International Journal of Innovative Technology and Interdisciplinary Sciences
https://journals.tultech.eu/index.php/ijitis
ISSN: 2613-7305
Volume 8, Issue 3
DOI: https://doi.org/10.15157/IJITIS.2025.8.3.595-623
Received: 08.08.2025; Revised: 09.09.2025; Accepted: 11.09.2025

Research Article

A Latent Dirichlet Allocation Framework to Analyse and Forecast Employability Skills

Milena Shehu[1*], Areti Stringa[1], Eralda Gjika Dhamo[2]

[1] Faculty of Economy, University of Tirana, Tirana, Albania
[2] Faculty of Computer Engineering and IT, Metropolitan University Tirana, Tirana, Albania
***milena.shehu@unitir.edu.al**

Abstract

Globalization, rapid technological advancement, Albania's EU integration process are reshaping labour market dynamics, creating urgent needs for timely skill intelligence. Traditional survey-based statistics often lag behind these changes, while online job postings provide a real-time source of employer demand. A Latent Dirichlet Allocation (LDA)-based framework is introduced in this paper, applied to 1,500 vacancies collected from five major Albanian job portals (July–September 2024), to extract, categorize, and forecast employability skills. The model is implemented in a rolling/windowed LDA Model, enabling the tracking of skill dynamics over time and alignment with the European Skills, Competences, and Occupations (ESCO) taxonomy. Findings show that Albanian employers predominantly demand transversal soft skills, especially Responsibility, Communication, Collaboration, Networking, and Presentation, while green and digital skills appear only gradually. An interactive "Skills Forecast" Shiny application operationalizes results, forecasting the top ten in-demand skills for specific vacancies that the user wants to test, and offering validation metrics for policymakers, educators, and employers. This study represents the first systematic application of topic modelling to Albania's labour market, providing a replicable, policy-relevant tool aligned with the National Employment and Skills Strategy 2023–2030 and the National Youth Strategy 2022–2029, while highlighting pathways for future integration of advanced NLP methods.

Keywords: Labour Market Intelligence; Skills Forecast; Latent Dirichlet Allocation; Topic Modelling; Natural Language Processing

INTRODUCTION

In recent years, migration flows, demographic shifts, and post-transition economic reforms have profoundly reshaped Albania's socio-economic landscape, with direct implications for the labour market [1, 2]. Shifts in skill demand across economic sectors highlight the need for policymakers, educators, and businesses to anticipate workforce requirements more effectively [3]. Traditional analyses of Albania's labour market have largely focused on macroeconomic trends, migration, and unemployment [4, 5], while less attention has been paid to employability skills and their evolution. Addressing this gap,

the present study applies Latent Dirichlet Allocation (LDA) algorithm to analyse and forecast soft skills and interests demanded in Albania's labour market.

Latent Dirichlet Allocation (LDA), introduced in 2003 [6], is a probabilistic topic-modelling algorithm designed to uncover latent semantic structures in large textual datasets. In this study, LDA is applied to online job vacancies dataset to extract the most relevant soft skills, classify them into categories, and identify underlying patterns that are not easily captured through conventional methods. Given the increasing complexity of labour market dynamics, predictive modelling approaches such as LDA provide valuable tools for generating timely labour market intelligence [6, 7].

Albania faces a persistent skills gap: the World Bank reports that the supply of graduates does not meet employer demand [1], while the International Labour Organization notes that these mismatches slow economic growth and exacerbate youth unemployment [2]. Existing research has begun to address employability in Albania. Fejzulla [4] stresses the importance of aligning private-sector needs with vocational education and training, authors at [5] statistically analyse the role of hard and soft skills and examine youth employment policies. More recently, authors at [8-11] reviewed topic-modelling approaches, highlighting their potential for analysing skills in the Albanian context. The analysis of labour market demand through online job vacancies has emerged as a critical research frontier, enabling timely insights into the skills employers require and how these evolve over time. Traditional labour market information systems, often based on surveys and administrative records, tend to lag behind real-time changes and may fail to capture emerging occupations or skills. By contrast, natural language processing (NLP) and machine learning methods applied to job postings provide scalable ways to extract and monitor skill demand at high frequency [8, 9].

Among topic modelling approaches, LDA remains one of the most widely applied unsupervised algorithms, due to its probabilistic structure and intuitive interpretability [12-14]. LDA has been employed extensively to extract latent themes from large corpora, including studies on employability and workforce development [14]. However, more recent advances in topic modelling have highlighted important limitations of LDA, particularly with respect to coherence and semantic richness [15-18]. Dynamic Topic Models (DTM) [15-17] extend LDA by explicitly modelling how topics evolve across time slices, making them highly suitable for forecasting skill dynamics in rapidly changing labour markets. In parallel, transformer-based methods such as BERTopic [19] leverage contextual embeddings from BERT to produce more coherent and semantically meaningful topics, often outperforming LDA in text mining tasks [19]. In the domain of labour market and employability research, several recent contributions have applied these innovations. Chiarello et al. [3] developed the "ESCO 4.0" approach, using text mining to align the European Skills, Competences, Qualifications and Occupations (ESCO) taxonomy with Industry 4.0 requirements. Their work illustrates how modern topic modelling can support EU-wide policy alignment. Similarly, El Sharkawy et al. [20] applied supervised machine learning to predict employability for IT graduates, demonstrating how labelled datasets

can inform targeted employability strategies. These studies highlight the methodological progress beyond classical LDA, as well as the potential of NLP algorithms for policymaking. Against this background, our contribution is not the introduction of a novel algorithmic technique, but rather the contextual adaptation and operationalization of an LDA based framework for the Albanian labour market, a context where applications of NLP to skill analysis are scarce. Specifically, we demonstrate how LDA can be combined with ESCO-aligned skill taxonomies and implemented in a rolling/windowed format to track skill dynamics over time. This enables the extraction of employability relevant skills from job postings and generates outputs that are directly actionable for policy. By doing so, the study addresses a significant gap in the Albanian and Western Balkans context, providing a methodological foundation for labour market intelligence that can inform education, training, and employment strategies.

Finally, this framework has direct policy relevance for Albania, as it supports the objectives outlined in the National Employment and Skills Strategy 2023–2030 [21], which emphasizes strengthening transversal, digital, and green skills to improve competitiveness and reduce skills mismatch. It also aligns with the National Youth Strategy 2022–2029 [22], which prioritizes employability, entrepreneurship, and skill development for young people entering the labour market. Furthermore, by integrating the ESCO classification, the framework contributes to EU integration efforts, ensuring that Albania's labour market monitoring tools are harmonized with European standards and methodologies.

THE LATENT DIRICHLET ALLOCATION (LDA) MODEL

Latent Dirichlet Allocation (LDA) is among the most widely applied unsupervised algorithms in the field of topic modelling. Its application to job vacancy data builds on the assumption that each vacancy description constitutes a document whose content can be represented as a mixture of latent topics. Each topic, in turn, is expressed as a probability distribution over words [6]. This probabilistic framework makes LDA particularly suitable for analysing large corpora of unstructured labour market texts, such as online job postings.

Generative Process

Formally, LDA models a corpus D consisting of M documents (job vacancies): we have D = [D1, D2, … , D_M], where each document Dd is constructed from a dictionary of LW unique terms: W = [w1, w2, … , wL_W]. The model seeks to uncover LT latent topics: T = [T1, T2, … , TL_T]. The generative process can be summarized as follows [6-8]:

1. For each topic Tt draw a word distribution $\phi t \sim Dir(\beta)$.
2. For each document Dd, draw a topic distribution $\theta d \sim Dir(\alpha)$.
3. For each word wn in document Dd:

- Select a topic $Tn \sim Mult(\theta d)$.
- Select a word $wn \sim Mult(\phi Tn)$.

Here, words are observable variables, while ϕ and θ are hidden variables. The hyperparameters α and β regulate the sparsity of document–topic and topic–word distributions, respectively. The probability of the observed corpus D is computed by marginalizing over these latent variables.

Model Outputs

The estimation produces two probability matrices, see equation (1) [6-8]:

$$\phi = [\phi t,w]LT \times LW , \quad \theta = [\theta d,t]M \times LT \tag{1}$$

- Φ - distribution of words across topics, characterizing the semantic content of topics.
- Θ - distribution of topics across documents, indicating the mixture of skills in each job vacancy.

Through LDA, words that frequently co-occur across vacancies are grouped into coherent topics, which can be interpreted as sets of employability skills. Importantly, terms are not uniquely assigned to topics but may contribute to multiple topics with different probabilities, reflecting the overlapping skill requirements of real-world job postings.

Mapping Topics to Skills

To map identified topics to specific skill categories, we define the set of categories as: C = [C1, C2, … , CLC]. Each LDA topic is compared to these categories using cosine similarity [11], producing a similarity matrix, see equation (2):

$$\Delta[\delta t,c]LT \times LC \tag{2}$$

This matrix quantifies the degree of association between topics and predefined skill groups, enabling the classification of job vacancies according to their underlying skill demands.

Temporal Adaptation with Rolling LDA

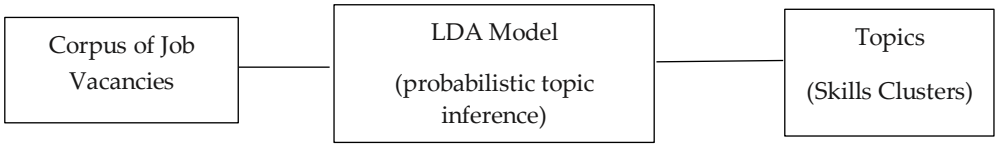
While dynamic LDA exists in its formal Bayesian state-space formulation, here we adapt a rolling/windowed LDA approach with topic alignment across time windows [23,24]. The dataset of job vacancies is segmented into periods (e.g., quarterly subsets), and an independent LDA model is estimated for each. This rolling approach provides a practical means of tracking emerging and declining skills in relatively sparse datasets, capturing localized demand patterns. Topic alignment across windows ensures comparability over time, thereby revealing how employers' requirements evolve in response to technological, economic, or policy changes.

