cloud virtual previous amazon
footprints operating workspaces
celestica citrix workstation testing
packaging concepts services etc
environments scripting azure google
endpoint terraform automated ansible
microsoft as infrastructure windows
server natting python imaging browser
web aws examples mfa vpcs regression
subnets may hosted ie typical explore
principals tasks desktops automating
deployed
```

Fig. 3. Example of a document with top 50 terms

3.4 Term Frequency

Based on the extracted keywords, we can analyze the term frequencies to identify the most important terms. This involves calculating the term frequency for each keyword and ranking them based on their frequency values.

3.5 Topic Modeling

For topic analysis, we employ a topic modeling algorithm, Latent Dirichlet Allocation (LDA) to examine job advertisements, aiming to uncover latent themes and patterns within the textual data [24]. This approach provides valuable insights into the skills and knowledge associated with various job advertisement. In previous research, LDA is used to cluster documents into topics, making it easier to organize and retrieve information. For example, Griffiths and Steyvers applied LDA to scientific literature to discover the structure of topics in the field of machine learning [25]. In another example Hong and Davison utilized LDA to analyze Twitter data, uncovering key topics and their temporal dynamics [26].

Specifically for this research we utilize the LDA algorithm, which is implemented via jsLDA by David Mimno [27]. The application of this algorithm is available in Voyant Tools. In this process, terms in each document are initially assigned randomly to a predefined number of topics, in this case 10 topics. The algorithm then iterates 50 times to refine the model, determining which terms best correspond to each topic based on their co-occurrence within the documents. Although each topic technically includes every word in the corpus, only the top ten words are displayed for clarity. The sequence of these words is significant, with the initial words contributing more substantially to defining the topic than the latter words.

3.6 Text Clustering

Text clustering is a text mining technique that involves grouping similar documents together based on their content. Text clustering also known as document clustering. The goal of text clustering is to automatically organize a large collection of text documents into meaningful clusters, where documents within the same cluster are semantically similar to each other. In this research we are using Lingo algorithm for text clustering. Lingo algorithm combines Latent Semantic Analysis (LSA) and K-means clustering to group similar documents or terms together in a semantic space [28]. Once clusters are formed, their labels can be utilized for summarization to provide insights into the underlying themes or topics present in the dataset. Each cluster obtained through Lingo represents a distinct topic or theme present in the dataset. By examining the terms within each cluster, we can gain insights into the main topics covered in the documents belonging to that cluster.

Lingo has been applied in various applications because of its effectiveness in producing meaningful and interpretable clusters. For example, in search engines

application. Lingo enhances search engines by clustering search results into coherent topics, making it easier for users to navigate through large sets of results [28]. Lingo also has been used to analyze social media data, grouping posts, tweets, and comments into clusters based on their content, The result helps in identifying trends and public sentiment [29].

