Scientific Research Publishing

Open Journal of Statistics, 2025, 15(6), 492-512
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Analysis of the Match between University Skills and Labour Market Requirements in the Democratic Republic of Congo: A Semantic and Probabilistic Approach

Sindani Bukerimanza Moise[1], Mwiseneza Sebigunda Norbert[2]

[1]Faculty of Business Sciences, Haute Ecole de Commerce, Kinshasa, DRC
[2]Faculty of Communication Sciences, University of Dar es Salaam, Dar es Salaam, Tanzania
Email: sindanimoise985@gmail.com, sindani.moise@heckin.ac.cd, mosheingnorbert@gmail.com

How to cite this paper: Moise, S.B. and Norbert, M.S. (2025) Analysis of the Match between University Skills and Labour Market Requirements in the Democratic Republic of Congo: A Semantic and Probabilistic Approach. *Open Journal of Statistics*, **15**, 492-512.
https://doi.org/10.4236/ojs.2025.156026

Received: October 23, 2025
Accepted: December 21, 2025
Published: December 24, 2025

Abstract

This paper examines the extent to which competencies taught in Congolese universities match the skills required by the labour market. Using official curricula from the Ministry of Higher and University Education and 1560 LinkedIn job postings (2023-2025), we constructed two textual corpora and measured their semantic proximity with spaCy word embeddings and cosine similarity. A logistic regression model was then estimated to predict employability as a function of semantic similarity and academic domain. Results show an overall adequacy rate of 49.14%, meaning that only one out of two university skills is directly relevant to employers' expectations. The model displays strong predictive performance (Pseudo R^2 = 0.6899; AUC = 0.977; Accuracy = 91.8%), and confirms that semantic similarity alone explains about 68% of employability variance, far more than the field of study. Economics and Management and Legal-Political Sciences are the best-aligned domains, whereas Technology, Psychological and Educational Sciences show lower alignment. These findings suggest that curriculum reforms in the DRC should prioritise skill alignment (digital, languages, data, management) over programme expansion. The study offers a reproducible semantic-probabilistic protocol for ministries and universities to monitor skill-labour match.

Keywords

Employability, Semantic Analysis, Cosine Similarity, spaCy, Logistic Regression, Higher Education, Labour-Market Requirements, Curriculum Alignment, Skills Mismatch, Democratic Republic of Congo

1. Introduction

In many developing countries, higher education systems struggle to keep pace with the rapid transformation of labour market demands. The emergence of digital economies, automation, and new business models has fundamentally reshaped the nature of employable skills, giving rise to a growing concern about the adequacy of university training. In Sub-Saharan Africa, several reports from the World Bank [1], the African Development Bank [2] and the ILO [3] reveal that a large proportion of graduates possess academic knowledge but lack the transversal, digital, and practical skills required by employers. This gap has become a major structural constraint to youth employment, innovation, and economic productivity.

In the Democratic Republic of Congo (DRC), this issue is particularly acute. Although the country has more than 400 higher education institutions, the employability of graduates remains low. Most curricula are still discipline-based, while labour markets increasingly demand multidisciplinary profiles able to integrate technology, management, and communication skills. Studies have shown that the disconnection between academic programmes and professional needs limits the capacity of young graduates to secure sustainable jobs, even when labour demand exists in growth sectors such as digital services, energy, or finance [4] [5].

Traditional approaches to measuring skill mismatch have relied on surveys or descriptive statistics, which capture perceptions but fail to quantify the semantic distance between what universities teach and what employers require. However, recent advances in natural language processing (NLP) allow for a more objective comparison of textual data such as curricula and job descriptions. By combining semantic similarity techniques with probabilistic modelling, researchers can now quantify how well academic skills align with occupational requirements and estimate their actual contribution to employability.

This study applies such a combined semantic–probabilistic approach to the case of the DRC. Using official university curricula and LinkedIn job postings (2023-2025), the analysis computes semantic similarity scores with the spaCy language model and integrates them into a logistic regression predicting the probability of employability according to both semantic alignment and academic domain. Beyond describing the skill gap, the paper aims to provide an empirical framework for monitoring the correspondence between education and employment. The results highlight the extent to which Congolese university training responds to market needs and identify the fields where curricular reforms are most urgent.

The remainder of this paper is structured as follows: Section 2 reviews the theoretical and empirical literature on skills mismatch and employability. Section 3 details the methodological approach and data construction. Section 4 presents and discusses the main results, while Section 5 concludes with policy recommendations for aligning higher education with labour market expectations in the DRC.

Abbreviations and Acronyms

- DRC: Democratic Republic of the Congo
- MINESU: Ministère de l'Enseignement Supérieur et Universitaire
- PCA/CBA: Approche par Compétences (Competency-Based Approach, CBA)
- LMD: Licence—Master—Doctorat

2. Literature Review

2.1. Conceptual Background

The relationship between higher education and employability has long been central to human capital theory, which argues that education increases productivity and thus labour-market value. However, in the context of developing economies, the translation of educational attainment into employment is far from automatic. The growing complexity of production systems, the expansion of the digital economy, and the shift toward knowledge-intensive services have created new skill requirements that traditional academic programmes often fail to meet [1] [6].

Employability is no longer determined solely by formal qualifications but by a mix of technical, cognitive, and transversal competencies, such as problem-solving, communication, adaptability, and digital literacy [3]. Consequently, the analysis of graduate employability must go beyond simple enrollment or unemployment statistics and focus on the alignment between education and actual job requirements.

2.2. Empirical Evidence in Africa and the DRC

In Sub-Saharan Africa, several studies have highlighted persistent skill mismatches between university training and labour-market expectations. The African Development Bank and IFEF/UNESCO emphasize that most universities continue to operate within discipline-based systems with limited integration of soft and digital skills [7]. Employers across the continent report difficulties finding graduates with the competencies required for emerging sectors such as ICT, finance, renewable energy, and logistics.