This temporal adaptation extends the static model by allowing the observation of trajectories of skill demand, for example, the persistence of transversal skills such as Communication and Responsibility, or the gradual emergence of Green and Digital skills.

Figure 1 illustrates the conceptual difference between static LDA, which extracts stable skill clusters without temporal information, and rolling LDA, which incorporates time slices to capture evolution in skill demand.

- Static LDA: treats all job postings as a single dataset, extracting stable skill clusters.
- Rolling/Windowed LDA: segments postings into time periods and links topic distributions across time, showing how skills emerge, grow, or decline.

Static LDA Workflow:



Rolling/Windowed LDA Workflow:

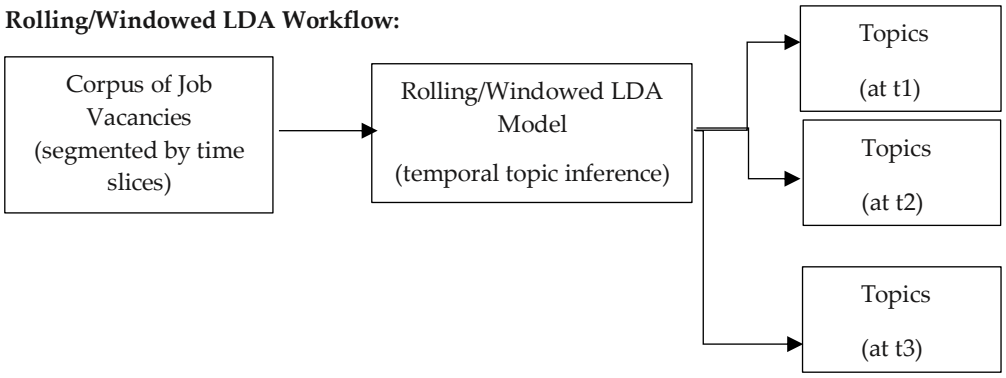


Figure 1. Static vs. Rolling/Windowed LDA Workflow, implementation from [23]

Innovation for the Albanian Context

The framework is built on the classical LDA model, which is a well-established technique in topic modelling. Our contribution is not algorithmic novelty, but rather the adaptation and operationalization of LDA for a labour market context where such applications are scarce [24-28]. Specifically, the innovations for the Albanian context lie in three areas:

1. Contextual adaptation: To our knowledge, this is the first systematic application of topic modelling to the Albanian labour market. By aligning outputs with the ESCO taxonomy, the framework produces results that are directly interpretable for employability policies and workforce strategies.
2. Temporal adaptation: Instead of static modelling, we employ rolling LDA with topic alignment across windows, enabling the practical tracking of skill dynamics over time.
3. Policy relevance: Unlike prior methodological studies [1, 26], our framework integrates unsupervised NLP with ESCO classification to generate outputs that are actionable for policy and practice in a developing labour market.

By combining these elements, the study advances labour market intelligence in Albania and the wider Western Balkans, providing policymakers and education providers with a

data-driven tool to anticipate skill needs, align curricula with future demand, and design more responsive employment strategies.

We do not claim algorithmic novelty, since LDA and its rolling variants have been established in the literature [6, 23]. Instead, our contribution lies in contextual and operational innovation: (i) integrating ESCO taxonomies with LDA outputs to ensure direct policy relevance, (ii) developing a rolling/windowed implementation suitable for relatively sparse datasets typical of smaller labour markets, and (iii) deploying results in an interactive Shiny application that functions as a practical policy toolkit. This combination provides a replicable framework for non-EU contexts such as Albania, where systematic NLP-based labour market intelligence is still largely absent.

MATERIALS AND METHODS

In this study, the Latent Dirichlet Allocation (LDA) algorithm was applied to extract and group employability-related skills from job vacancy descriptions, subsequently aligning them with the ESCO classification codes for skills and competences. For each skill category, the LDA model produces a list of representative terms, which are then reviewed, validated, and refined by labour market experts. As LDA is an unsupervised learning method, expert supervision is indispensable to ensure that the outputs are both reliable and policy relevant. To operationalize this approach, we designed a structured sequence of steps that form the basis of the LDA-based framework, as illustrated in the workflow, see Figure 2.

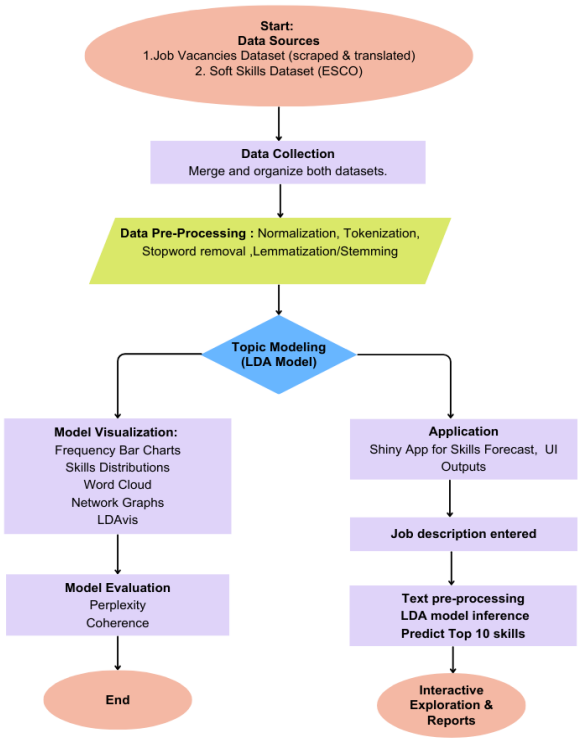


Figure 2. Workflow of the LDA-based Framework

Data Collection

This study employed two datasets: one comprising online job vacancies and another containing soft skills organized into predefined categories.

The first dataset consisted of 1,500 job vacancies collected from five major private job portals in Albania between July and September 2024. Data collection was performed using Python-based web scraping scripts, customized for the structure of each portal and executed in the Google Collab environment to produce dataset with job vacancies in CSV format. The resulting datasets were subsequently merged into a single corpus containing the following fields: job title, job category, date of scraping, job description, job qualifications, and the URL of the posting. The dataset provides a snapshot of labour demand in Albania during the study period, allowing us to test and operationalize the LDA-based framework.

Due to the limited timeframe, multi-year statistical trend analysis was not feasible. Nevertheless, to capture temporal variation, the analysis incorporated a rolling/windowed LDA design, which enables tracking of emerging and declining skills across shorter time intervals. Future research, based on larger longitudinal datasets, will expand this approach to test temporal dynamics and conduct hypothesis-driven validation of skill trends. The dataset captures postings across multiple economic sectors, with representation in services, information technology, administration, construction, and trade. However, the coverage is not exhaustive. Vacancies advertised through public employment services or informal channels were excluded, and international platforms such as LinkedIn were deliberately omitted. This decision reflects both technical and methodological constraints: (i) LinkedIn imposes restrictions on large-scale automated scraping, limiting reproducibility; and (ii) many LinkedIn postings target international rather than Albanian labour markets, which would reduce the dataset's representativeness. Focusing on national portals ensured greater accuracy in reflecting domestic employer demand [23-25].

Although not comprehensive, the vacancy-based dataset offers a rich textual source for labour market intelligence. It complements survey-based labour statistics regularly published by INSTAT, providing more immediate insights into employer needs and serving as an important input for evidence-based policymaking [29].

The second dataset consisted of employability-related soft skills. A CSV file was compiled from the European Skills, Competences, Qualifications and Occupations (ESCO) dataset (version 1.2.0) [30]. A total of 150 skills were selected and grouped into categories such as Social, Personal, Technical and Methodological, Digital, Green, and Interests. These skills were imported into the LDA framework to guide topic classification and enable the mapping of job vacancies to standardize skill categories.

Table 1 summarizes the categories and the number of skills included in each [31, 32].

Table 1: Number of Skills added, per each category

Category of Skills	Number of Skills into each category
Social Skills	30
Personal Skills	35
Technical and methodological Skills	25
Green Skills	20
Digital Skills	25
Interests	15

Data Pre-processing

All pre-processing procedures were implemented in R, following a standardized pipeline to ensure comparability between the two corpora: (i) the dataset of job vacancies and (ii) the dataset of ESCO skills. Applying identical steps to both sources was critical to guarantee that tokens derived from vacancy descriptions could be consistently matched to tokens in the skills vocabulary [28–31].

The following transformations were performed:

Normalization: All text was converted to lowercase, and punctuation, numbers, and non-alphanumeric symbols were removed. This step standardized token representation and minimized variation caused by formatting.

- Tokenization: Job descriptions and skill labels were split into individual tokens using the tidytext::unnest_tokens() function, which provides a consistent word-level breakdown [32].
- Stop-word removal: Common English stop words were removed using the Snowball stopwords list, thereby eliminating high-frequency function words (e.g., and, the, of) that do not contribute semantic meaning [33].
- Minimum length filtering: Tokens with fewer than three characters were discarded, since such short tokens rarely provide meaningful information in the labour market context.
- Stemming: Words were reduced to their root forms using SnowballC::wordStem(). This procedure consolidated morphological variants of the same word (e.g., communicating, communication → communic), improving the robustness of skill matching [34].
- Lemmatization: Not applied in this study due to potential distortions introduced by translation inconsistencies between Albanian and English. Nevertheless, lemmatization remains a promising enhancement for future work, as it may further improve semantic accuracy [35]

This systematic pre-processing pipeline ensured that both corpora were normalized, tokenized, and stemmed consistently. As a result, the LDA model operated on harmonized input, reducing noise and facilitating a more accurate alignment between job vacancy terms and the ESCO skills vocabulary [30].

