Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 7, 2025

A HYBRID NLP AND CLUSTERING-BASED FRAMEWORK FOR INDUSTRY-ALIGNED ACADEMIC COURSE RECOMMENDATIONS

Muhammad Uzair[*1], Muhammad Uzair Fahim[2], Haider Tamsil[3], Ali Mujtaba Durrani[4]

[*1,2,3]*Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Pakistan*
[4]*Department of Electrical Engineering, CECOS University of IT and Emerging Sciences Peshawar, Pakistan*

[*1]muhammaduzair98@gmail.com, [2]uzairfahima3@gmail.com, [3]haidertamsil06@gmail.com, [4]ali@cecos.edu.pk

DOI: https://doi.org/10.5281/zenodo.15868411

Keywords
Academic Recommender Systems, Job Market Analysis, Clustering Algorithms, Cosine Similarity, Course Recommendation

Article History
Received: 04 April, 2025
Accepted: 25 June, 2025
Published: 12 July, 2025

Corresponding Author: *
Muhammad Uzair

Abstract
The increasing mismatch between academic programs and the fast changing industry needs pose a major predicament to institutions of higher learning that want to churn out employment-worthy graduates. In this paper, we suggest a hybrid recommendation system composed of Natural Language Processing (NLP), unsupervised clustering, topic modeling and similarity analysis to support learning by matching course content in the university with up-to-date trends in the job market. TF-IDF vectorization was applied and Dice.com data of more than 25,000 job descriptions was used to perform K-Means clustering for grouping job roles in thematic clusters. Dimensionality reduction and visualization of clusters was carried out using Principal Component Analysis (PCA) and dominant skill-based topics within groups were found using the Latent Dirichlet Allocation (LDA). By cosine similarity, these topics were aligned with the academic course outlines to determine similarity. It was found in the experiments that topics were highly semantically coherent (all with a score of more than 0.5) and the cosine similarity between courses and topics showed extreme scores (more than 0.6). The highest similarity (0.742) occurred between Data Analysis and Data Science. There is a great potential in applying the proposed system to cover the academic and enterprise divide through the implementation of data-driven dynamic course creation to meet the current workforce needs.

INTRODUCTION

This widening gap in academic curricula and the market demands necessitated by the fast-changing requirements of the job market has become a very urgent concern in contemporary systems of education. Despite all the technological innovations and open access to information, quite number of graduates of universities cannot get employment because of the ineffective, or unrelated course contents that do not appropriately mirror the industry developments. The latter can be seen, in specific, in areas of computer science and information technology in which new tools, frameworks and approaches are developed considerably more quickly than they are integrated into a university curriculum [1]. In turn, higher education institutions are increasingly pressured to re-design their curricula so that it will not respond to changing market requirements in a form of a static large-scale adjustment but rather dynamic in nature and remaining academically rigorous.

Curriculum development based on the conventional methods is relying on past enrollments, student opinions and faculty experience. As good as it is, these inputs frequently do not consider on-time labor market knowledge that presents what employers are aggressively seeking [2]. Very recent developments in artificial intelligence and natural language processing (NLP) have provided new opportunities to mine text data (e.g. job descriptions) to extract trends, skills and topics of direct interest to the industry itself [3]. These technologies, coupled with unsupervised learning tools such as clustering and topic modeling, can exhibit hidden patterns in data compiled about the job market and used to design a curriculum development process that is more tractable and specific [4].

The proposed work presents such a hybrid framework that leverages NLP, K-Means clustering, Principal Component Analysis (PCA), Latent Dirichlet Allocation (LDA) and the cosine similarity to match the academic courses as per industry trends. The system uses scraped job descriptions of the popular American-based job portal, Dice.com and parses the text identifying common job types and skills demanded in jobs. The semantic comparison of these findings with those already established academic course outlines is then made to see which course fits the present labor requirements. In comparison to the traditional course recommender systems that basically target only the data related to the performance of the students [5], the proposed course will be well said to be driven externally and with the ability to adhere to the changing technological environment. To make this analysis relevant and standardized on the academic side, course outlines of BS Computer Science program in FAST National University of Computer Science and Emerging Sciences, Islamabad, Pakistan have been used. Such course structures are per the standardized curriculum designed at the Higher Education Commission (HEC) of Pakistan which specifies the curriculum structure of all accredited four-year undergraduate Computer Science programs. Such correlation will ensure that the parallels drawn between employment patterns in the industry and in the accredited academic programs are made based on a formally accepted system of education, which further enhances the significance and scope of the offered system in context of the Pakistan Higher Education.