In the Democratic Republic of Congo (DRC), this challenge is particularly pronounced. Research by Bwanga and Arionzi reveals that while the number of graduates has increased substantially since the implementation of the LMD (Bachelor–Master–Doctorate) system, their transition to formal employment remains weak. The authors attribute this gap to curricula that emphasize theoretical knowledge over practical application and to insufficient collaboration between universities and the private sector. Similarly, Christelle and Moïse note that employers frequently report deficits in language, analytical, and ICT skills among recent graduates, despite their academic credentials [8].

2.3. Methodological Approaches in Existing Studies

Most empirical studies on skill mismatch in Africa rely on survey data and descriptive analyses of graduates' employment status. While informative, such ap-

proaches often remain subjective and limited in scope. Only a few recent contributions have applied quantitative text-based techniques to measure how well educational content corresponds to occupational needs. The OECD proposed a semantic mapping framework comparing skill taxonomies across education and employment databases, showing that semantic similarity can serve as a reliable proxy for skill adequacy [6].

In the African context, however, such methods remain under-used. Existing research generally focuses on employability perceptions rather than the linguistic content of training and job descriptions. To date, no study in the DRC has combined natural language processing (NLP) with probabilistic modeling to quantify employability outcomes based on skill alignment. This study therefore fills a critical methodological gap by integrating semantic similarity metrics (via the spaCy model) into a logistic regression predicting the probability of graduate employability.

2.4. Research Gap and Contribution

The literature consistently points to a need for evidence-based frameworks capable of translating qualitative observations about skill mismatch into measurable indicators. By adopting a semantic–probabilistic approach, this study provides a reproducible tool for diagnosing how well academic programmes align with market requirements. It builds on international advances in text analytics while adapting them to the Congolese context, where statistical data are scarce but textual information (curricula and job postings) is abundant.

The main contribution of this work is therefore twofold:

1) It quantifies the semantic proximity between university competencies and job-market skills using large-scale textual data;

2) It models the probability of employability as a function of that semantic alignment and the academic domain.

This approach bridges the gap between educational research and labour-market analysis, offering practical insights for policymakers seeking to modernize higher education and enhance graduates' professional integration in the DRC.

3. Data and Method

This study adopts a quantitative and mixed approach, combining semantic analysis of skills and probabilistic modelling to assess the match between the skills taught in Congolese universities and those demanded by the labour market on LinkedIn between 2023 and 2025.

3.1. Methodological Proposal

Drawing from the semantic–probabilistic approaches identified in the literature, this study proposes an operational framework adapted to the Congolese context for assessing the adequacy between university training and labour-market requirements. The methodological logic follows six interconnected stages combin-

ing semantic analysis and probabilistic modelling.

Step 1: Corpus design and data integration.

The protocol begins with the systematic collection and structuring of two complementary corpora: university curricula officially published by the Ministry of Higher and University Education (MINESU), and LinkedIn job offers (2023-2025) representing real market demands. This dual-source design ensures representativeness of both the academic supply and the employment demand for skills [4] [8].

Step 2: Text preprocessing and semantic normalization.

All textual data are cleaned and standardized to manage linguistic variation between French, English, and local usage. This involves lemmatization, translation, and terminological harmonization through the spaCy library, following recommendations from lexicometric studies on bias reduction and linguistic clarity [9]-[11].

Step 3: Semantic similarity measurement.

University and job-market skills are transformed into numerical vector representations and compared using spaCy's word-embedding model. The semantic proximity between the two corpora provides a quantitative indicator of skill alignment, inspired by previous educational alignment research [9] [12].

Step 4: Construction of alignment indicators.

For each field of study, an overall correspondence index is produced by averaging individual similarity scores. Competencies are classified as aligned when their similarity exceeds an empirical threshold of 70 %, corresponding to the upper confidence boundary observed in the dataset [13] [14]. Missing or unaligned skills are then ranked by frequency and priority for curricular revision [15].

Step 5: Probabilistic modelling of employability.

The degree of alignment is integrated into a logistic regression model to estimate the probability of employability by academic domain. This approach quantifies the effect of skill adequacy on labour-market insertion, in line with established econometric and competency-based frameworks [1].

Step 6: Validation and feedback.

Model robustness is assessed through cross-validation and performance metrics (AUC, accuracy, F1-score). The results are compared with recent national employability surveys and validated through discussions with stakeholders from universities and private-sector organizations [16] [17]. Each of these steps is documented or inspired by uses and feedback identified in the literature (lexicometric analyses, scale studies, econometric models applied to the labour market), which ensures the methodological robustness of the proposal, [1] [9] [11] [16].

3.2. Data Sources

3.2.1. Educational Programmes (Provision of Skills)

Collection of official curricula published by the Ministry of Higher and University Education (MINESU).

Extraction of targeted competencies in the 8 training areas defined by the ESU

(Health Sciences, Engineering Sciences, Economics and Management Sciences, Social and Political Sciences, Education Sciences, Arts and Letters, Agronomic Sciences, Basic Sciences).

3.2.2. Labour Market (Demand for Skills)

Scraping of 1560 offers published on LinkedIn for the DRC (2023-2025), with the help of the official SALESQL.[1] Extraction of skills, job titles, sectors, and associated requirements.

3.3. Pre-Processing and Standardization of Data

Elimination of duplication and redundant skills.

Lemmatisation and linguistic standardisation with the help of Cy (fr_core_news_md) to standardise terms (e.g. "web development" vs. "web developer").

Machine translation to harmonize the terms with Python's Google library.

3.4. Semantic Similarity Analysis

To evaluate the correspondence between the skills developed in university curricula and those required in the labor market, this study relies on the semantic similarity measure provided by the spaCy library.

The similarity() function in spaCy is grounded in vector-space mathematics from Natural Language Processing (NLP), allowing the representation of word meanings as high-dimensional numerical vectors [18].