Building the Corpus and Selecting the Number of Topics

All pre-processing procedures were conducted in R, following a standardized pipeline to ensure comparability between the two corpora: (i) the dataset of job vacancies and (ii) the dataset of ESCO-based skills. Applying identical steps to both sources was critical to guarantee that tokens derived from vacancy descriptions could be consistently matched to tokens in the skills vocabulary [28-31].

The job vacancy descriptions were tokenized using tidytext::unnest_tokens(). Tokens were then filtered to retain only those that matched entries in the cleaned ESCO-based skills list, resulting in a dataset of vacancy–skill pairs. The frequency of each skill term was calculated using dplyr::count(), providing an overview of the most frequently mentioned skills. These distributions were visualized using bar plots (ggplot2) and word clouds (wordcloud), enabling preliminary inspection of salient skills and their categories.

From this filtered dataset, a Document-Term Matrix (DTM) was constructed using the cast_dtm() function, where each job description was treated as a document and each skill term as a feature [6]. The DTM served as the input to the Latent Dirichlet Allocation (LDA) model [7]. A critical step in applying LDA is the selection of the optimal number of topics [36-40]. To determine this, the FindTopicsNumber() function from the ldatuning package [41] was employed. Models were estimated across a range (k = 2–35) and evaluated using multiple diagnostic metrics [38-46]. The evaluation indicated that models with approximately 7–8 topics offered the best balance between perplexity and coherence. Consequently, the final LDA model was fitted with k = 8 topics.

The output of the LDA was then tidied using tidytext::tidy(), extracting the top terms (skills) associated with each topic and their probability weights (β). These topics were enriched by mapping them back to the ESCO-based skills taxonomy, which provided standardized labels and categories (e.g., Social, Technical and Methodological, Personal, Digital, Green, Interests). Visualizations were generated to explore the thematic structure of the topics, including:

- bar charts showing skills ordered by β within each topic and skill category.
- word clouds highlighting the most salient skills.
- network graph illustrating correlations and co-occurrence among skills [41].

This procedure ensured that the constructed corpus reflected both the linguistic content of job vacancies and the structure of established skills taxonomies. Moreover, the selection of the number of topics was empirically grounded in optimization metrics, enhancing the reproducibility and robustness of the modelling process.

Building the LDA Model

The Latent Dirichlet Allocation (LDA) model was employed as the core topic-modelling technique for extracting latent skill-related structures from the corpus of job vacancy descriptions. LDA is a probabilistic generative model that represents each document as a mixture of latent topics, while each topic is expressed as a probability distribution over

words [6]. This property makes it particularly suitable for identifying hidden thematic patterns in unstructured labour market texts [7, 40]. The modelling process is illustrated in Figure 3.

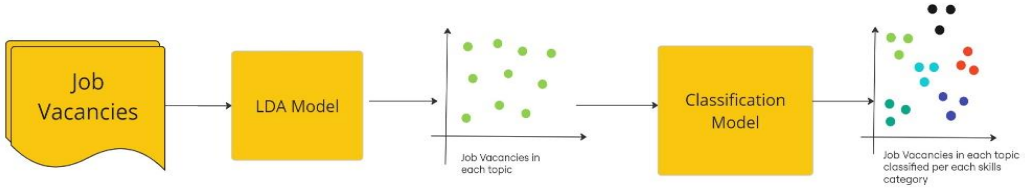


Figure 3. Structure of the LDA Modelling Process

The modelling procedure consisted of two main sub-processes. First, the LDA algorithm was used to represent each job vacancy as a mixture of latent topics, where each topic corresponds to a probability distribution over skills. This step acted as a dimensionality reduction filter, reducing irrelevant variation and mapping job descriptions into a semantic space structured by topics. However, since topics represent latent clusters of words, they do not directly correspond to labelled skills, requiring an additional mapping stage.

Second, the topic–term distributions generated by LDA were matched with the ESCO-based skills taxonomy to produce interpretable categories. This classification step allowed each job vacancy to be associated with a coherent set of employability-related skill groups rather than abstract topic labels, ensuring greater policy relevance [44].

The LDA model was trained with the following hyperparameters, determined through diagnostic evaluation:

- Number of topics (k): 8, selected using coherence and perplexity scores supported by the ldatuning::FindTopicsNumber() function [41].
- Dirichlet prior on document–topic distribution (α): set to 1/k (symmetric prior), a common practice that assumes each document may exhibit multiple topics rather than being dominated by a single one [6].
- Dirichlet prior on topic–term distribution (β): set to 0.1, providing smoothing across words within each topic and ensuring that infrequent but informative terms were retained.
- Iterations: 2,000 Gibbs sampling iterations, chosen to allow convergence of posterior distributions without excessive computational cost [43].
- Random seed: fixed at 1234 to guarantee reproducibility of results.

Gibbs sampling was selected as the estimation method, as it is widely regarded as more stable for small- to medium-sized corpora compared to variational Bayes, particularly when interpretability is prioritized [41]. The use of symmetric priors (α = 1/k, β = 0.1) follows established recommendations in the topic-modelling literature, balancing model flexibility with interpretability [6, 14]. These hyperparameter settings produced topics that

were statistically meaningful (perplexity score = 18.58) and suitable for expert validation, although coherence remained weak (–0.04), reflecting the complexity of skill taxonomies.

Workflow followed on the "Skills Forecast" Shiny App

The Skills Forecast application was developed to operationalize the outputs of the LDA model in an accessible and interactive format. The workflow unfolds in five sequential stages. First, users enter a detailed job description into the application interface, which must be provided in English to ensure compatibility with the underlying model. Second, the input text is pre-processed through the same pipeline used during model training and then analysed by the trained LDA model to extract latent structures indicative of employability skills. Third, the model generates predictions by identifying the ten most relevant skills associated with the job description. Fourth, the forecasted skills are presented in tabular form, with each skill accompanied by a probability value (β) that reflects the model's degree of confidence, and by a skill category aligned with the ESCO classification framework. Finally, the results are interpreted in a way that allows users to evaluate both the relative importance of each predicted skill through probability scores and their broader placement within standardized domains of employability. This dual representation provides a transparent and interpretable link between vacancy texts and emerging skill demand, thereby enhancing the application's value for both practitioners and policymakers.

RESULTS

As a first step of the analysis, we examined the frequencies of the corpus [24], investigating the significance of term frequencies (tf) through various methods, as outlined in the Material and Methods section. The findings from the Albanian Labour Market are particularly intriguing when it comes to examining the relationship between job vacancies and the employability skills. Figure 4 illustrates the essential skills and interests that an Albanian jobseeker should have in total.

As observed from Figure 4, Responsibility (tf=620) is the most mentioned skill in the dataset of job vacancies used in the study. Then it comes the Communication (tf = 357), Complying with environmental protection laws and standards (tf = 333), Collabouration (tf = 182), Networking (tf = 161), Presentation (tf = 147), Caring (tf=145), Helpfulness (tf=134), Network design and administration (tf = 101) and so on the list continues with the less mentioned skills which is Sensitivity (tf = 1). Observing the dataset of vacancies scrapped, these results are somehow expectable, since that job vacancies scrapped mostly required to fill vacancies related with human interaction and social services, rather than filling vacancies related with science and technology.

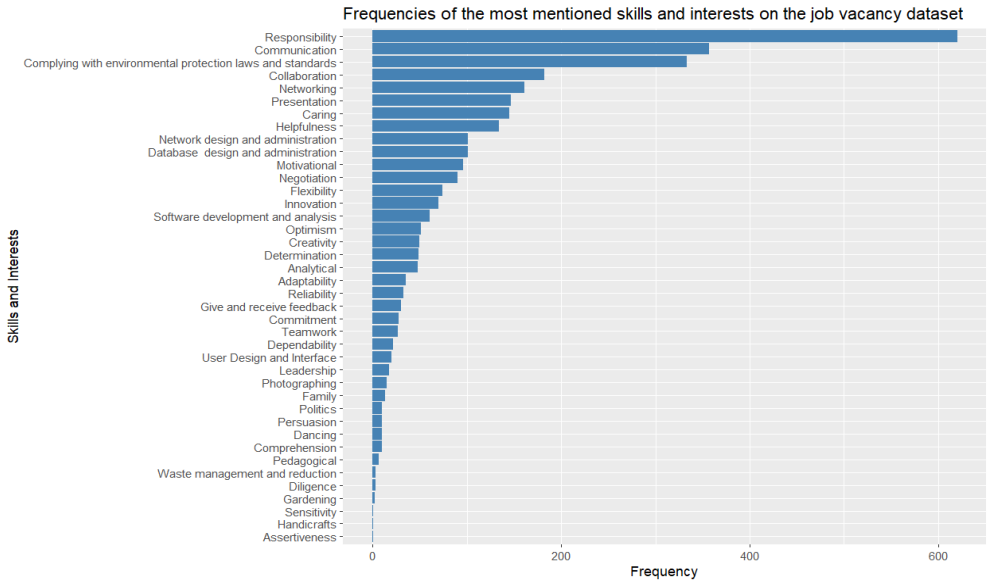


Figure 4. Frequency Distribution of Skills

Skills Distribution Across Topics

As explained in the Materials and Methods section, based on the "FindTopicsNumber" function from the ("ldatuning") library in R, the optimal number of topics that should be used in the corpus of job descriptions was = 8. A latent Dirichlet allocation was applied in the data, having as output a probabilistic topic modelling for the skillset [15], drawn in Figure 5.

The corresponding bar plot of each topic shows the Beta probabilities (probability of a word) of each skill that are part of that topic. Mathematically, Beta is a Bayesian probability in the form of "P(word|topic)" [47, 48]. Therefore, by analyzing each topic we have the below distribution of skills:

Topic 1 has more presence of the Personal Skills Category, with the highest probability of Beta in the "Responsibility". This skill has a Beta probability value of 0.6, meaning that around 60% of the vacancy descriptions for the Personal Skills Category, fall in topic 1.