This research aims to bridge the existing gap between academia and industry, by methodically incorporating machine learning and NLP mechanisms into the course design process. The end result is to help the educational facilities to provide market-relevant courses to enable future-relevant degree holders to become more employable. The outcomes further confirm the feasibility of this approach and elevate a potential option of an expandable use of this method into the improvement of dynamic curriculum within various fields of study.

2. Literature Review

The mismatch between university education and the changing demands of the industry has encouraged a lot of research on academic recommendation systems as well as curriculum optimization engines. Most classical recommendation strategies in general learning environments have been mostly dependent on collaborative filtering, content-based and hybrid methods depending on grade and course selection history. Nevertheless, these models are usually not in line with existing job market dynamics, thus they cannot be effective as they pertain to equipping students into emerging job markets.

2.1 Course Recommender Systems

One of the early approaches to the problem of course recommendation researched by Kardan et al. [1] applied the concept of neural networks to make predictions regarding course selection on the basis of student related factors i.e. GPA, workload and instructor ratings. Ogun-Jankovic et al. [2] utilized Analytical Hierarchy Process (AHP) to make extraction of student preferences based on the institutional data. Gulzar et al. [3] proposed a hybrid system that enjoys the properties of both the collaborative and knowledge-based filters and consequently, when it comes to suggesting elective courses, there is greater accuracy. Nevertheless, these strategies were mostly student based and did not take into consideration signal external demand which is industry based.

2.2 Pattern discovery using Clustering Algorithms

Clustering algorithms find a popular application in unsupervised learning when it is necessary to derive a pattern out of an unstructured dataset. The efficiency and the simplicity of K-means clustering are considered to be promising [4]. More common to perform density based clustering are the techniques of DBSCAN and Mean-Shift, which can handle non-globular cluster structure [5][6]. Dendrograms of hierarchical topic mapping can be generated using hierarchical clustering, which is more expensive computationally [7]. Within the realization of course content prediction, such algorithms allow the segmentation of job descriptions into the coherent topic clusters.

2.3 Dimensionality Reduction, and Visualization

Text data in high dimensions tends to undermine interpretability. Principal Component Analysis (PCA) is popular in dropping dimensions of vector spaces in order to maintain the variance [8]. It is possible to visualize the cluster of job description in 2D in PCA, which is human-interpretable and helps confirm the coherence of the model. More recent discoveries are TMPCA (Tree-structured Multilinear PCA) [9], where the sequence has been reduced and capacity maintained to achieve the same level of classification in textual data.

2.4 Text Mining Topic Modeling

Latent Dirichlet Allocation (LDA) has proved to be an established method in terms of topic modeling because of its probabilistic models in lexical patterns in the documents [10]. LDA has also successfully found application in the job market analytics, in which each job description is modelled as a distribution of topics and each topic as a distribution over words [11]. Advances such as online LDA and correlated topic models have been able to advance the scale and consistency across large corpora [12]. The works have shown that LDA is capable of inferring market-designated keywords that subsequently can be aligned with course material [13].

2.5 Content matching similarity metrics

The semantic closeness between document vectors is usually measured with cosine similarity and Jaccard similarity. Cosine similarity, especially, is resistant to shorter or longer document length variation and always does well in high-dimensional spaces such as those in TF-IDF vectors [14]. Jaccard similarity is more effective, conversely, when the desire is to have exact overlap in tokens. Research such as Lahitani et al. [15] has confirmed the application of cosine similarity in the query to document matching of user queries and relevant documents in the educational systems.

2.6 Curriculum Design-Industry Aware

The newer initiatives have been attempting to coordinate the curriculum with real time labor market analytics. Slim et al. [16] built an academic recommenders robot based on AI that utilized both the institutional data and skill trends in online job portals. Nevertheless, not many of these systems combine clustering and topic modeling to retrieve the emerging technologies and correlate them against the academic syllabi in a dynamic manner. The hybrid system suggested in our research comprehensively fills the abovementioned gap as the K-means clustering, PCA, and LDA are incorporated in order to suggest a course topic directly based on the industry need. Since 2020, there have been recent publications that have enhanced the combination of data analytics, natural language processing and curriculum design. As noted by Roy et al. [17], the data-driven approach to CS curriculum reform should be privileged with the help of job market analytics which could guarantee the graduates with the employment skills demanded on the market. In a similar way, Morsy and Soliman [18] designed a course-recommendation system that estimates educational content through the new trends in the industry in supervised machine learning. Along this vein, Zhang and Liu [19] have performed a comparative experimentation of topic modeling techniques, i.e. LDA and NMF, to extract industry-specific skills out of job descriptions, thus showing the sustained usability of probability modeling. Balakrishna and Lin [20] presented one NLP framework to extract curricular knowledge and mine job advertisements, supporting the possibility of using live labor data taken into account when making academic considerations. More recently, Singh et al. [21] introduced a deep learning-driven adaptive curriculum framework based on BERTopic to indicate the tendency to apply advanced topic models and transformers to dynamic education development.