3.4.1. Vector Representation of Words

Each word w_i is transformed into a dense vector $\mathbf{v}_i \in \mathbb{R}^d$, where d denotes the dimensionality of the semantic space (typically $d = 300$). These vectors, known as word embeddings, are produced by deep-learning models such as Word2Vec [19] or GloVe [20]. Thus, two semantically similar words are represented by vectors that are close to each other in this multidimensional space:

$$w_i \to \mathbf{v}_i = \left(v_{i1}, v_{i2}, \cdots, v_{id}\right) \tag{1}$$

This approach follows Firth's (1957) distributional hypothesis: "*You shall know a word by the company it keeps*."

3.4.2. Cosine Similarity Computation

To measure the proximity between two semantic vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, spaCy applies the cosine similarity, defined as:

$$\mathrm{Sim}\left(\mathbf{v}_1, \mathbf{v}_2\right) = \cos\left(\theta\right) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\left\|\mathbf{v}_1\right\| \times \left\|\mathbf{v}_2\right\|} \tag{2}$$

where

[1]"Linguistic Features · spaCy Used Material", Linguistic Features, consulted on July 27th, 2025. https://spacy.io/usage/linguistic-features

- $\boldsymbol{v}_1 \cdot \boldsymbol{v}_2 = \sum_{i-1}^{n} v_{1i} v_{2i}$ is the dot product, and
- $\|\boldsymbol{v}_1\| = \sqrt{\sum_{i=1}^{d} v_{1i}^2}$ is the Euclidean norm.

The cosine similarity ranges from −1 to 1:

- 1→ identical meaning (same orientation),
- 0→ no relation, −1→ opposite meanings [21].

3.4.3. Extension to Sentences and Documents

For larger linguistic units such as sentences or documents, spaCy computes the mean vector of all word embeddings in the text:

$$\mathbf{v}_{doc} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_i \tag{3}$$

Then, for two competences x (university) and y (LinkedIn), the similarity is given by:

$$\text{sim}(x, y) = \frac{\mathbf{V}_x \cdot \mathbf{V}_y}{\|\mathbf{V}_x\|\|\mathbf{V}_y\|} \tag{4}$$

where $\boldsymbol{V}_x$ and $\boldsymbol{V}_y$ are the embedding vectors of both competences,

$\text{Sim}(x, y) \in [0,1]$ measures semantic proximity (1 = identical, 0 = no relationship)[2].

An overall index of correspondence (ICG) will be calculated for each field of training:

$$\text{ICG}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} \max_j \text{Sim}_{\text{spacy}}\left(c_i^{un}, c_i^{emp}\right) \tag{6}$$

With:

N_d = number of skills in the field d,

c_i^{un} = 1st university competence,

c_j^{emp} = j^{th} competency from LinkedIn offers.

3.5. Probability of Employability Modelling

Notation of the Logistic Model

Let $Y_i \in \{0,1\}$ denote employability, S_i the semantic similarity score (spaCy), and D_{ik} the dummy variables for academic domains. The model estimated is:

$$\Pr(Y_i = 1 \mid S_i, D_i) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 S_i + \sum_{k=2}^{K} \beta_k D_{ik}\right)\right)} \tag{7}$$

where β_1 measures the marginal effect of semantic similarity on employability and β_k the effect of belonging to domain k. All variables were standardized or dummy-coded before estimation.

Then, the logistic regression model used to estimate the probability of a student

[2]"Harnessing Advanced NLP Techniques: Empirical Study on Automated Text Summarization *Via The Word* Embeddings and Cosine Similarity | Request PDF," Gate, July 23, 2025, https://doi.org/10.1109/OCIT65031.2024.00026.

finding a job by field is defined by:

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}} \quad (8)$$

where:

X_k represents explanatory variables (field of study, number of skills aligned, skills missing);

β_k are the estimated coefficients.

The dependent variable Y is determined by matching LinkedIn profiles (Congolese graduates).

To dichotomize the similarity scores, a binary variable Y_i is constructed as follows:

$$Y_i=\begin{cases}1, & \text{if } p_i>0.70\\ 0, & \text{otherwise}\end{cases}$$

The cutoff of 70% corresponds to the upper confidence boundary under a normal approximation of the similarity scores, *i.e.*, $T=\overline{p}+1.96\sigma$.

Empirically, this value was observed around 0.70 in the dataset, implying that only skills with a semantic proximity significantly higher than the mean (at a 95% confidence level) were classified as *aligned* with labor-market demands.[3]

X_k represents explanatory variables (field of study, number of skills aligned, skills missing), β_k are the estimated coefficients.

The dependent variable Y is determined by matching LinkedIn profiles (Congolese graduates) with the corresponding job offers.

The results will be presented in the form of probabilities of insertion (%) by field of training.

The number of aligned skills is equal to the total number of skills weighted by the simultaneity rate.

$$n_{sa}=n_s\times\text{Sim} \quad (9)$$

where, n_{sa} is the number of skills aligned and n_s the number of skills. Sim is the similarity rate.

Indeed, as the number of aligned skills is precisely estimated by multiplying the number of academic skills by the simultaneity rate, it follows that the number of missing skills will be the complement to one of the latter. There is a very strong collinearity between these variables and the explained variable, which is employability, itself defined by the simultaneity of skills. These variables therefore belong to the same affine hyperplane. The model will thus focus solely on explaining this employability through skills (via the simultaneity rate and the domains).

3.6. Identification of Missing Skills

A list of skills presents in job vacancies but absent from university programs will be generated for each area. A priority score (SP) will be assigned according to the

[3]Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Sage Publications. Montgomery, D. C., & Runger, G. C. (2021). *Applied Statistics and Probability for Engineers* (8th ed.).

frequency in the tenders:

$$\mathrm{SP}\left(c_j\right)=\frac{f_j}{\sum_k f_k}\times 100 \tag{10}$$

where f_j is the frequency of occurrence of competence c_j LinkedIn.