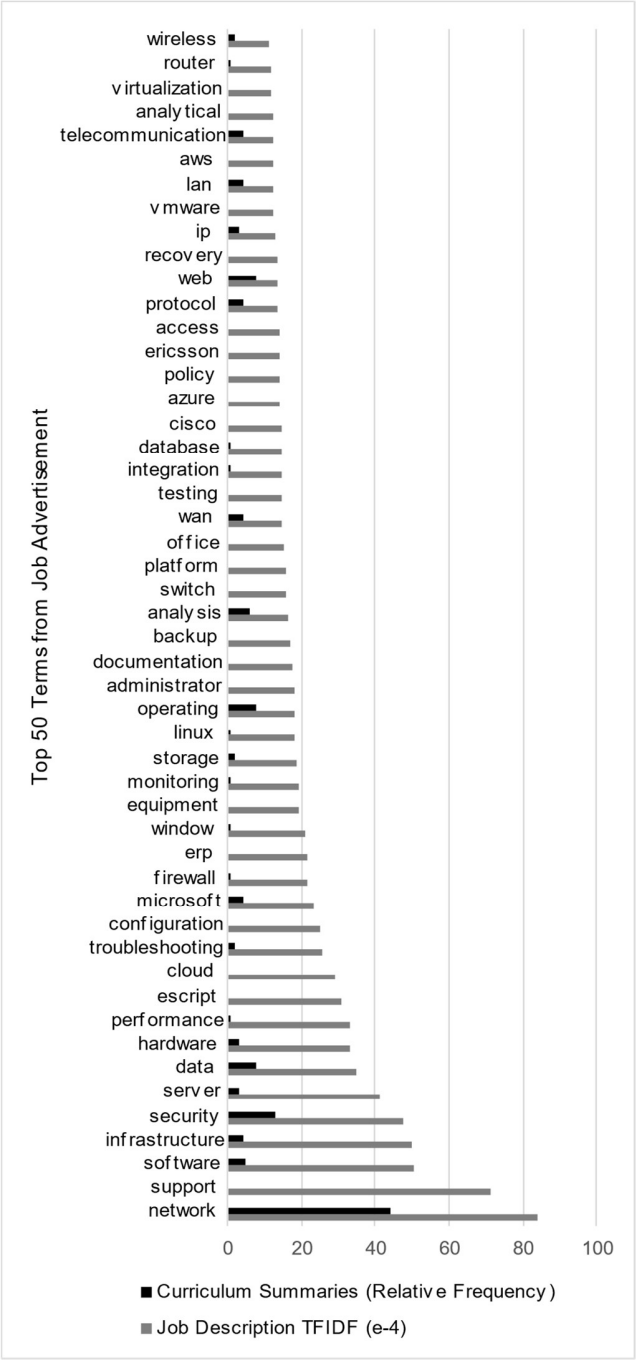

Fig. 4 Terms frequency in job advertisement and curriculum summaries

4. Result and Discussion

We performed a text mining analysis utilizing Voyant Tools, a web-based platform for text analysis and visualization designed to facilitate the exploration and interpretation of digital texts. To derive the corpus, we used a summarized version of job advertisements (top 50 terms) in conjunction with curriculum summaries, allows us to analyze term frequency, term overlaps, term gaps, and thematic topics. For term clustering, we employed Carrot2, a tool for organizing large collections of textual documents. Figure 4 presents the result of frequency analysis of keywords from both job advertisements and curriculum content. This data allows us to compare these term frequencies with the curriculum content to identify alignment with industry requirements.

As expected, "network" is the most frequent term in the text collection since the job advertisement that we search is related to computer networking. Following this are various terms representing the fundamental knowledge that correspond to a common computer networking knowledge base. By comparing the top terms in job advertisement, text collection and the curriculum summaries we can see that there are overlaps and gaps in the terms for both datasets. Next, in Figure 5, we visualize the results of these overlaps and gaps by using topic analysis that displays the top 50 terms in job advertisements, sorted by relative frequency. The white cells indicate overlapping terms, while the grey cells highlight gaps between industry needs and current curriculum content.

Figure 6 shows the result of cluster analysis with a visualization of the topics derived from the job advertisement corpus. Each topic consists of a group of terms that frequently appear together, indicating an underlying theme within the corpus. The percentages score shows how prevalent the topic is for the document, in this case the curriculum content. This visualization helps in understanding which areas are emphasized more and which ones are less covered. Results in Figure 6 are meant to help the curriculum developers, by showing which areas are heavily emphasized in industry-related documents and may suggest areas where the curriculum should focus or expand. For example, the term “firewall” has a weight of 26.9% while term “cloud” has a weight of 2.4%. The large percentage shows higher demand in the industry which could be due to several reasons, including the technology usage in many related jobs which reflects could also be the stability of the technology usage in market. A lesser percentage shows less demand in industry which could be due to less demand for the technology in the industry, that may indicate the technology is new to it. Curriculum developers may decide to maintain the high percentage category of term or technology in their syllabus and to start introducing the less percentage category in their curriculum if they never had it before.