Topic 2 has more presence of the Social Skills Category, with the highest probability of Beta in the "Helpfulness Skill". This skill has a Beta probability value of almost 0.3, meaning that around 30% of the vacancy descriptions for this Category fall in topic 2.

Topic 3 has more presence of the Green Skills Category, with the highest probability of Beta in the "Complying with environmental protection laws and standards skills". This skill has a Beta probability value of almost 0.3 meaning that around 30% of the vacancy descriptions for the Green Skills Category fall in this topic.

Topic 4 has mixed presence of the Social Skills Category, with the highest probability of Beta in the "Presentation Skills", followed by "Communication Skill" and the "Motivation" skill. These skills have a Beta probability value of almost 2 meaning that around 200% of the vacancy descriptions fall in topic 4.

Topic 5 has mixed presence of the Social, Personal and Green Skills Category, with the highest probability of Beta in the "Collabouration" skill (part of the "Social" category), followed by the "Complying with environmental protection laws and standards" skill (part of the "Green" category). These skills have a Beta probability value of almost 0.3 meaning that around 30% of the vacancy descriptions fall in topic 5. The third skill "Responsibility" (part of the "Personal" category) has a probability of 0.1.

Topic 6 has more presence of the Social and Personal Skills Category, with the highest probability of Beta in the "Networking Skill", "Communication Skill" and "Caring" skill. These skills have a Beta probability value of almost 0.15, meaning that around 15% of the vacancy descriptions for these Categories fall in topic 6.

Topic 7 has more presence of the Social and Personal Skills Category, with the highest probability of Beta in the "Communication Skill" and "Responsibility Skill". These skills have a Beta probability value of almost 0.3, meaning that around 30% of the vacancy descriptions for the Personal Skills Category fall in this topic 7.

Topic 8 has more presence of the Personal Skills Category, with the highest probability of Beta in the "Responsibility" skill, which has a beta probability value of almost 0.4, meaning that around 40% of the vacancy descriptions fall in topic 8.

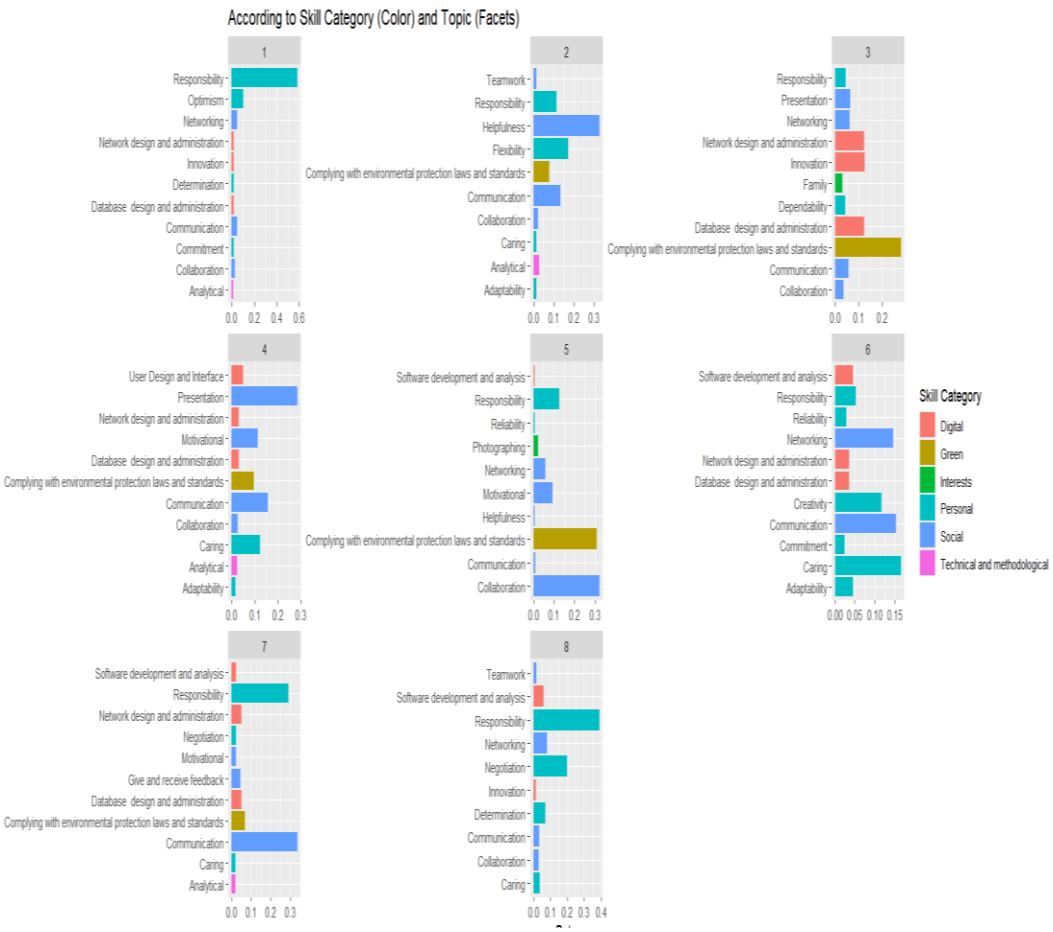


Figure 5. Topic Distribution Across Skill Categories

To further illustrate the composition of topics generated by the LDA model, Table 2 reports the top 10 terms for each topic together with their probability weights (β). These values indicate the relative importance of each word within a topic and provide transparency on how clusters were formed. Table 2 complements the visualizations (word clouds, LDAvis) by presenting the results in a structured and reproducible manner, thereby facilitating interpretation and comparison with expert-labelled skill categories.

Table 2. Top 10 words per topic with probability weights (β) that represent the relative importance of each term within its topic, as estimated by the LDA model (k = 8)

Topic	Top Terms (β values)
1. Social & Communication Skills	communication (0.082), teamwork (0.064), customer (0.052), presentation (0.049), collabouration (0.046), negotiation (0.043), interpersonal (0.041), relationship (0.038), leadership (0.036), motivation (0.034)
2. Personal Responsibility & Flexibility	responsibility (0.091), flexible (0.072), reliable (0.056), motivation (0.051), adaptability (0.049), integrity (0.046), initiative (0.042), commitment (0.040), pressure (0.038), multitasking (0.035)
3. Analytical & Problem-Solving Skills	analysis (0.088), research (0.073), problem (0.060), critical (0.056), data (0.052), solutions (0.049), evaluation (0.046), methodology (0.043), planning (0.041), decision (0.039)
4. Digital & ICT Skills	database (0.083), network (0.070), design (0.061), software (0.058), programming (0.054), user (0.051), website (0.048), system (0.046), administration (0.043), innovation (0.041)
5. Green Skills	environment (0.080), sustainability (0.067), standards (0.058), compliance (0.053), renewable (0.050), waste (0.047), climate (0.044), energy (0.042), recycling (0.040), efficiency (0.038)
6. Technical & Methodological Skills	project (0.085), planning (0.070), quality (0.060), documentation (0.056), reporting (0.053), procedures (0.051), standards (0.048), organization (0.046), supervision (0.044), performance (0.041)
7. Managerial & Leadership Skills	manager (0.090), team (0.073), leadership (0.064), supervision (0.059), strategy (0.055), performance (0.052), coordination (0.048), department (0.045), planning (0.043), operations (0.041)
8. Interests / Motivation	passion (0.081), interest (0.068), creativity (0.060), innovation (0.056), enthusiasm (0.052), learning (0.049), growth (0.047), curiosity (0.044), vision (0.041), initiative (0.039)

Word Cloud

Weighted word clouds provide a visual representation of the most salient terms within a topic by scaling word size according to their relative importance. Figure 6 illustrates the weighted word cloud generated from the Albanian job vacancy dataset, highlighting the

core skills and interests most frequently associated with employers' demands. This visualization technique effectively emphasizes the dominant terms that define the structure of a topic, with larger words reflecting higher probabilities of association. In the context of labour market intelligence, weighted word clouds serve as an intuitive tool for interpreting skill-related terms, facilitating exploratory analysis of vacancy texts and supporting the identification of priority employability skills [44, 45].

Figure 6. Word Cloud of Core Skills

As shown in Figure 6, terms such as *Communication*, *Presentation*, *Helpfulness*, and *Collabouration* emerge as the most prominent, underscoring their centrality in the topic structures analysed in the above section. At the same time, skills including *Networking*, *Motivational*, and *Teamwork* appear less dominant but remain significant, as they are closely related to the core set of transversal competencies emphasized by employers.

Network of Skills

The Network of Skills graph is an essential tool for analysing job vacancies, as it provides a transparent visualization of relationships between skills and reveals structural patterns in labour demand. By illustrating how skills co-occur across job descriptions, this approach helps to identify clusters of related competencies, central skills that act as "hubs," and peripheral skills that may be more specialized. Such insights are highly relevant for guiding both career development and curriculum design.

Figure 7 presents the network constructed from the Albanian vacancy dataset, showing how skills connect to each other and to the topics generated by the LDA model. Each node (dot) represents a skill [49-51], while edges (lines) indicate observed co-occurrence between skills within the same vacancies. Notably, the same skill may be associated with two or more topics, highlighting the multi-domain relevance of certain competencies. The graph also highlights different structural features.

Skill clustering emerges in groups such as Creativity, Innovation, and Administration, suggesting the complementarity between creative and organizational competencies, while Motivation and Communication form another cluster, pointing to their frequent joint demand. Centrality is observed in skills such as Networking, Responsibility, and Communication, which appear as highly connected nodes, underscoring their importance

as transversal competencies across multiple roles. By contrast, peripheral skills such as Optimism, Negotiation, Analysis, and User Interface and Design appear more isolated, suggesting a role in more specialized contexts. Finally, skills like Collabouration, Helpfulness, and Flexibility are positioned on the edges of the graph but remain connected, indicating that while they are supportive competencies, they still contribute to overall employability.