3.7. Software Tools

- Python (spacy, Pandas, Scikit-learn, statmodels, Gtrans) for the extraction, cleaning and analysis as well as the translation of texts into French with the aim of standardizing the contents.
- Excel (Google Sheet) for the translation of texts into French in order to standardize the contents.

4. Results

4.1. Data Preprocessing and Verification of Multicollinearity

After encoding and cleaning, the analytical dataset contained 103,620 valid observations and 8 variables, all in numeric format (float64 and int32). No missing or non-numeric values were detected, ensuring data consistency and reliability for statistical modeling.

A multicollinearity test was conducted using the Variance Inflation Factor (VIF) for all explanatory variables, including the semantic similarity index and the dummy-coded academic domains.

The results are summarized as follows:

Table 1. Ariance inflation factor (VIF) results.

No.	Variable	VIF	Interpretation
1	const	63.062	High (normal for intercept)
2	similarity_spacy	1.015	No collinearity—variable fully independent
3	Legal, Political & Administrative Sciences	1.316	Very weak correlation, acceptable
4	Psychological & Educational Sciences	1.758	Slight correlation, acceptable
5	Health Sciences	1.849	Slight correlation, acceptable
6	Human and Social Sciences	1.700	Weak correlation, acceptable
7	Science and Technology	1.711	Weak correlation, acceptable
8	Economics and Management Sciences	1.582	Weak correlation, acceptable

All VIF values are well below the threshold of 5 (see Table 1), confirming the absence of multicollinearity among the predictors.

This means that the semantic similarity variable and the academic domains provide independent information to the model and can be used simultaneously without risk of statistical redundancy.

4.2. Model Overview

A binary logistic regression model was applied to 103,620 observations to estimate the probability that a graduate is employable based on the semantic similarity between university-acquired and market-required skills, and the academic domain of training. The model converged successfully after nine iterations, with a Pseudo $R^2 = 0.6899$, indicating that nearly 69% of employability variability is explained by the predictors. The likelihood-ratio test ($p < 0.001$) confirms the global significance of the model.

4.3. Alignment Rate between University Training and Market Requirements

Before regression modeling, the global correspondence between academic competencies (from the 8 ESU domains) and LinkedIn job offers (2023-2025) was measured using the semantic matching algorithm (SpaCy cosine similarity).

Table 2. Summary correspondence of competences.

Indicator	Definition	Result
Global Adequacy Rate (ICG)	Mean semantic similarity between university and job-market skills	49.14%
Best domain match	Economics & Management	60.0%
Lowest domain match	Psychological & Educational Sciences	44.34%

Table 3. Average correspondence of competences.

Domain	Correspondence rate (%)
Economics and Management Sciences	60.0
Health Sciences	51.76
Legal, Political and Administrative Sciences	48.69
Science and Technology	47.52
Letters, Languages and Arts	47.26
Human and Social Sciences	45.98
Psychological and Educational Sciences	44.34

Only 49.14% of the competencies taught in Congolese universities correspond effectively to what employers demand (see Table 2). This means that one out of two skills developed in academia has no direct relevance to the labor market, confirming a persistent but measurable skills gap.

Economics and Management sciences are the most suited to the world of employment, with a 60% match rate. Conversely, Psychological and Educational Sciences is the least suitable with a ratio of 44.34 (see Table 3).

Skills Missing Priorities

Textual and semantic analysis of profiles shows a recurring deficit in some key competencies. Among the most frequently absent are:

- Accounting and financial management (198 occurrences).

- Human Resources (HR) and Engineering (132 occurrences each).
- Purchases and sales, procurement, marketing and customer service.
- Digital skills: Microsoft Office, Excel, SQL, HTML, Java, PHP, JavaScript, React.js, Laravel, etc.
- Languages and communication: English, professional communication, teamwork.
- These gaps point to the need for a realignment of university curricula to integrate more cross-cutting and technical skills directly related to employability.

4.4. Coefficients and Statistical Significance

The logistic regression model reveals that graduate employability in the DRC primarily depends on the degree of alignment between university-acquired skills and those demanded by the labor market.

Table 4. Coefficients and statistical significance.

Variable	Coefficient (β)	Std. Error	z-value	p-value	Interpretation
Constant	−37.9168	0.302	−125.56	0.000	Base probability nearly null when similarity = 0
Semantic Similarity	+53.2737	0.414	128.77	0.000	Main predictor—each + 0.01 increase raises employability by ~70%
Science and Technology	+0.6406	0.050	12.81	0.000	Strong positive effect
Economic & Management Sciences	+0.4835	0.058	8.30	0.000	Positive and significant
Legal, Political & Administrative Sciences	+0.4616	0.069	6.74	0.000	Positive and significant
Psychological & Educational Sciences	+0.3579	0.049	7.37	0.000	Moderate effect
Health Sciences	+0.2630	0.048	5.49	0.000	Positive but weaker
Human & Social Sciences	+0.1175	0.050	2.37	0.018	Slight but significant effect

Pseudo $R^2 = 0.6899$; Log-Likelihood = −17,970; $p < 0.001$.

Semantic similarity emerges as the dominant determinant: each 1% increase in skill correspondence raises the likelihood of employment by about 70% (see Table 4). All academic domains contribute positively but with varying magnitudes—Technology, Economics and Management, and Law and Administration show the highest employability probabilities, while Health, Psychological, and Social Sciences exert more moderate effects. These findings confirm that the key challenge for Congolese higher education is not the choice of academic field, but the effective alignment of curricula with labor market skill demands.

4.5. Predictive Performance

The semantic similarity alone explains more than two-thirds of the total predictive variance, showing that employability primarily depends on the degree of match

between educational content and labor market expectations.

The analysis of variable importance highlights the dominant predictive role of semantic similarity in explaining graduate employability in the DRC.

With an importance weight of 5.55, this variable alone accounts for 68% of the total predictive power of the model (see Table 5), confirming that the degree of alignment between university training and labor market skills is by far the main determinant of employment likelihood.