network	equipment	azure
support	monitoring	policy
software	storage	ericsson
infrastructure	linux	access
security	operating	protocol
server	administrator	web
data	documentation	recovery
hardware	backup	ip
performance	analysis	vmware
escript	switch	lan
cloud	platform	aws
troubleshooting	office	telecommunication
configuration	wan	analytical
microsoft	testing	virtualization
firewall	integration	router
erp	database	wireless
window	cisco	

	Gap between industry needs and current curriculum content
	Overlap between industry needs and current curriculum content

Fig. 5 Gap and overlap terms between current curriculum and industry

Fig. 6 Topic Analysis

In summary, the outcome of frequency analysis, topic analysis and cluster analysis strongly indicate that text mining and extraction can effectively inform the customization of curricula to match job demand requirements. Figure 7 depicts the results of a cluster analysis performed using the Lingo algorithm in Carrot2 tool. Each rectangle in the Treemap represents a cluster of documents grouped together based on their content. The size of each rectangle indicates the number of documents within that cluster. Larger rectangles correspond to clusters with more documents. Each cluster is labeled with a representative term or phrase that summarizes the main topic of the documents within the cluster. The largest cluster which is labeled as "Operating Systems" indicating a significant focus on operating systems within the dataset. Another substantial cluster VMware, reflecting documents related to VMware, a prominent virtualization technology. Another result in Fig. 7, shows the clustering of term from the industry-related document.

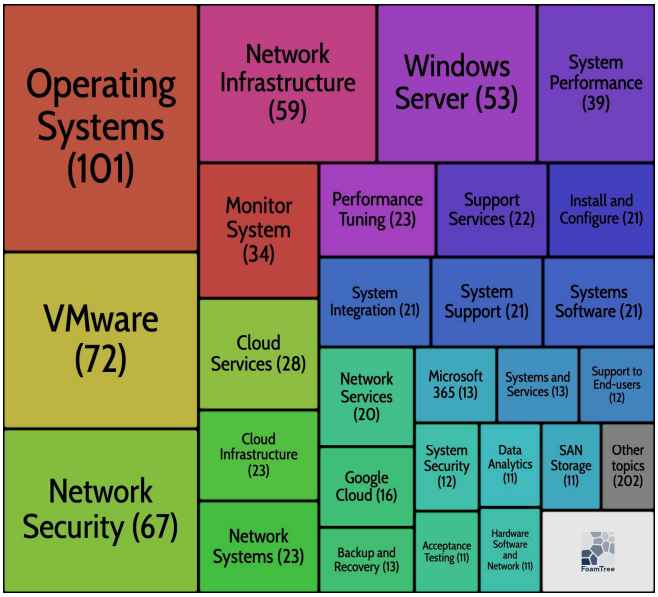

Fig. 7 Term Clustering

Based on the results in Figures 5, 6 and. 7 clearly, we can see that the gaps in the curriculum are in the area of cloud technology, data analytic and visualization. It also shows that the current curriculum is aligned with the current industry need in area of network infrastructure such as knowledge in network routing, switching and security. The result in Fig. 5 shows that out of 50 terms, 21 are not aligned between the two documents, industry-related documents and course summaries and 29 terms are aligned. This indicates that more than half of the course summaries' terms are listed in the industry-related documents, which is an important indicator as it determines that the course content could align with the industry's needs.

5. Conclusion

In conclusion, curriculum developer can gain valuable insights into the specific skills and knowledge areas that employers prioritize by using a systematic analysis of job advertisements. This process involves leveraging text mining techniques to extract keywords from job postings using TF-IDF scores. After keyword identification, text analysis, including term frequency, topic modeling, and text clustering are performed. The findings from these analyses are then compared against the existing curriculum content. By identifying the most frequently occurring terms in job advertisements, educators can evaluate the relevance of their existing curricula and instructional modules. This comparison highlights areas where the curriculum gaps and overlaps, providing a clear indicator of its alignment with industry expectations. This process involves applying text mining techniques to extract salient keywords from job postings, primarily using TF-IDF scores to highlight the most significant terms. This technique aligns with the findings of [16], which demonstrated that TF-IDF-based keyword extraction is effective for comparing course content across different universities to identify curriculum gaps.