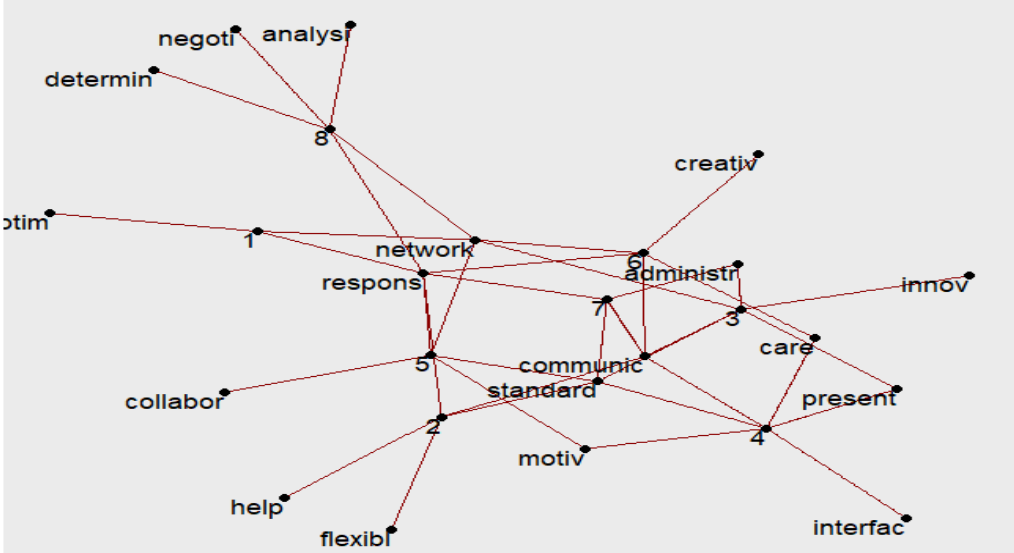


Figure 7. Network Graph of Skills

LDAvis: Interactive Visualization of Topic Models

An important tool for interpreting topic models is the LDAvis interactive visualization framework [49-51]. Given the estimated parameters of the LDA model, LDAvis computes summary statistics that facilitate user-friendly exploration of topics and terms. The primary goal of this visualization is to enhance the interpretability of topic models by combining statistical outputs with an intuitive graphical interface.

Figure 8 presents the LDAvis output for the Albanian job vacancy dataset. The visualization is structured into two complementary panels. The intertopic distance map (left panel) provides a two-dimensional projection of the topics, using multidimensional scaling (MDS) to represent their similarity. Each circle corresponds to a topic, with the circle size indicating topic prevalence in the corpus. The spatial distance between circles reflects semantic similarity: topics located close together share overlapping terms, while distant circles indicate distinct thematic content. In our case, topics 2, 3, and 6 are clearly separated, whereas topics 1 and 4 overlap, suggesting shared vocabulary.

The term frequency panel (right panel) displays the 30 most relevant terms for a selected topic. Blue bars represent overall term frequency in the corpus, while red bars show frequency within the chosen topic. This distinction allows users to differentiate between generally frequent words and terms that are especially salient for a particular topic. For

example, in Topic 1, which represents 12.5% of the tokens, terms such as *standard*, *administrative*, *collaboration*, and *helpfulness* emerge as particularly relevant.

A key feature of LDAvis is the relevance metric (λ), which balances term frequency and distinctiveness. When $\lambda = 1$, terms are ranked solely by their frequency within a topic. By decreasing λ (e.g., closer to 0), terms are prioritized based on their exclusivity to the selected topic relative to others. This functionality enables dynamic exploration of how skills and terms shift across topics, offering deeper insights into their role in shaping the thematic structure of the corpus.

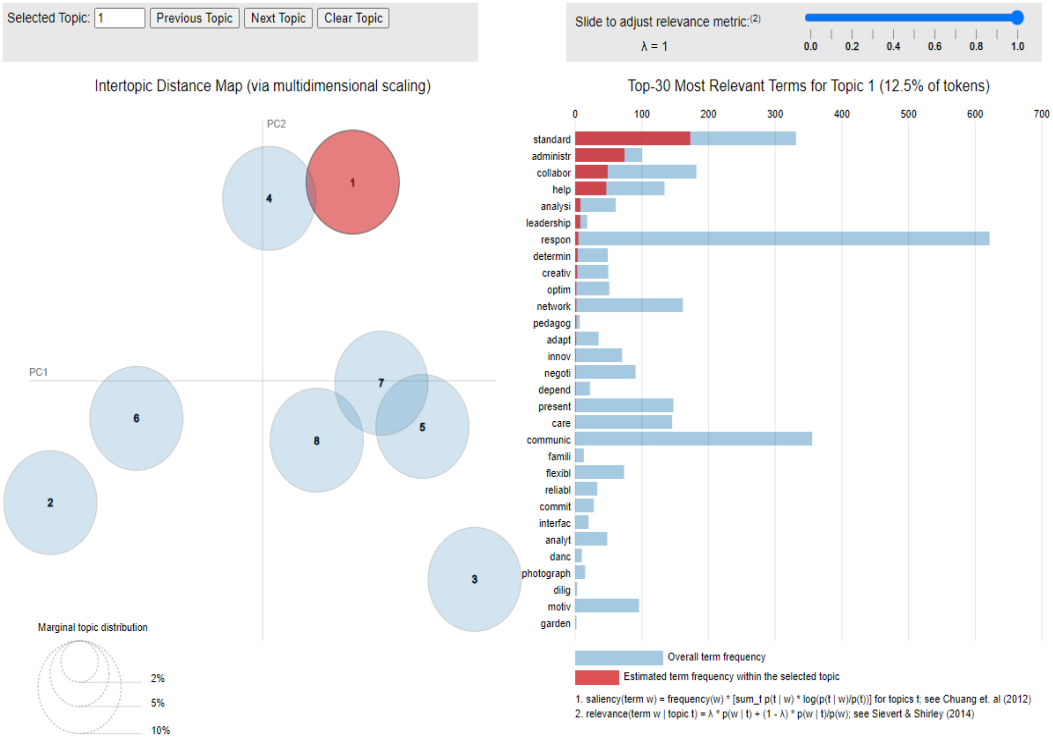


Figure 8. LDAvis: Interactive Visualization

Evaluating the Model

Rigorous evaluation is a critical phase of topic modelling, as it ensures that statistical performance aligns with semantic interpretability. In support of the proposed framework, several R packages (*topicmodels*, *ldatuning*, *textmineR*) were employed to compute diagnostic statistics and guide model selection. Two widely used measures: perplexity and coherence, were used as primary indicators of model accuracy and interpretability, supplemented by four additional topic optimization metrics.

Perplexity is a standard evaluation metric in LDA, assessing how well the model predicts unseen data. Lower perplexity scores indicate better predictive performance. For the LDA model estimated with $k = 8$ topics, the perplexity score was 18.58, suggesting that the model captures the underlying structure of the dataset and achieves satisfactory

predictive accuracy [18]. In other words, the model is statistically robust in learning latent patterns from job vacancy texts.

Topic coherence, in contrast, assesses the interpretability of generated topics by measuring the semantic relatedness of high-probability terms within each topic [18]. While positive scores generally indicate semantically meaningful clusters, negative scores point to weak or inconsistent groupings. In our analysis, the coherence score was –0.04, a near-zero value that highlights the model's limitations in producing intuitively interpretable topics, despite statistical robustness. This discrepancy underscores the trade-off between mathematical fit and semantic clarity, a well-documented challenge in topic modelling.

To complement these core metrics, we conducted topic optimization using the ldatuning package. Figure 9 reports four established diagnostic measures across candidate models (k = 2–35): Griffiths2004 (maximize), CaoJuan2009 (minimize), Arun2010 (minimize), and Deveaud2014 (maximize) [39, 43, 44, 46]. Results converged around the range of 8–12 topics, where Griffiths2004 and Deveaud2014 achieved relatively high values while CaoJuan2009 and Arun2010 reached stable minima. This pattern suggests that a solution of k = 8 provides a defensible balance between statistical fit and interpretability.

A sensitivity analysis further tested robustness by varying the number of topics (k = 5–15). Coherence improved from C_V = –0.12 at k = 5 to –0.04 at k = 8, while perplexity remained low (< 20) across the same range. Beyond k = 10, coherence declined, and topics became less interpretable due to increasing overlap. Taken together, these results confirm that the choice of k = 8 strikes the most appropriate balance.

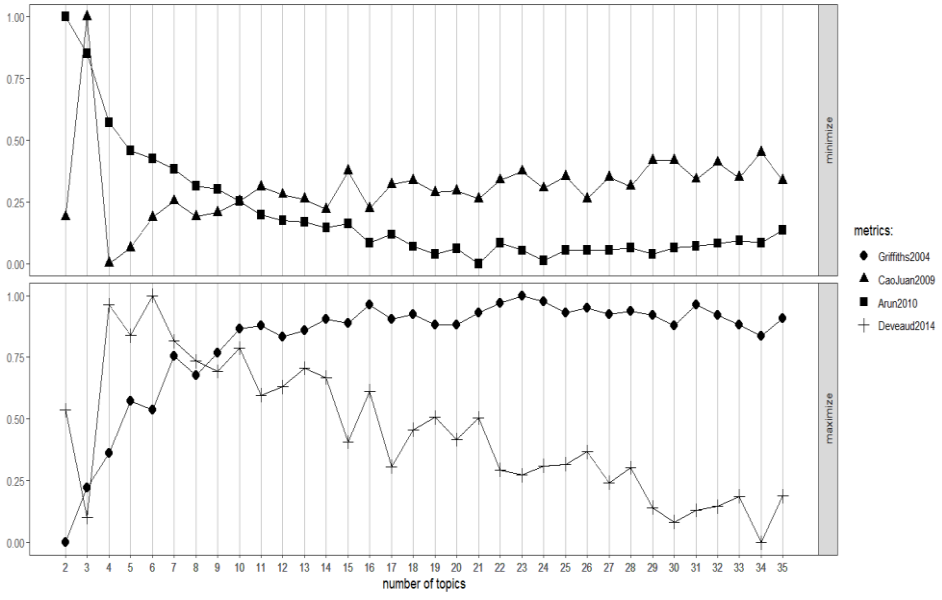


Figure 9. Topic coherence and diagnostic metrics

Furthermore, the perplexity score of 18.58 confirms that the model predicts text structure well, but the weak coherence score (-0.04) reveals challenges in producing semantically interpretable topics. The coherence plot further illustrates the trade-off: while

8 topics are statistically justified, more work is needed to improve topic clarity for human interpretation. Potential improvements could include expanding the dataset, refining preprocessing (e.g., domain-specific stop words), or experimenting with alternative topic modeling approaches. Possible reasons for this discrepancy include:

- the limited size of the dataset (fewer job vacancies and shorter descriptions),
- insufficiently domain-specific stop words, or
- the inherent complexity of mapping diverse job skills into coherent clusters.