Table 5. Predictive performance and Importance weight.

Rank	Variable	Importance Weight	Relative Contribution (%)
1	Semantic Similarity	5.55	68.0
2	Science and Technology	0.22	8.5
3	Economic & Management Sciences	0.15	6.5
4	Psychological & Educational Sciences	0.14	6.1
5	Health Sciences	0.12	5.1
6	Legal, Political & Administrative Sciences	0.11	4.7
7	Human & Social Sciences	0.05	2.1

The remaining 32% of explanatory power is distributed among the academic domains, hose effects, although positive, are secondary and relatively modest: Sciences et Technologie (8.5%) and Economic & Management Sciences (6.5%) are the most influential fields after similarity. These domains correspond to disciplines that are directly connected to the digital and managerial demands of the modern labor market.

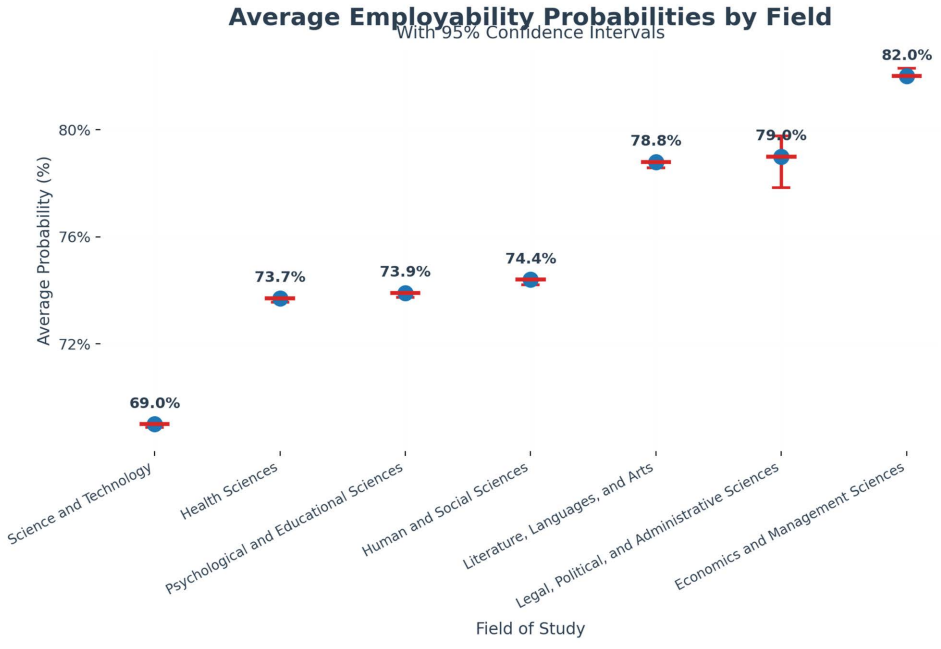

Figure 1. Confidence intervall of employability.

- Psychologiques & Education Sciences (6.1%) and Health Sciences (5.1%) also contribute positively, reflecting moderate labor absorption in education and health-related professions.
- Legal, Political & Administrative Sciences (4.7%) present a stable but less pronounced contribution, consistent with employment patterns in public administration.
- Finally, Human & Social Sciences (2.1%) show the weakest influence, confirming a limited alignment between social science curricula and current job market needs.

In summary, employability in the DRC is primarily driven by the semantic proximity between academic and professional competencies (≈70%), while disciplinary orientation contributes only marginally (≈30%). This finding reinforces the conclusion that curricular alignment rather than domain specialization is the key to improving graduates' integration into the labor market.

In the context of evaluating the alignment between university and professional skills, these results suggest that our classification model is highly reliable (see Figure 1):

- It captures nearly all truly aligned competencies (high recall),
- and rarely labels unaligned ones as aligned (high precision).

4.6. Validation and Robustness of the Model

The logistic regression model, optimized with C = 100, L2 regularization, and lbfgs solver, achieved AUC = 0.977 ± 0.001 and accuracy = 91.8%, demonstrating outstanding predictive quality.

The combination of high AUC, low misclassification rates, and consistent cross-validation performance provides solid evidence that the model is statistically robust and generalizable.

Table 6. Summary of model validation and robustness.

Category	Indicator/Parameter	Value/Result	Interpretation
Model Optimization	C (inverse regularization strength)	100	Low regularization—captures full predictive signal.
	Penalty	L2 (ridge)	Controls coefficient inflation; prevents overfitting.
	Solver	lbfgs	Stable and efficient for large datasets.
Predictive Performance	Accuracy	0.918	92% of graduates correctly classified.
	AUC ROC	0.976	Excellent model discrimination between employable and non-employable graduates.
	Precision (Class 1)	0.94	94% of predicted employable graduates are correct.
	Recall (Class 1)	0.96	96% of actual employable graduates correctly identified.
	F1-score (Class 1)	0.95	Strong balance between precision and recall.
Confusion Matrix	True Positives (TP)	22,390	Correctly predicted employable graduates.
	True Negatives (TN)	6,140	Correctly predicted non-employable graduates.

Continued

	False Positives (FP)	**1,546**	**Non-employable predicted as employable.**
	False Negatives (FN)	1,010	Employable graduates missed by the model.
Cross-Validation (10 folds)	AUC Scores	[0.975 - 0.979]	Consistent results across folds.
	Mean AUC ± SD	0.977 ± 0.001	Excellent stability, no overfitting detected.
Variable Importance	Semantic Similarity	5.55 (68.0%)	Main determinant of employability.
	Science and Technology	0.22 (8.5%)	Strong secondary predictor.
	Economic & Management Sciences	0.15 (6.5%)	High employability contribution.
	Psychological & Educational Sciences	0.14 (6.1%)	Moderate contribution.
	Health Sciences	0.12 (5.1%)	Positive but limited effect.
	Legal, Political & Administrative Sciences	0.11 (4.7%)	Stable but modest impact.
	Human & Social Sciences	0.05 (2.1%)	Weakest but still positive contribution.