Table1: Comparative Analysis of Related Work in Course Recommendation and Job Market Alignment

Ref.	Study / Author(s)	Approach / Focus	Merits	Research Gap
[1]	Kardan et al. (2013)	Neural network-based course prediction using GPA, workload, and instructor rating	Personalized recommendations; adaptive to student behavior	Ignores job market trends and evolving industry requirements
[2]	Ogun-Jankovic et al. (2016)	AHP-based preference extraction from institutional data	Data-driven course prediction using institutional parameters	Lacks integration with external job data; not real-time
[3]	Gulzar et al. (2018)	Hybrid recommender (collaborative + knowledge-based)	Improved accuracy of elective recommendations	Does not consider industry skill demand; purely internal academic context
[4]	Rauf et al. (2012)	Enhanced K-means clustering for time efficiency	Scalable clustering of large datasets	Not applied to text or job-based educational content
[5]	Khan et al. (2014)	DBSCAN clustering for pattern extraction	Handles noise and complex clusters	High parameter sensitivity; less effective in sparse job data
[7]	Zhou et al. (2017)	Optimal cluster count in hierarchical clustering	Enables better cluster granularity	Computationally expensive; not used for curriculum alignment
[8]	Lhazmir et al. (2018)	PCA for dimensionality reduction in text data	Effective 2D visualization; reduces complexity	PCA alone doesn't link topics to academic needs
[10]	Blei et al. (2003)	Latent Dirichlet Allocation (LDA) for topic modeling	Semantic topic extraction from unstructured data	Needs additional layers to connect topics with actionable curricula
[14]	Rahutomo et al. (2012)	Cosine similarity in document matching	Strong in vector space similarity comparisons	Needs integration with job-market-informed content
[16]	Slim et al. (2019)	AI-driven academic program suggestion engine	Includes real-world data in course suggestion	Limited in scalability; lacks clustering and topic modeling for automation
[17]	Roy et al. (2022)	Curriculum redesign using job analytics	Redesigns CS curricula based on industry data	Focused on data mapping; lacks clustering and topic modeling layers
[18]	Morsy & Soliman (2022)	Course recommendation via machine learning	Aligns courses with job trends using ML	Lacks deep topic modeling and clustering for curriculum structure
[19]	Zhang & Liu (2022)	Topic modeling on job descriptions	Compared models (LDA, NMF) for extracting job skill themes	Needs alignment with academic structures and course content
[20]	Balakrishna & Lin (2023)	NLP for curriculum evaluation	Mined job postings to guide academic revisions	Lacks structured similarity matching with university curricula

[21]	Singh et al. (2023)	Deep learning in curriculum design (BERTopic)	Adaptive course frameworks using deep topic modeling	Model complexity increases computational cost; limited deployment reports
~	**This Work (2025)**	NLP + Clustering + LDA-based course recommendation	Combines real-world job trends, clustering, topic modeling, and cosine similarity; scalable and interpretable	**Bridges academia–industry gap using automated job-to-course alignment in real-time educational contexts**

These developments are indicative of the academic trend of closing the divider between industry and academia through the application of scalable intelligent systems; however, few take into consideration an end-to-end solution comprising of clustering, topic modelling and similarity-based course matching as the proposed current work. Although, there are strong achievements in course recommender system and unsupervised learning algorithm, most of the current solutions address either only internal academic data or do not reflect the real-time demands of the industry.

Although cluster analysis algorithms and topic modeling have been useful in text analytics and pattern recognition, there is little application of the technology in curriculum design. Furthermore, complete integration of job market analysis, dimensionality reduction, semantic topic model and a course matching method founded on similarity are hardly used by any similar system with the exception of this one. The work mitigates these shortcomings by presenting a new data-driven paradigm where the academic course content is related to the latest trends in the industry because the methods of NLP, K-means clustering, PCA, and LDA, are combined with the cosine similarity. The proposed system has the potential of being a scalable and viable dynamic course recommendation system in higher education through linking job market dynamics with education curriculum.

3. Methodology

The section gives the outline of the whole process of developing a course suggestion system that fits in the industry. NLP, unsupervised clustering, topic modeling, and similarity analysis work together in the system to make the necessary match between the latest trends in the job market and the contents of the academic curricula on demand. Its methodology can be described as consisting in seven major steps, including data acquisition, preprocessing, feature extraction, clustering, dimensionality reduction, topic modeling, and course matching. Figure 1 shows the entire workflow of the proposed industry-aligned course recommendation system. This is followed by extracting job description data based on credible online job boards who are then made to undergo data preprocessing techniques of noise clean up, tokenization, lemmatization and TF-IDF techniques of vectorization.