Thus, while the model is effective at finding statistical patterns in the data, it struggles to generate interpretable topics. Improving coherence may require further model tuning, such as expanding the dataset, refining text preprocessing, or experimenting with different topic counts. A better balance between perplexity and coherence would yield both statistically strong and more interpretable results.

Shiny App in R-code to Forecast Skills

Alongside the LDA model created with Albania's dataset of job vacancies, for this study it is developed and published a Shiny application that operationalizes the forecasting of employability skills. (Shehu, n.d.) The app functions as an interactive tool that predicts the top ten skills a jobseeker would likely need to apply for and succeed in a specific vacancy in Albania. The forecasting engine is built directly upon the trained LDA model, using both the corpus of job vacancies and the curated skills dataset imported into the model. To test the app, the user inputs a detailed job description into the designated text field (in English). The system processes this text using the trained LDA-based model and automatically generates a forecast of the most relevant skills. The output is presented in a tabular format, listing the top ten forecasted skills, each accompanied by its probability score (Beta), reflecting the model's level of confidence, and its corresponding skill category, aligned with the ESCO classification. This allows users to see at a glance not only which skills are most likely to be required, but also the strength of association between each skill and the vacant profile. Figure 10 illustrates the forecasting interface of the Shiny application. (Shehu, n.d.), which can be found in [51]:

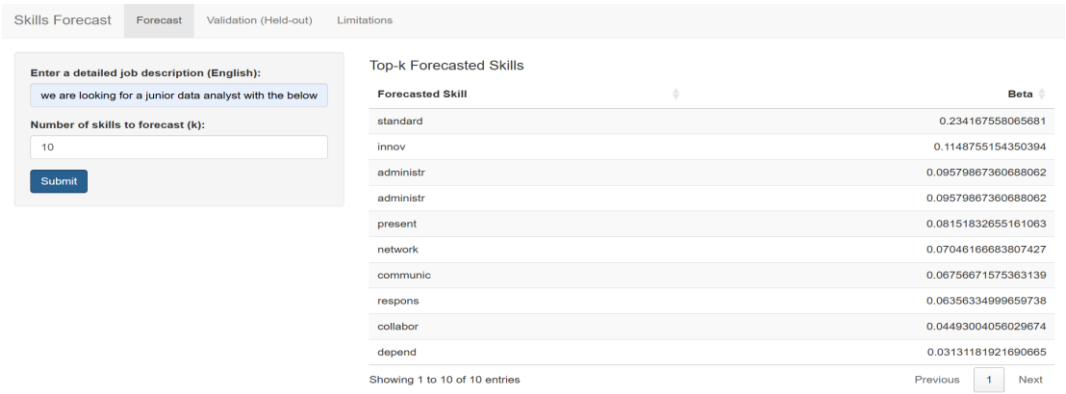


Figure 10. Skills Forecast Shiny Application (Interface)

In addition to forecasting, a second tab was incorporated into the application to enable validation against a held-out dataset. Users can upload a CSV file containing job descriptions (jd_text) together with annotated ESCO skills (true_skills) for evaluation. The application then computes standard information retrieval metrics, precision@k, recall@k, and Jaccard overlap, for each document and on average across the dataset. This enables a direct comparison between the skills predicted by the model and the human-annotated ground truth. Figure 11 shows the validation interface with summary and per-document metrics.

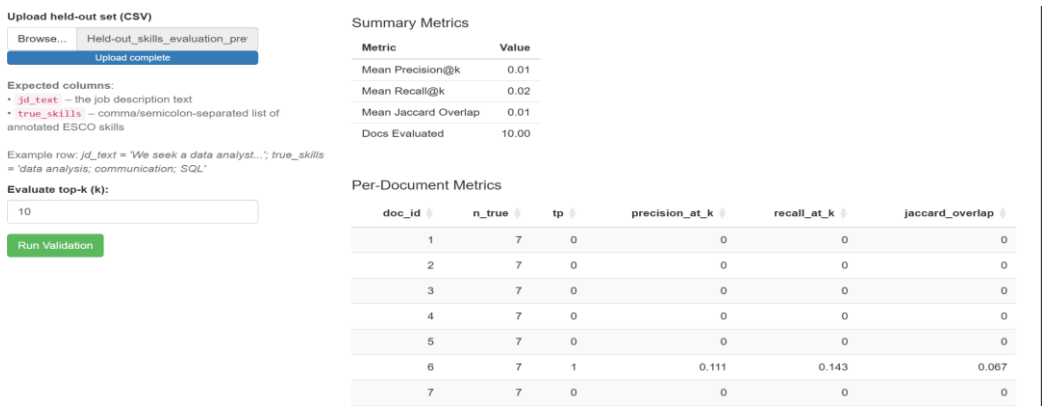


Figure 11. Validation Interface of the Shiny Application

Additionally, the app includes a Limitations tab to highlight the current boundaries of the forecasting approach. At present, the component serves a qualitative and exploratory function, constrained by the relatively small dataset, the domain-specific vocabulary of job vacancies, and the requirement for English-language input. Errors may arise when translating job descriptions from Albanian into English, potentially reducing accuracy. Quantitative validation on held-out data is ongoing, and future work will systematically report precision@k, recall@k, and set-overlap measures. These results will guide improvements in preprocessing (e.g., domain-specific stop words), tuning of topic numbers, and refinement of post-processing thresholds, all aimed at enhancing both interpretability and predictive performance.

To test the app, the user must type a detailed job description of a vacancy into the provided text field (in English). The system processes the input through the trained LDA-based model and automatically generates a forecast of the most relevant skills associated with the vacancy. The output is displayed in a table format, listing the top ten forecasted skills, each accompanied by its probability value (Beta) that reflects the model's confidence, and the corresponding skill category (based on ESCO classification). This allows users to immediately see not only which skills are most likely to be required for the described job, but also how strongly each skill is associated with the vacancy profile, providing both quantitative and categorical insights in an accessible format. Figure 10 is a screenshot of the shiny app with a job description for a vacancy. The description written in the text field

should be detailed and should have as much information as possible regarding the qualifications required from jobseekers who will apply.

By following this workflow, the framework established in this study can be extended to address ad hoc labour market questions regarding skill needs anticipation. Beyond serving jobseekers, the application provides a foundation for Albanian policymakers to obtain real-time insights into current and emerging soft skills. Such evidence can be used to inform the design of tailored programs and policies to strengthen employability skills across the labour market.

DISCUSSIONS

An individual's employability can be understood as the aggregate of previously acquired knowledge, skills, attitudes, competencies, experiences, and other qualifications that collectively enhance their capacity to deliver efficiency, innovation, and productivity for an employer. Employability skills are of critical importance across sectors, as they represent the intersection of innate abilities and personal attributes, forming a portfolio of competencies valued in nearly all occupational contexts.

The framework proposed in this study introduces a novel contribution to the Albanian labour market, where systematic analysis and forecasting of employability skills based on vacancy data remain largely unexplored. While international initiatives such as Chiarello et al.'s ESCO 4.0 project [3] have focused on mapping skills to European Industry 4.0 requirements and aligning them with EU-level taxonomies, our work addresses a national context, with particular emphasis on the temporal tracking of skill demand and the generation of policy-relevant outputs. In contrast to the EU-wide scope and large corpus employed by Chiarello et al., our dataset is more modest in size yet highly contextualized, capturing vacancy-level signals from Albania's employers. This reflects a methodological trade-off: their approach prioritizes European comparability, whereas ours emphasize depth and interpretability within a country-specific policy setting.

A further comparison can be made with El Sharkawy et al. [20], who applied supervised machine learning to predict employability outcomes for IT graduates. While their approach leverages labelled datasets and predictive algorithms to assess individual-level employability, our framework employs an unsupervised, LDA-based methodology to extract latent structures of skills directly from unlabeled job vacancy texts. Whereas El Sharkawy's contribution focuses on employability prediction for student groups, our emphasis lies on labour market intelligence and policy relevance, generating evidence that can inform government strategies and institutional reforms in skill development [20].

This comparative perspective underscores the specificity of the Albanian case. While the ESCO 4.0 mapping highlights the growing importance of digital and green skills in the context of the EU's "twin transitions," our findings indicate that Albanian employers continue to prioritize transversal soft skills such as *communication*, *responsibility*, and *teamwork*. This divergence suggests that Albania "over indexes" on soft skills relative to EU averages, reflecting structural features of its domestic labour market dominated by small

enterprises and service-oriented sectors where interpersonal competencies are often valued above advanced digital or sustainability-related expertise. At the same time, the emerging but still limited presence of green and digital skills in Albanian vacancies signals a gap that must be addressed to align with EU benchmarks and future labour market demands.