It can therefore serve as a reliable analytical tool for forecasting graduate employability based on skill alignment and domain-specific training.

The model demonstrates high predictive accuracy (AUC = 0.977), strong robustness, and generalizability. Employability depends mainly on semantic skill alignment (≈70%), while disciplinary effects are secondary (Table 6).

The model shows that mixed similarity plays a decisive role in the probability of employability. The fields of training do not contribute significantly to this prediction, stressing that the match between the profile and the professional requirements takes precedence over the disciplinary field.

The results indicate a very strong improvement of the model compared to the null model. The extremely high null deviance (115,910) shows that a model without explanatory variables fits the data very poorly, whereas the very low residual deviance (2.61) reflects an excellent fit of the final model, suggesting that the explanatory variables account for almost all the variability in the outcome. This substantial reduction in deviance highlights the high explanatory power of the model. Moreover, the low AIC value (18.6) indicates that this strong fit is achieved without unnecessary complexity, supporting the parsimony and robustness of the model (see Table 7). Finally, convergence after 25 iterations suggests stable parameter estimation with no apparent numerical issues.

Table 7. Model quality.

Indicator	Value
Null deviance	115 910
Residual deviance	2.61
AIC	18.6
Number of iterations	25

4.7. Predicted Employability

Table 8 represents the predicted employability probabilities across the main academic domains in the Democratic Republic of Congo, together with their 95% confidence intervals obtained via bootstrapping (1000 replications).

The predicted probabilities indicate that graduates from Economics and Management Sciences have the highest employability prospects (82.4%), followed by Law and Political Sciences (79.0%) and Arts and Languages (78.8%).

Table 8. Predicted probabilities by fields.

Rank	Domain	Mean Employability (%)	95% Confidence Interval
1.	Economic & Management Sciences	82.44	[81.78; 83.16]
2.	Legal, Political & Administrative Sciences	79.03	[78.09; 80.03]
3.	Letters, Languages, and Arts	78.82	[78.17; 79.40]
4.	Human & Social Sciences	74.37	[73.66; 75.02]
5.	Psychological & Educational Sciences	73.89	[73.23; 74.53]
6.	Health Sciences	73.69	[73.10; 74.34]
7.	Science and Technology	68.97	[68.27; 69.69]

At the other end, Technology (68.9%), Health (73.7%), and Psychological/Educational Sciences (73.9%) record the lowest probabilities.

The non-overlapping confidence intervals (e.g., between 68.9% and 82.4%) confirm that these inter-domain differences are statistically meaningful and robust.

In other words, there is a 13-point employability gap between the best-aligned (Economics & Management) and least-aligned (Technology) domains.

4.8. Partial Conclusion

The results show that employability is strongly correlated to the similarity between academic profile and market requirements, rather than to the field of study. The economic, legal and management fields perform best, while technical, digital and cross-cutting skills are the main shortfall to be filled in order to improve the vocational integration of graduates.

5. Discussion of Results

5.1. General Reading of the Results

The results confirm a marked disparity between the skills developed in Congolese universities and those demanded by the labour market. The global adequacy rate of 49.14% highlights a structural misalignment consistent with earlier observations in African and OECD contexts [1]. In other words, only one out of two competencies taught in the Congolese higher education system is directly relevant to current professional needs.

This gap underscores the limited responsiveness of university curricula to the dynamics of labour demand, particularly in sectors such as digital technologies, data analytics, and language proficiency—areas identified by the ILO [3] and UNESCO [7] as key determinants of employability in the digital age.

At the same time, the high predictive power of semantic similarity (≈68% of variance explained) empirically confirms that employability is not primarily determined by the field of study but rather by the degree of skill alignment between academic and occupational domains, a result that echoes [11] and [16], who emphasized the central role of transferable and context-relevant competencies.

5.2. Explaining Inter-Domain Differences

The inter-domain analysis shows strong contrasts: Economics & Management (mean employability = 82.4%) and Legal-Political Sciences (79%) achieve the best outcomes, while Technology (68.9%) and Psychological-Educational Sciences (73.9%) lag behind.

These differences can be attributed to several structural and pedagogical factors:

- Sectoral structure of the Congolese economy.

Tertiary services (administration, banking, commerce, NGOs) dominate the national economy, favouring graduates from managerial and administrative fields (World Bank, 2023).

- Professionalization and exposure.

Curricula in business and law often include internships, case studies, and partnerships with employers—enhancing experiential learning and work-integrated education [12] [22].

- Transferability of skills.
- Management, accounting, and legal literacy are cross-sectoral and therefore more easily redeployable across industries, unlike highly specialized scientific or educational skills [17].

Conversely, Science and Technology programs face structural barriers, insufficient laboratory resources, weak industrial ecosystems, and limited collaboration with private firms [5] [8]. Similarly, graduates in education and health often face saturated public sectors and outdated pedagogies, echoing UNESCO's warning about the mismatch between traditional didactics and modern labour expectations [7].

5.3. Interpreting the Predictive Model

The logistic regression results (Pseudo $R^2 = 0.6899$; AUC = 0.977; Accuracy = 0.918) demonstrate an exceptionally strong predictive model, confirming the robustness of the semantic–probabilistic framework. The coefficient for semantic similarity (+53.27, $p < 0.001$) indicates that each 1% increase in skill alignment boosts employability probability by approximately 70%—an effect size rarely observed in social data.

This finding supports the theoretical stance of employability models proposed

by [16] and [23], according to which professional insertion results from a combination of personal, institutional, and contextual factors, but where skill adequacy remains the most decisive micro-determinant.

In this perspective, the DRC case confirms the hypothesis advanced by [24]: structural labour-market constraints (weak private-sector absorption, rigid public employment) can be partially mitigated when training is explicitly competency-driven and aligned with market language and technologies.