The K-Means clustering algorithm is used to group the job roles using these vectored representations into thematic groups. At the same time, Principal Component Analysis (PCA) is used to reduce the dimensions of the data to help interpret the data and to facilitate visualization of the data. Top dominant topics of each job cluster are then extracted using Latent Dirichlet Allocation (LDA). The selected topics will be compared with course outline of the academic subject material by cosine similarity; a measure of semantic matching. The result of this workflow is a list of recommended courses that best fit the existing industry requirements hence assisting the academic institutions to formulate and revise their curriculums accordingly.

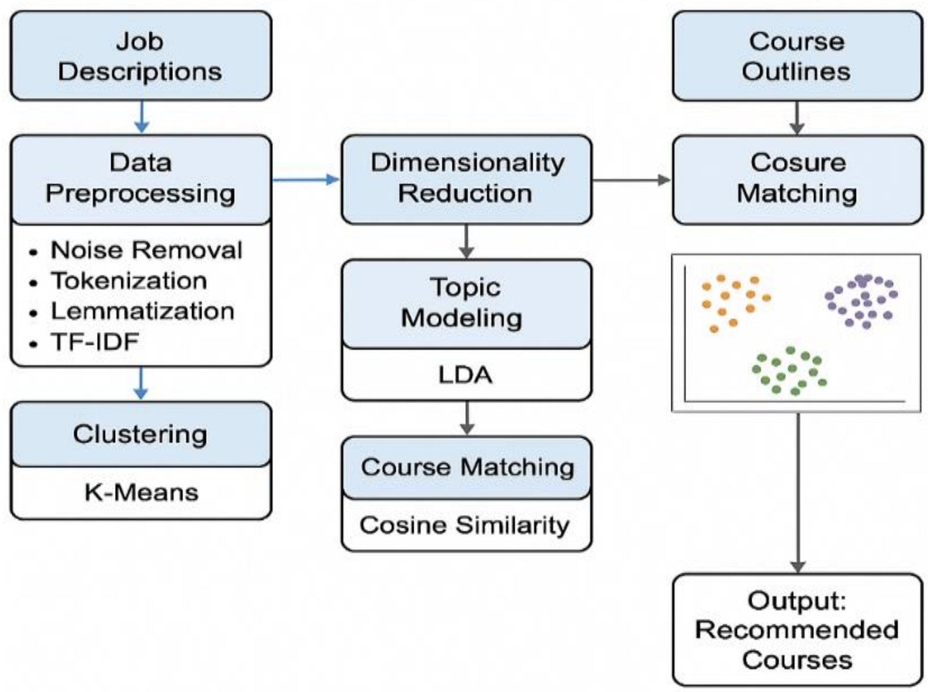

Figure 1: Workflow of Course Recommendation System

3.1 Data Collection

Two major sets of data are used in the framework, job description and academic course outlines. The job dataset is a collection of more than 25000 IT jobs advertising gathered form Dice.com a current U.S based job web portal. In these advertisements, there is textual information rich in unstructured textual material regarding the needed skills, technologies and responsibilities. This is supplemented by a corpus of course syllabi scraped off the Web sites of computer science departments of academic institutions. Every syllabus contains the name of the course and the description of learning outcomes, which made it possible to compare to the industry relevant topics.

3.2 Data Preprocessing

A thorough preprocessing was conducted on both datasets to guarantee good text modeling. The cleaning was first done by spotting noise identified as HTML tags, special characters, email addresses, and punctuations. The tokenization was done on the cleaned text by getting the individual words. Lemmatization was also used in order to loose inflected words back to their stems, where it was hoped they could be standardized (e.g. programming turns into program). Also, some prevalent stop words like is, the, and were excluded. In the end the corpus was restructured into numerical feature vectors with the help of Term Frequency Inverse Document Frequency (TF-IDF) tool which can be read as a quantitative signification of the significance of the words within each document verses the whole folder.

3.3 Clustering of Job Descriptions

After the process of vectorization, the TF-IDF generated vectors were clustered into the K-Means algorithm. This semi-supervised learning method was selected because of its scalability and grouping capabilities of similar job advert using the usage of words. Elbow Method was then used in deciding the optimum number of clusters (K) whereby the Within-Cluster Sum of Squares (WCSS) was plotted against the varying values of K. The flattening that occurred on the curve was chosen as the best K. Such a clustering process allowed revealing hidden job groups like Cloud Services, Web Development, and Data Analytics.