For further analysis, curriculum developers can manually map the terms identified by their frequencies to relevant courses. This detailed mapping allows for a specific qualitative analysis of coverage, helping educators to prioritize high-impact areas for immediate curriculum updates and plan long-term educational strategies accordingly. This approach is consistent with findings in [9], which highlight that analyzing job vacancy advertisements is an effective way to identify employability skills that can inform curriculum enhancement. In summary, this approach not only confirms the curriculum's alignment with industry standards but also ensures its continuous improvement. By harnessing job advertisements through text mining, higher education providers can stay ahead of industry trends, providing students with a cutting-edge education that enhances their employability and career readiness.

Acknowledgments

The authors would like to thank the Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM) and Center of Advanced Computing Technology (C-ACT) for their incredible supports in this project.

References

[1] Malaysian Qualifications Agency, “Guidelines to Good Practices: Programme Design and Delivery First Edition: 2013 Second Edition: 2023,” 2023

[2] D. Messum, L. Wilkes, K. Peters, and D. Jackson, “Content analysis of vacancy advertisements for employability skills: Challenges and opportunities for informing curriculum development,” *Journal of Teaching and Learning for Graduate Employability*, vol. 7, no. 1, 2017, doi: 10.21153/jtlge2016vol7no1art582.

[3] N. R. Aljohani, A. Aslam, A. O. Khadidos, 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,” *Journal of Innovation & Knowledge*, vol. 7, no. 3, p. 100190, Jul. 2022, doi: 10.1016/J.JIK.2022.100190.

[4] D. A. Jones, J. W. Shultz, and D. S. Chapman, “Recruiting through job advertisements: The effects of cognitive elaboration on decision making,” *International Journal of Selection and Assessment*, vol. 14, no. 2, 2006, doi: 10.1111/j.1468-2389.2006.00342.x.

[5] Y. Ding, B. Luo, Y. Feng, and X. Mei, “Method research and system design of automatic acquire recruitment information based on Internet,” in *2016 2nd IEEE International Conference on Computer and Communications, ICCC 2016 - Proceedings*, 2017. doi: 10.1109/CompComm.2016.7925212.

[6] L. D. Spiller, D. W. Marold, H. Markovitz, and D. Sandler, “50 Ways to Enhance Student Career Success in and out of Advertising and Marketing Classrooms,” *Journal of Advertising Education*, vol. 15, no. 1, 2011, doi: 10.1177/109804821101500110.

[7] M. G. Luchs, K. S. Swan, and M. E. H. Creusen, “Perspective: A Review of Marketing Research on Product Design with Directions for Future Research,” in *Journal of Product Innovation Management*, 2016. doi: 10.1111/jpim.12276.

[8] M. Maghsoudi, “Uncovering the skillsets required in computer science jobs using social network analysis,” *Educ Inf Technol (Dordr)*, 2023, doi: 10.1007/s10639-023-12304-4.

[9] A. Gardiner, C. Aasheim, P. Rutner, and S. Williams, “Skill Requirements in Big Data: A Content Analysis of Job Advertisements,” *Journal of Computer Information Systems*, vol. 58, no. 4, 2018, doi: 10.1080/08874417.2017.1289354.

[10] M. Pejic-Bach, T. Bertoncel, M. Meško, and Ž. Krstić, “Text mining of industry 4.0 job advertisements,” *Int J Inf Manage*, vol. 50, 2020, doi: 10.1016/j.ijinfomgt.2019.07.014.

[11] U. Raja, T. Mitchell, T. Day, and J. M. Hardin, “Text mining in healthcare. Applications and opportunities.,” *J Healthc Inf Manag*, vol. 22, no. 3, 2008.