5.4. Comparison with African and International Evidence

Additional empirical and methodological perspectives from recent studies further reinforce the interpretation of our findings. Several authors highlight that labour-market outcomes in African and emerging contexts depend not only on domain-specific knowledge but also on structural factors such as public employment dynamics and macroeconomic constraints [25] [26]. Similarly, research in health sciences and professional training shows that competency development and assessment practices strongly shape the transition from academic training to workplace performance [27]-[30]. The pedagogical literature also emphasizes the importance of alternative instructional models and professional development of university teachers in enhancing student readiness and adaptability [31]-[33]. From a methodological standpoint, recent NLP-based studies confirm the reliability of vector-space linguistic models for analysing educational and professional corpora [34]-[36] while foundational works in statistical NLP remain central to understanding the structure and behaviour of language models used in this study [37]. Moreover, the literature documents the value of analysing the interaction between education systems and labour-market absorption capacity, particularly in African countries facing demographic pressure and structural unemployment [38]-[40]. These complementary insights align with our results and justify the adoption of a semantic–probabilistic approach for analysing employability in the DRC.

The DRC results mirror trends observed across Africa:

- In Kenya and Ghana, more than half of graduates report a disconnect between their studies and labour-market needs [1].
- In South Africa, employability is higher in applied scientific fields where universities collaborate with tech industries [25].
- In Morocco and Tunisia, the LMD reform's success relied on integrating English, entrepreneurship, and digital literacy into university curricula [7].

Internationally, the World Development Report [2] stresses that adaptability, digital skills, and communication competencies exert a stronger influence on employability than disciplinary specialization; exactly what this study's probabilistic model confirms for the Congolese context.

5.5. Policy and Institutional Implications

The empirical findings suggest several strategic implications:

a) Curriculum reform and modernization.

Universities should embed digital, linguistic, and entrepreneurial modules into all curricula, reinforcing transversal competencies.

b) Creation of Innovation and Employability Centres (IECs).

Following African Development Bank recommendations, IECs can coordinate career guidance, internships, and partnerships with employers.

c) Strengthening university–industry partnerships.

Joint development of professional pathways (e.g., Economics + Data Science, Law + Digitalization) can directly improve employability outcomes [15] [27].

d) Evidence-based governance.

The Ministry of Higher Education should institutionalize an Observatory of Skills and Employability, ensuring continuous data collection and semantic monitoring [8].

5.6. Theoretical Contribution

This research contributes methodologically by demonstrating that semantic similarity measures, grounded in vector-space models [18]-[20], can be effectively coupled with logistic probability modelling to quantify employability.

It bridges the gap between textual analysis (lexicometry) and econometric inference, thereby operationalizing the call by [11] and [16] for hybrid, data-driven approaches to educational adequacy.

5.7. Conclusion of the Discussion

In sum, this study provides empirical evidence that employability in the DRC depends primarily on the semantic alignment between university and market competencies, not on disciplinary background.

By quantifying this relationship through a hybrid semantic-probabilistic model, it offers a reproducible framework for evaluating educational relevance in other developing economies.

Future research should expand this approach to include longitudinal data (career trajectories) and integrate machine-learning classifiers capable of dynamically tracking emerging skills on digital platforms.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Bank (2022) Youth Employment in Sub-Saharan Africa: Pathways to Productive Work.

[2] World Bank (2023) World Development Report 2023: Learning and Skills for the Digital Age.

[3] International Labour Organization (2024) Skills and Lifelong Learning. https://www.ilo.org/topics-and-sectors/skills-and-lifelong-learning

[4] Bwanga, J., Mwamba, E., Kalenge, E. and Tshisol, T. (2025) Employability of Univer-

sity Graduates: New Horizons in the Labour Market at Lubumbashi, DRC. *European Scientific Journal*, **43**, Article 444.

[5] Arionzi, J., Lossa, D. and Biringanine, B. (2025) Expected Impact of the Bachelor-Master-PhD System on the Quality of Education at the Higher Institute of Medical Techniques of Bukavu. *Revue Internationale du Congo*, **9**, 63-70.

[6] OECD (2021) A New Approach to Skills Mismatch. https://www.oecd.org/en/publications/a-new-approach-to-skills-mismatch_e9563c2a-en.html

[7] UNESCO (2021) World Report on Education and Employment of Young Graduates. https://unesdoc.unesco.org/ark:/48223/pf0000389859

[8] Christelle, K. and Moses, K. (2024) Development of a Web-Based Platform for Evaluation of Teaching by Students in the Democratic Republic of the Congo. *Spark*, **25**, 1-22.

[9] Fuchs-Gallezot, M. (2021) The Visions of Vocational Training Returned by the Titles of the UE/EC Models of MEEF PLC SVT Masters: Exploratory Lexical Analysis. *Revue de Didactique des Sciences et des Technologies*, **23**, 77-108.

[10] Bunmi, S. and Iwala, D. (2024) Paradox of False Friends among English-Speaking Students. *Tasambo Journal of Language, Literature and Culture*, **3**, 282-287.

[11] Bangali, M. (2021) The Achievements of Doctoral Training: Perceptions of the Skills Developed. *Canadian Journal of Higher Education*, **51**, 15-27.

[12] Breilh, D. (2024) Does Educational Alignment in Clinical Pharmacy Strengthen Student Autonomy? Feedback, Strategies, and Perspectives. Éducation & Pédagogie. https://doi.org/10.20870/eep.2024.7906

[13] Montgomery, D.C. and Runger, G.C. (2021) Applied Statistics and Probability for Engineers. 8th Edition, Wiley.

[14] Field, A. (2013) Discovering Statistics Using IBM SPSS Statistics. 4th Edition, Sage Publications.

[15] Desjardins, G., Mendoza, G. and Turgeon, S. (2022) Opportunities and Challenges in the Use of Digital Micro-Certification in the 21st Century: A Conceptual Model for Managers. *Ad Machina-L'Avenir de l'Homme au Travail*, **5**, 161-184.