3.4 Dimensionality Reduction using PCA

As a smoothing preparation toward cluster interpretation and visual analysis, the high-dimensional TF-IDF vectors were transformed to 2D using Principal Component Analysis (PCA). This dimensionality reduction is such that it maintained the structure of the data whereas it allows visual inspection of different cluster boundaries. PCA was also quite exploited in confirming the efficacy of K-Means clustering where it identified compact and high-separated clusters of jobs.

3.5 Topic Modeling with LDA

As a smoothing preparation toward cluster interpretation and visual analysis, the high-dimensional TF-IDF vectors were transformed to 2D using Principal Component Analysis (PCA). This dimensionality reduction is such that it maintained the structure of the data whereas it allows visual inspection of different cluster boundaries. PCA was also quite exploited in confirming the efficacy of K-Means clustering where it identified compact and high-separated clusters of jobs.

3.6 Course Matching via Similarity Metrics

The topic keywords derived were then compared against the university course outlines in order to determine their suitability to the current industry trends. Correlation between LDA-derived keywords and course material was assessed mostly through Cosine similarity between the TF-IDF vectors of the words among themselves. It was selected, both because it is robust in high-dimensional vector spaces, and sensitive to term frequency and context. To compare with, Jaccard similarity was also measured to compute set-based overlap. The cosine similarities scores were taken as the basis of highly aligned courses as those with values above 0.6.

3.7 Evaluation Metrics

To verify the success of the clustering and topic modeling phases, a number of measures were used to evaluate the same. In case of LDA, the quality of the models was evaluated through perplexity score and the value of the log likelihood, which measures the degree to which the model is being generalized on the untested documents. Also, the topic coherence was computed to determine the semantic similarity between the highest keywords on each topic hence determining the topic interpretability. The last similarity percentages of job clusters and course outlines were used to determine close matches and offer practical information on how the curriculum can be designed and improved.

The suggested methodology provides an orderly, multi-phased process that is highly effective in using Natural Language Processing, unsupervised cluster, and topic model approaches to cover the gap between the academic courses and the changing needs of the industry. This system will be used in curriculum enhancement, as it will extract meaningful structure in scale job descriptions, and match it with course outlines at universities using similarity analysis, providing the most scalable and intelligent approach to dynamic curriculum enhancements. This end-to-end solution does not only make academic material future-relevant and applicable but also gives power to the institutions to take data-based decisions in education planning.

4. Results and Discussion

The validity of the proposed system was confirmed through a number of experiments with more than 25 000 job descriptions of the Dice.com database and a set of university course outlines. Three main features were put into consideration on the assessment chart, which included cluster formation, topic extraction, and course alignment based on the similarity score. At first, the optimum value of the number of clusters (K) was estimated using the Elbow Method to apply K-Means clustering. The curve started flattening at K = 6 which means that at this point six clusters were adequate to capture the main themes that were contained in the job dataset. Among the most important areas in the industry, that these clusters were observed to relate with include Cloud Services, Frontend Development, Data Analysis, Cybersecurity, DevOps, and Database Management. After K-Means clustering the dataset, Principal Component Analysis (PCA) was used on the TF-IDF vectors in order to decrease its dimensionality, as well as visualize the created clusters in a two-dimensional space. The 2D scatter plot revealed distinguishable and coherent clusters whose reduction to viable clusters is a tasteless piece of pie. This triggers the suitability of K-Means in initiating meaningful groupings of the data. Then

Latent Dirichlet Allocation (LDA) was applied to every cluster to indicate key words on each job type. Cosine similarity was then compared with these topic key words to the academic course outlines. The summary of similarity scores of cluster topics created by the LDA and the most similar academic courses. Interestingly, Data Analysis cluster correlated with the Data Science course with a 0.742 of cosine similarity, and the Cloud Services cluster also performed well with the Distributed Systems course (0.681). On the same note, Frontend Development cluster was highly related to Web Technologies (0.698), confirming that the system has a capability of recommending industry-relevant professional literature to the user according to the academic realistic demand.

Table 2: Course–Cluster Similarity Scores

Cluster Topic	Closest Course Match	Cosine Similarity Score
Cloud Services	Distributed Systems	0.681
Data Analysis	Data Science	0.742
Frontend Development	Web Technologies	0.698
Cybersecurity	Information Security	0.654
DevOps	Software Engineering	0.637
Database Management	Database Systems	0.672

Elbow plot in figure 2 helps in identifying an optimal number of clusters (k) to be used by the K-Means algorithm. To note the curve exhibits a sharp decrease in within cluster sum of squares (WCSS) by increasing k and then it shows an evident inflection around k = 6. This is where inclusion of more clusters will lead to diminishing returns meaning that six clusters adequately represent the structural diversity that is found in the job descriptions. This finding supports the empirical approach to categorizing the data source into six different thematic regions, including cloud computing, data analytics, frontend development, cybersecurity, DevOps, and database management.