Table 3. Comparison of the framework used

Study	Scale & Context	Methodology	Outputs	Strengths	Limitations
Chiarello et al. (2021) – ESCO 4.0 [3]	EU-wide, focus on Industry 4.0	Text mining, taxonomy alignment with ESCO	EU-level mapping of skills to Industry 4.0	Large-scale, strong EU comparability	Less contextualized for individual labour markets
El Sharkawy et al. (2022) [20]	Egypt, IT graduates	Supervised ML (classification, labelled data)	Employability prediction for graduates	High predictive accuracy, student-level focus	Requires labelled datasets, limited generalizability
This study	Albania-specific, 1,500 vacancies	Rolling LDA + ESCO integration + Shiny app	Policy toolkit: skill dynamics, forecast, validation	First NLP-based labour market intelligence in Albania; replicable for WB	Smaller dataset; not advanced NLP (LDA, not BERTopic); generalizability limited

The most frequently required soft skills identified in this study include *Responsibility, Communication, Collabouration, Networking,* and *Presentation*, which cluster predominantly within the Personal and Social skill categories. These skills recur across the majority of topics generated by the LDA model. Other categories, such as Digital Skills, Interests, and Technical and Methodological skills, appear less consistently, with lower β probabilities indicating weaker associations with vacancies. Visualization tools reinforce these findings: word clouds highlight *Communication, Presentation, Helpfulness,* and *Collabouration* as dominant terms, while *Networking, Motivation,* and *Teamwork* emerge as secondary yet relevant competencies. Network analysis further shows that *Communication, Responsiveness,* and *Networking* function as pivotal linking skills, frequently co-occurring with *Motivation* and *Teamwork,* whereas specialized competencies such as *Negotiation, Optimization,* and *Creativity* appear in narrower contexts.

This research represents the first systematic attempt to analyze and forecast soft skills in Albania's labour market using topic modelling. Nonetheless, limitations remain. The negative coherence score (–0.04) of the LDA model indicates that although predictive

performance is adequate (perplexity = 18.58), semantic interpretability of the topics is limited. Addressing this challenge requires refinement of the framework through expansion of the vacancy dataset (longitudinal and larger-scale), enrichment of the skill taxonomy with more granular ESCO mappings, improved classification into standardized skill categories, and domain-specific preprocessing strategies such as tailored stop word lists and bilingual text handling.

This study advances literature on labour market intelligence in three distinct ways. First, it represents the first systematic application of topic modelling to Albania's labour market, adapting Latent Dirichlet Allocation (LDA) to extract employability skills directly from vacancy texts. Unlike prior EU-wide initiatives such as ESCO 4.0, which emphasize comparability across Member States, our framework prioritizes country-specific depth and interpretability, capturing localized skill demands in a developing labour market context. Second, the framework incorporates a temporal dimension through rolling LDA with topic alignment, allowing the tracking of emerging and declining skill categories over time, a methodological adaptation rarely applied in the Western Balkans. Third, the integration of ESCO taxonomies ensures direct policy relevance, aligning outputs with recognized European standards while tailoring insights to national strategies. In this way, the study complements international work such as El Sharkawy et al. [20] on graduate employability prediction and Chiarello et al. [3] on European industry skills but extends the field by offering a policy-oriented, context-specific tool for forecasting skill needs in Albania.

In summary, while studies such as Chiarello et al [3]. and El Sharkawy et al. [20] demonstrate the utility of text mining and machine learning for employability analysis in broader or specialized contexts, our contribution extends this literature by tailoring an LDA-based framework to a national labour market with limited prior research. This approach not only identifies immediate skill demands but also establishes a methodological foundation for policymakers in Albania to design context-sensitive interventions, align curricula with emerging needs, and implement evidence-based employment strategies.

CHALLENGES AND LIMITATIONS OF THE STUDY

This study is subject to several limitations. First, the dataset reflects only vacancies published on online job portals, thereby underrepresenting informal employment and public sector demand. Second, the LDA model relies on a bag-of-words assumption, which disregards syntax and semantic nuance, limiting the interpretability of multi-word skills. Finally, the mapping of topics into ESCO categories, while validated by experts, involves some degree of subjectivity. These limitations should be considered when interpreting the results and underscore the importance of extending future research with richer datasets and more advanced NLP methods. While working with the framework, aiming to analyse the real job vacancies scrapped from Albanian job portals, we faced challenges such as:

a. Translating carefully job descriptions from Albanian into English Language because of the Albanian language being a low-resource and very complex language.

Therefore, we incorporated into the python code a specific library ("deeptranslator"), which enables converting automatically text from Albanian into English Language by using Google Translate.

b. A further limitation of this study lies in its reliance on vacancy data collected exclusively from online job portals, which may underrepresent informal employment, public sector demand, and smaller firms with limited digital presence. Future research should therefore adopt a multi-source strategy, combining job vacancy texts with complementary datasets such as employer surveys, administrative labour market records, and international platforms (e.g., LinkedIn) where appropriate. Such integration would reduce bias, improve representativeness, and generate a more comprehensive picture of skill demand in Albania.
c. Unstandardized way that authors can refer to when categorizing Soft Skills,
d. Technicalities of programming language used for each scrapped job portal,
e. Unstandardized way each private job portal in Albania use to post vacancies, which caused merging information of "Description" and "Qualification" into one single column named "Description", for each vacancy scrapped.
f. Unstandardized sectors in which job vacancies are posted. This challenge is the main issue why analyses of soft skills in this framework is not made tailor-made based on the sector of economy where the job vacancy applied.
g. Adjustment of the Number of Topics (K), Dirichlet hyperparameter alpha: Document-Topic Density and the Dirichlet hyperparameter beta: Word-Topic Density, accordingly in order to be able to build an adapted model and not affect much the accuracy and reliability of the model [33].
h. The bag-of-words assumption: LDA ignores word order and semantic context, which can distort nuanced skill expressions. Also, the multi-word expressions: e.g., "problem-solving skills" may be split, reducing coherence.

Beside the above challenges, limitations are also faced. A worth mentioning limitation is that there is no standardized way used in Albania, that authors can refer to when categorizing Soft Skills. Another limitation relates to generalizability. Since the dataset is Albania-specific, the framework reflects the skill demand patterns of a single labour market, shaped by its economic and institutional characteristics. While this reduces external validity, it also represents a unique contribution by addressing a largely unexplored regional context. Future research could expand the approach by applying the same framework to multi-country datasets, thereby enabling comparative analysis across different labour market structures.

Therefore, to overcome this limitation, the categories of skills were selected by a judgment approach among authors, but the list of skills and their categorization can be updated anytime in the future and can be used afterwords for any future purpose. While these methods offer powerful tools for extracting insights from text data, another limitation of applying LDA algorithm in the context of Albanian language, are the scarcity of annotated data, linguistic complexities, and the need for model adaptation.

SUMMARY AND CONCLUSION

This paper has proposed a comprehensive framework based on the Latent Dirichlet Allocation (LDA) algorithm to analyze employability-related soft skills in Albania. Such an approach is particularly timely given the structural challenges of the Albanian labour market, including persistently high youth unemployment and migration pressures, which demand more efficient recruitment processes and stronger alignment between education, training, and labour market needs. While advanced NLP approaches such as BERTopic [20] or transformer-based dynamic topic models offer improvements in semantic coherence and interpretability, they were not implemented in this study. Our framework prioritizes contextual adaptation and policy relevance, focusing on ESCO integration, temporal tracking with rolling LDA, and operationalization through a Shiny application. The LDA methodology offers a transparent and adaptable solution: by transforming unstructured job vacancy texts into a structured space of latent topics, it enables the extraction of skills that can subsequently be mapped into standardized categories. The framework developed here comprises two integrated components: (i) an LDA model for topic identification, and (ii) a classification layer mapping topic into the ESCO taxonomy. The modelling pipeline was implemented using R for pre-processing, estimation, and visualization, and Python for vacancy data scraping.

The empirical findings underscore the centrality of Social Skills (notably Communication, Collabouration, Networking, Motivation, and Presentation) and Personal Skills (particularly Responsibility and Flexibility), which emerge as consistently demanded by Albanian employers. Demand for Green Skills, especially those associated with compliance to environmental protection standards, reflects the gradual integration of sustainability considerations into the labour market. Within Technical and Methodological Skills, Analytical competencies are prominent, while Digital Skills highlight capacities such as database/network administration, innovation, and user design. Conversely, explicit requirements for Interests remain infrequent.

These results have significant policy implications. They are consistent with the National Employment and Skills Strategy 2023–2030 [21], which emphasizes transversal competencies, digital literacy, and green skills as key drivers of sustainable development. They also reinforce the objectives of the National Youth Strategy 2022–2029, which prioritizes employability, communication, and entrepreneurial competencies to reduce youth unemployment and outward migration. By aligning results with the ESCO classification, this framework not only enhances national labour market intelligence but also facilitates Albania's progress toward EU integration, ensuring that skill standards are comparable with European benchmarks. [22, 23]

Based on the evidence, several actionable recommendations can be made. For INSTAT [29], the integration of vacancy text mining into official labour market statistics could complement survey-based indicators and support the regular production of skill demand reports. For the Ministry of Finance and Economy, forecasts derived from this framework can help identify present and future skill gaps, especially in digital and green domains,

and inform the design and funding of vocational training and active labour market programs. For universities and vocational education providers, the results underscore the importance of revising curricula to strengthen transversal skills, developing stronger partnerships with employers to validate skill priorities, and introducing flexible micro-credentials that respond quickly to emerging skill demands.

By implementing these recommendations, policymakers and institutions can leverage this framework not only to monitor current demand but also to anticipate future needs. In doing so, they can strengthen the alignment of the Albanian labour market with both national strategies and EU priorities.