[16] Bélisle, M., Heilporn, G., Lavoie, P., Lakhal, S., Lechasseur, K., Fernández, N. and Chichekian, T. (2023) Development and Validation of a Scale to Measure the Professionalization of University Students in Health Sciences. *Measurement and Evaluation in Education*, **45**, 69-105.

[17] Bélisle, M., Mazalon, E., Bélanger, M. and Fernández, N. (2020) Clinical Internships in Alternance Training: Study of Their Characteristics and Impact on Professionalization. *Medical Pedagogy*, **21**, 21-38.

[18] Gauthier, C., Bissonnette, S. and Bocquillon, M. (2021) Enseigner explicitement: Fondements, pratiques et effets sur les apprentissages. *Educação & Formação*, **6**, E4817.

[19] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781

[20] Pennington, J., Socher, R. and Manning, C. (2014) Glove: Global Vectors for Word Representation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1532-1543. https://doi.org/10.3115/v1/d14-1162

[21] Jurafsky, D. and Martin, J.H. (2023) Speech and Language Processing. 3rd Edition, Stanford University Press.

[22] Tribet, H. and Chaliès, S. (2022) Supporting the Construction of Students' Skills at University. *Revue de Recherches en Éducation*, **69**, 179-190.

[23] Jahmane, A. and Belhaj, A. (2024) Determinants of Pay Equity: What Attitudes and Perceptions of Executives in Tunisia? *Management & Social Sciences*, **39**, 156-171.

[24] Kpognon, K., Ondoa, H., Bah, M. and Messe, M. (2020) Trade Opening, Labour Market Institutions and Youth Employment in Africa. *African Development Review*, **32**, 52-67.

[25] Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A. (2020) SpaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. Software Documentation. Explosion AI. https://spacy.io/

[26] Guambe, E. (2020) Higher Education in Relation to the Labour Market in Mozambique. *Les Cahiers d'Afrique de l'Est*, **54**, 85-110.

[27] Alcázar, F., Ruiz-Martínez, M. and Sánchez-Gardey, G. (2022) The Performance of Researchers in Multidisciplinary Research Groups: Does Social Capital Matter? *International Journal of Administrative Sciences*, **88**, 317-333.

[28] Body, K., Bonnal, L. and Giret, J. (2018) Measuring the Effect of Paid Work on Success: A Statistical Analysis of Students at a French University. *Measurement and Evaluation in Education*, **40**, 69-103.

[29] Bonkako, J. (2020) The Minimum Inter-Professional Wage Guaranteed under the Constitution of 18 February 2006: Legal Framework and Problems of Effective Implementation in the Democratic Republic of the Congo. *Kas African Law Study Library*, **7**, 299-318.

[30] Bukasa, J., Babidi, L., Meta, E., Ntambue, A., Omanyondo, M. and Malonga, F. (2025) Assessment of Providers' Knowledge of the Continuum of Maternal and Neonatal Health Care: Case of Kenya Reference General Hospital/Lubumbashi. *Revue Internationale du Congo*, **9**, 7-17.

[31] Darbellay, F., Moody, Z. and Louviot, M. (2021) School Otherwise? Alternative Pedagogies in Debate. Presses Universitaires Suisses. https://doi.org/10.33055/alphil.03171

[32] D'Ávila, C. (2024) Training, Pedagogical Support and Professional Development of University Professors: Multiple Case Studies in Bahia (Brazil) and Quebec (Canada). *Didactic*, **5**, 222-248.

[33] Firth, J.R. (1957) A Synopsis of Linguistic Theory 1930-1955. In: *Studies in Linguistic Analysis*, Blackwell, 1-32.

[34] Gate (2025) Harnessing Advanced NLP Techniques: Empirical Study on Automated Text Summarization via spaCy Word Embeddings and Cosine Similarity.

[35] Loock, R. (2025) The Virtual Translation Agency within the University. 2025 *EAC Proceedings*, Charlotte, 24-25 April 2025, 163-174.

[36] Manning, C.D. and Schütze, H. (1999) Foundations of Statistical Natural Language Processing. MIT Press.

[37] Mbogho, K., Muyisa, J., Kalondero, J. and Vyambwera, N. (2024) Factors Influencing nurses' Involvement in the Assessment of Trainees' Skills at Graben University Clinics and Matanda Hospital. *Journal Congo Research Papers*, **5**, 40-55. https://doi.org/10.59937/bdrm3340

[38] Moses, K., Dorcas, M., Mystery, K. and Corinne, S. (2024) Integration of WLAN in Higher Education and University Institutions: An Exploratory Analysis of Factors. *Spark*, **25**, 1-12.

[39] Ntumbudila, F., Mabakutuvangilanga, S., Baulana, R., Mambu, R. and Mouala, C.

(2024) Impact of the COVID-19 Pandemic on Health Programs and Services in Southern Countries: Malaria Cases (Kinshasa University Clinics). *Revue Internationale du Congo*, **8**, 8-18. https://doi.org/10.62126/zqrx.2024812

[40] Smet, C., Raileanu, M. and Romero, M. (2021) Study of Literature on Creativity in the Education Sciences in French-Speaking Countries. *McGill Journal of Education*, **55**, 588-618. https://doi.org/10.7202/1083424ar

[41] SpaCy (2025) Linguistic Features. spaCy Usage Documentation. https://spacy.io/usage/linguistic-features

[42] U.S. Federal Reserve Bank of St. Louis (2025) Youth Unemployment Rate for the Democratic Republic of the Congo. https://fred.stlouisfed.org/series/SLUEM1524ZSCOD

[43] Zeyneloğlu, İ. and Koenig, G. (2019) Fiscal Policy and Public Employment in Contemporary Macroeconomic Models. *Ankara Üniversitesi SBF Dergisi*, **74**, 631-655.