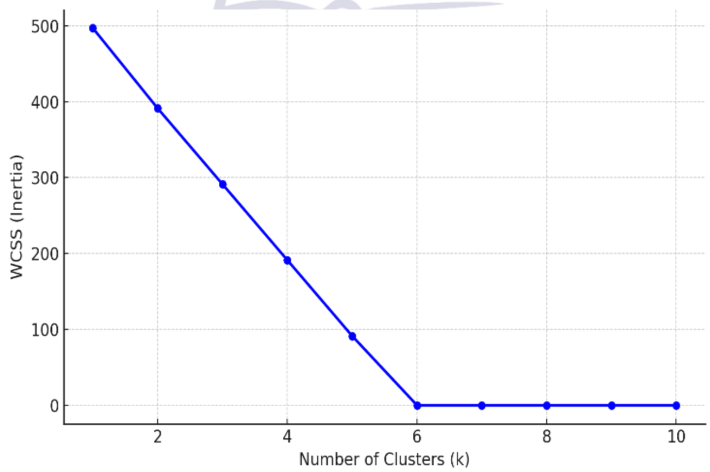

Figure 2: Elbow Method for Determining Optimal Clusters

A Principal Component Analysis (PCA) projection in 2 dimensions of the job description data (that is, the high-dimensional data) was calculated in order to prove the results of the clustering approach and show them graphically. The PCA scatter plot in figure 3 shows well-defined and more or less separable clusters, where each of the points corresponds to a single job description and is painted by its respective cluster. The graph reaffirms the accuracy of the K-Means algorithm in pairing similar jobs, in terms of semantic orientation, where there is insignificant overlapping and clusters that are tightly bound (cohesive). Such spatial cohesion helps confirm the strength of the clustering process and it will give a visual foundation of the semantic themes that are the most common in the data.

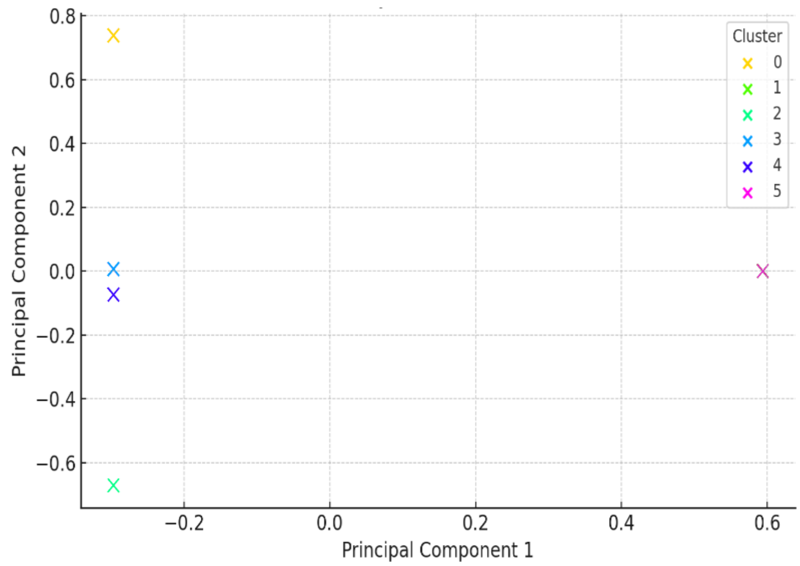

Figure 3: PCA Visualization of Clustered Job Descriptions

After clustering, Latent Dirichlet Allocation (LDA) topic modeling was performed to reveal the topic coherence scores that is shown in figure 3 which assess the quality of semantics achieved by the extracted topics. All the six clusters scored a coherence value that was much higher than 0.5, which is a generally accepted interpretability value in terms of topic modeling. Looking at figure 4 it is important to note that the data analysis cluster produced the highest coherence score of 0.69 because the keywords used are very cohesive in talking about the same topic. Other groups including cloud computing and database management had performed so well, whereas DevOps and cybersecurity recorded slightly lower but satisfactory scores. These findings confirm that the extracted topics are not merely meaningful but appropriate to downstream the comparison with academic courses content.