[12] W. B. van Dijk *et al.*, “Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study,” *J Clin Epidemiol*, vol. 132, 2021, doi: 10.1016/j.jclinepi.2020.11.014.

[13] O. Kovalchuk, S. Banakh, M. Masonkova, K. Berezka, S. Mokhun, and O. Fedchyshyn, “Text Mining for the

Analysis of Legal Texts," in *Proceedings - International Conference on Advanced Computer Information Technologies, ACIT*, 2022. doi: 10.1109/ACIT54803.2022.9913169.

[14] Y. L. Chen, Y. H. Liu, and W. L. Ho, "A text mining approach to assist the general public in the retrieval of legal documents," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 2, 2013, doi: 10.1002/asi.22767.

[15] J. West, "Validating curriculum development using text mining," *Curriculum Journal*, vol. 28, no. 3, 2017, doi: 10.1080/09585176.2016.1261719.

[16] K. Kawintiranon, P. Vateekul, A. Suchato, and P. Punyabukkana, "Understanding knowledge areas in curriculum through text mining from course materials," in *Proceedings of 2016 IEEE International Conference on Teaching, Assessment and Learning for Engineering, TALE 2016*, 2017. doi: 10.1109/TALE.2016.7851788.

[17] S. Gottipati, V. Shankararaman, and J. R. Lin, "Text analytics approach to extract course improvement suggestions from students' feedback," *Res Pract Technol Enhanc Learn*, vol. 13, no. 1, pp. 1–19, Dec. 2018, doi: 10.1186/S41039-018-0073-0/FIGURES/6.

[18] N. Gronberg, A. Knutas, T. Hynninen, and M. Hujala, "Palaute: An online text mining tool for analyzing written student course feedback," *IEEE Access*, vol. 9, pp. 134518–134529, 2021, doi: 10.1109/ACCESS.2021.3116425.

[19] S. and G. R. Sinclair, "Voyant Tools." Accessed: May 29, 2024. [Online]. Available: http://voyant-tools.org/

[20] S. W. D. Osinski, "Carrot2 - Open Source Search Results Clustering Engine." Accessed: Jun. 04, 2024. [Online]. Available: https://search.carrot2.org

[21] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int J Comput Appl*, vol. 181, no. 1, 2018, doi: 10.5120/ijca2018917395.

[22] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani, "A novel text mining approach based on TF-IDF and support vector machine for news classification," in *Proceedings of 2nd IEEE International Conference on Engineering and Technology, ICETECH 2016*, 2016. doi: 10.1109/ICETECH.2016.7569223.

[23] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct. 2011.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4–5, 2003, doi: 10.7551/mitpress/1120.003.0082.

[25] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc Natl Acad Sci U S A*, vol. 101, no. SUPPL. 1, 2004, doi: 10.1073/pnas.0307752101.

[26] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*, 2010. doi: 10.1145/1964858.1964870.

[27] David Mimno, "jsLDA: In-browser topic modeling." Accessed: May 28, 2024. [Online]. Available: https://mimno.infosci.cornell.edu/jsLDA/jslda.html

[28] S. Osiński and D. Weiss, "A concept-driven algorithm for clustering search results," *IEEE Intell Syst*, vol. 20, no. 3, 2005, doi: 10.1109/MIS.2005.38.

[29] S. Poomagal, P. Visalakshi, and T. Hamsapriya, "A novel method for clustering tweets in Twitter," *International Journal of Web Based Communities*, vol. 11, no. 2, 2015, doi: 10.1504/IJWBC.2015.068540.

ZURINA SAAYA is a senior lecturer in Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM) where she involved with teaching computer networking topics. Her research focuses on technologies for information retrieval, data mining and recommender systems. She received PhD in Computer Science and Informatics.

NOR HAFEIZAH is a senior lecturer in Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM) where she involved with teaching software engineering subjects. Her research focuses secured software frameworks. She received PhD in Information and Communication Technology.