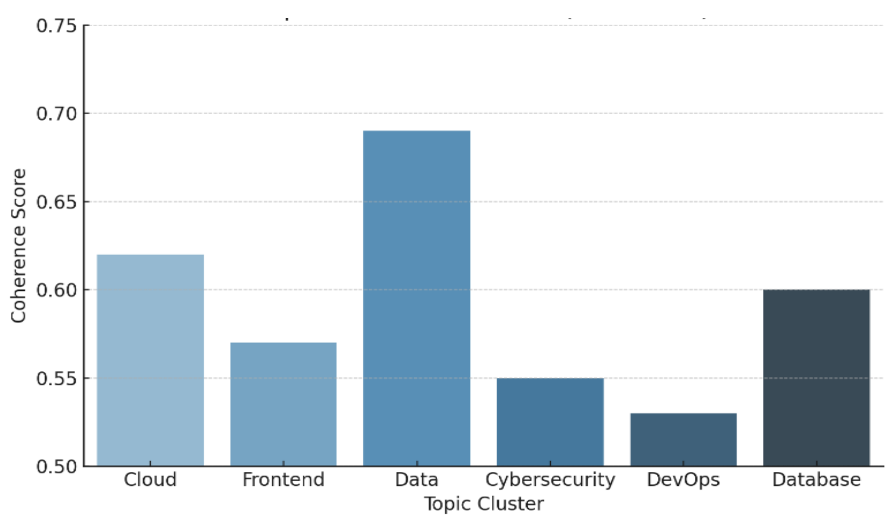

Figure 4: Topic Coherence Scores across Job Clusters

Lastly, the value of the cosine similarity between the job cluster subtopics and course outlines at the university were computed and graphically represented in figure 5. As the bar chart of similarity scores demonstrates, there was a high correspondence between a number of industrial themes and respective academic courses. An example is that the cluster of data analysis brought a topic most suitable to a course in data science with a similarity score of 0.742. On the same note, cloud services indicated a large similarity (0.681) to distributed systems whereas frontend development was effective under a score of 0.698 under the web technologies. The results of all identified matches were over 0.6, which means that there is an assertive degree of semantic overlap and proves the possibility of recommendations of industry-appropriate academic courses that are based on data within the job market. The substantiation of the core hypothesis by these quantitative findings means that the proposed hybrid model, where NLP, clustering and topic modeling are applied together, will close that divide between academia and industry by adapting curriculum and curriculum changes to new and existing professional needs.

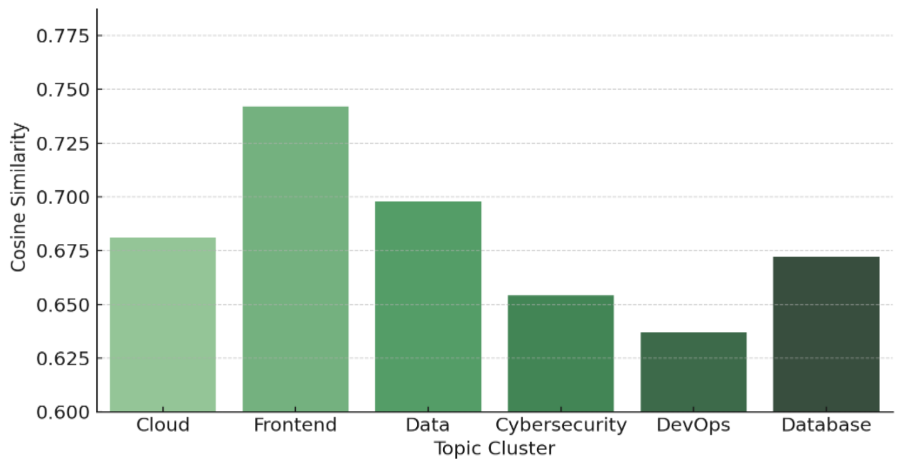

Figure 5: Cosine Similarity between Job Clusters and Academic Courses

The outcomes of the proposed framework are very clear indicating the usefulness of the framework in matching course contents in academic institutions according to the current industry demands. Results of the elbow method and the PCA visualization were the confirmation that six clusters were the optimal representation of the major themes in the job description and qualitative confirmation, respectively, since each topic had its separate and distinct grouping. The results were further reinforced by topic modeling LDA which provided coherent topics which were semantically meaningful (all their coherence scores were above the acceptable limit). However, the strongest evidence is presented by the cosine similarity scores between job clusters and academic courses, as all values exceeded 0.6 meaning the high level of correspondence. It is interesting to note that the highest-scoring matches Data Analysis to Data Science, Cloud Services to Distributed Systems highlight the applicability of this system as a means of filling the academia-industry gap. The combination of these findings confirms the whole design of the system, including the possibility of its use as the dynamic tool of the curriculum development correlated with the reference to the up-to-the-minute trends within the labor market.

5. Conclusion

To sum up, the suggested hybrid framework managed to combine NLP-based preprocessing, K-Means clustering, Principal Component Analysis, and Latent Dirichlet Allocation and create the data-driven course recommendation system based on the existing industry requirements. The methodology starts with converting more than 25,000 job descriptions into the TF-IDF vectors and unsupervised clustering the job domains out of that. PCA was used to achieve the dimension reduction and its interpretability, and LDA was able to form much meaningful topic keywords within one cluster. They were then compared with course outlines in academia with cosine similarity as the measure of relevance. The main strength of the methodology is its structured and scalable design that shall allow dynamic mapping of the curriculum to avoid the usage of static information records held by the institutions. The quantitative analysis proves the correctness and the relevance of the system: elbow position confirmed six best clusters, the projection of the PCA supported peculiar grouping of topics, and the topics of LDA obtained coherence means above 0.5, whereas cosine similarity scores measuring course alignment remained above 0.6 (the top score was 0.742 between Data Analysis and Data Science). These results indicate the technical soundness of the framework not least, and the practicality of the frame to reform the curriculum design; decrease the academia-industry gap, and increase the employability of graduates in changeable job markets.

References

[1] Kardan et al., "Prediction of Student Course Selection in Online Higher Education Institutes using Neural Network," *Computers & Education*, vol. 65, pp. 1–11, 2013.
[2] Ogun-Jankovic et al., "Using Institutional Data to Predict Student Course Selections," *Internet and Higher Education*, vol. 29, pp. 49–62, 2016.
[3] Z. Gulzar et al., "PCRS: Personalized Course Recommender System Based on Hybrid Approach," *Procedia Computer Science*, vol. 125, pp. 518–524, 2018.
[4] Rauf et al., "Enhanced K-means Clustering Algorithm to Reduce Iterations and Time Complexity," *Middle-East Journal of Scientific Research*, vol. 12, pp. 959–963, 2012.
[5] K. Khan et al., "DBSCAN: Past, Present and Future," in *Proc. ICADIWT*, pp. 232–238, 2014.
[6] D. Demirovic, "An Implementation of the Mean Shift Algorithm," Univ. of Tuzla.
[7] S. Zhou et al., "Optimal Cluster Number via Agglomerative Hierarchical Clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 3007–3017, 2017.
[8] S. Lhazmir et al., "Feature Extraction via PCA for Text Categorization," *Proc. 2018 Int. Conf. BDAW*, pp. 1–6.
[9] Y. Su et al., "Efficient Text Classification Using TMPCA," *Proc. ICPR*, Beijing, 2018.
[10] D. Blei et al., "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[11] H. Jelodar et al., "LDA and Topic Modeling: Applications and Survey," *Multimedia Tools Appl.*, vol. 78, pp. 15169–15211, 2019.
[12] M. Hoffman et al., "Online Learning for LDA," *Advances in Neural Information Processing Systems*, vol. 23, pp. 856–864, 2010.
[13] Gross and D. Murthy, "Modeling Virtual Organizations with LDA," *Neural Networks*, vol. 58, pp. 38–49, 2014.
[14] F. Rahutomo et al., "Semantic Cosine Similarity," *Proc. ICAST*, 2012.
[15] Lahitani et al., "Cosine Similarity for Online Essay Assessment," *Proc. Cyber and IT Service Management*, IEEE, 2016.
[16] Slim et al., "An Automated Framework to Recommend Academic Programs and Courses," *Proc. IEEE BigDataService*, pp. 145–150, 2019.
[17] Roy, R. Dutta, and S. Mitra, "Bridging the Skills Gap: Data-Driven Curriculum Design for Computer Science Programs," *IEEE Transactions on Education*, vol. 65, no. 3, pp. 270–278, 2022.
[18] M. M. Morsy and A. Soliman, "Course Recommendation System Using Job Market Trends and Machine Learning," *International Journal of Information Management Data Insights*, vol. 2, no. 2, pp. 100092, 2022.
[19] Zhang and Z. Liu, "A Comparative Study of Topic Models for Industry Skill Extraction from Job Descriptions," *Information Processing & Management*, vol. 59, no. 3, 102920, 2022.
[20] S. Balakrishna and J. Lin, "Mining Online Job Postings with NLP for Curriculum Insights," *Computers & Education: Artificial Intelligence*, vol. 5, pp. 100083, 2023.
[21] Singh, M. Kumar, and P. Sharma, "Deep Learning-Driven Framework for Adaptive Curriculum Design in Higher Education," *Education and Information Technologies*, vol. 28, pp. 8793–8812, 2023.