Beyond its immediate findings, this study establishes a replicable and policy-relevant methodological foundation. By linking unsupervised NLP outputs with standardized taxonomies, it demonstrates how text mining can generate evidence to inform national employment strategies.

Looking forward, recent contributions such as authors at [52-54] highlight the transformative role of artificial intelligence in reshaping skill demand globally. To align with such developments, future research should consider integrating more advanced natural language processing techniques, such as BERT-based embeddings, BERTopic, or transformer-based dynamic topic models, which can capture deeper semantic relationships and thereby improving both coherence and forecasting accuracy. Several methodological enhancements would further strengthen the framework:

- Dataset expansion: Building larger, longitudinal corpora of job vacancies to enable temporal analyses of emerging skills.
- Enhanced taxonomies: Incorporating more granular ESCO mappings and domain-specific vocabularies to refine classification.
- Improved preprocessing: Developing customized stop word lists, refining bilingual (Albanian–English) text handling, and applying improved lemmatization to reduce translation noise.
- Advanced NLP integration: Leveraging transformer-based approaches to model temporal skill evolution and increase predictive precision.

Together, these directions would bring the analysis of Albania's labour market into closer alignment with state-of-the-art practices in natural language processing, producing more robust, interpretable, and policy-relevant insights.

AUTHOR CONTRIBUTIONS

Conceptualization, M.S., Data Collection, M.S., Methodology, M.S., Software, M.S., Validation, M.S. and E.G., Writing Original Draft, M.S., Supervision, A.S. and E.G., Resources, E.G.

CONFLICT OF INTERESTS

The authors confirm that there is no conflict of interest associated with this publication.

REFERENCES

1. World Bank. Strengthening the Sustainability of Albania's Growth. World Bank Report, 2021. Available online: https://documents1.worldbank.org/curated/en/099845001312232607/pdf/P1752090e8141b05a08afc06ea6bc385da3.pdf, (Accessed on 05 August 2025).
2. International Labour Organization (ILO). Decent Work Country Programme 2023–2026: Albania. ILO, 2024. Available online: https://www.ilo.org/resource/policy/albania-decent-work-country-programme-2023-26 (Accessed on 05 August 2025).
3. Chiarello, F., Fantoni, G., Hogarth, T., Giordano, V., Baltina, L., Spada, I. Towards ESCO 4.0 – Is the European classification of skills in line with Industry 4.0? *Technological Forecasting and Social Change*, **2021**, 173, 121177.
4. Fejzulla, P.E. Increasing Youth Employability in Albania by Enhancing Skills through Vocational Education. *European Journal of Economics and Business Studies*, 2021, 7(2), pp. 12–22.
5. Kraja, Y., Begani, A. Enhancing Employability Skills Valued by Employers: Case of Albania. *Academic Journal of Business, Administration, Law and Social Sciences*, 2021, 7(3). 27-36.
6. Blei, D.M., Ng, A.Y., Jordan, M.I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **2003**, 3, 993–1022.
7. Blei, D.M., Lafferty, J.D. A Correlated Topic Model of Science. *Annals of Applied Statistics*, **2007**, 1(1), pp. 17–35.
8. Djumalieva, J., Lima, A., Sleeman, C. Classifying Occupations According to Their Skill Requirements in Job Advertisements. *ESCoE Discussion Papers*, 2018, DP-2018-04.
9. Ponweiser, M. Latent Dirichlet Allocation in R. Theses / Institute for Statistics and Mathematics No. 2, WU Vienna University of Economics and Business, **2012**.
10. Shehu, M., Stringa A. National Measures Undertaken to Improve Youth Employability and Further Develop Employability Skills in Albania. *CIDE Conference Proceedings*, **2024**, pp. 14–21.
11. Shehu, M., Gjika, E. A Comprehensive Review of the Three Main Topic Modeling Algorithms and Challenges in Albanian Employability Skills. *European Scientific Journal*, **2024**, 20(12), 31–44.
12. Muchene, L., Safari, W. Two-stage Topic Modelling of Scientific Publications: A Case Study of University of Nairobi, Kenya. *PLOS ONE*, **2021**, 16(12), e0243208.
13. Liu, Q., Chen, Q., Ba, W., Shen, J., Wu, M., Sun, M., Ming, W-K. Data Analysis and Visualization of Newspaper Articles on Thirdhand Smoke. *JMIR Med Inform*, **2019**, 7(1), e12414.
14. Jelodar, H., Wang, Y., Yuan, C., Feng, X. Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. arXiv preprint arXiv:1711.04305. **2017**.
15. Buenaño-Fernández, D., Gonzalez, M., Gil, D., Luján-Mora S. Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. *IEEE Access*, **2020**, 8, 1–10.
16. Verma, M. K., & Yuvaraj, M. AI-Based Literature Reviews: A Topic Modeling Approach. *Journal of Information and Knowledge Management (JIKM)*, **2023**, 60(2), 97–104.
17. Röder, M., Both, A., Hinneburg A. Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, **2015**, pp. 399–408.
18. Wallach, H.M., Mimno, D.M., McCallum A. Rethinking LDA: Why Priors Matter. *Advances in Neural Information Processing Systems*, **2009**, 22, 1973–1981.

19. Grootendorst M. BERTopic: Neural Topic Modeling with Contextual Embeddings. arXiv preprint, **2022**.
20. El Sharkawy, G., Helmy, Y., Yehia, E. Employability Prediction of Information Technology Graduates Using Machine Learning Algorithms. *International Journal of Advanced Computer Science and Applications*, **2022**, 13(10), 0131043
21. Ministry of Finance and Economy. National Employment and Skills Strategy 2023–2030. Available online: https://arkiva.financa.gov.al/wp-content/uploads/2023/10/National-Employment-and-Skills-Strategy-2030_EN.pdf (Accessed on 05 August 2025)
22. Ministry of Education and Youth. National Youth Strategy 2022–2029. Available online: https://riniafemijet.gov.al/wp-content/uploads/2023/04/SKR29_Anglisht.pdf (Accessed on 05 August 2025)
23. Rieger, J., Jentsch, C., & Rahnenführer, J. Rolling. LDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data. *EMNLP Findings*, **2021**, pp. 2337–2347.
24. Boselli, C., Cesarini, M., Mercorio, R. and Mezzanzanica, M. Classifying Online Job Advertisements through Machine Learning. *Future Generation Computer Systems*, 2018, 86, 319-328.
25. Hossain, A., Karimuzzaman, M., Hossain, M.M., Rahman, A. Text Mining and Sentiment Analysis of Newspaper Headlines. *Information*, **2021**, 12, 414.
26. Kureková, L.M., Beblavý, M. & Thum-Thysen, A. Using online vacancies and web surveys to analyse the labour market: a methodological inquiry. *IZA J Labour Econ*, 2015, 4, 18.
27. European Commission: Eurostat, Beręsewicz, M. and Pater, R., Inferring job vacancies from online job advertisements – 2021 edition, Publications Office, 2021 https://data.europa.eu/doi/10.2785/963837
28. Askitas N., Zimmermann K.F. The Internet as a Data Source for Advancement in Social Sciences. International Journal of Manpower, 2015, 36(1), 2–12.
29. Albanian Institute of Statistics (INSTAT). Administrative data on labour market. Available online: https://www.instat.gov.al/media/14498/60administrative-data-on-labour-market_final_esms_en.pdf. (Accessed on 15 July 2025).
30. European Commission. ESCO. European Skills, Competences, Qualifications and Occupations (Version 1.2.0). Available online: https://esco.ec.europa.eu/en (Accessed on 7 July 2025).
31. Eurostat. Labour Market and Skills Indicators: Methodologies and Data Sources. Eurostat Technical Report, 2022. Available online: https://www.cedefop.europa.eu/files/esi_2022_technical_report.pdf (Accessed on 7 July 2025).
32. Silge, J., Robinson D. Text Mining with R: A Tidy Approach. O'Reilly Media: Sebastopol, CA, USA, **2017**.
33. Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge University Press: Cambridge, UK, **2008**.
34. Silge, J., Robinson, D. Tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *Journal of Open Source Software*, **2016**, 1(3), 37.
35. Porter, M.F. An Algorithm for Suffix Stripping. *Program*, **1980**, 14(3), 130–137.
36. Bouma G. Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, **2009**, pp. 31–40.
37. Spacy.io. Industrial-Strength Natural Language Processing in Python. Available online: https://spacy.io (accessed on 7 September 2025).

38. Griffiths, T.L., Steyvers, M. Finding Scientific Topics. *Proceedings of the National Academy of Sciences* USA, **2004**, 101(Suppl. 1), 5228–5235.
39. Nikita, M., Grün, B. Ldatuning: Parameters Tuning of Topic Models. R Package Version 1.0.2, 2019. Available online: https://cran.r-project.org/package=ldatuning (Accessed on 7 July 2025).
40. Wallach, H.M., Mimno, D.M., McCallum, A. Rethinking LDA: Why Priors Matter. *Advances in Neural Information Processing Systems*, **2009**, 22, 1973–1981.
41. Asuncion, A., Welling, M., Smyth, P., The, Y.W. On Smoothing and Inference for Topic Models. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, **2009**, pp. 27–34.
42. Griffiths, T.L., Steyvers, M. Probabilistic Topic Models. In Handbook of Latent Semantic Analysis; Landauer T., McNamara D., Dennis S., Kintsch W., Eds.; Psychology Press: Mahwah, NJ, USA, **2007**, pp. 427–448.
43. Deveaud, R., SanJuan, E., & Bellot, P. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, **2014**, 17(1), 61-84.
44. Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2010. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. **2010**, 6118.
45. Mezzanzanica, M., Boselli, R., Cesarini, M., Mercorio, F. and Moscato, V. Labour Market Intelligence through Text Mining: Building a Skills Forecasting System. *Computers in Industry* **2018**, 100, 275–285.
46. Liu, Y., Tang, J., Han, J., Jiang, M., & Yang, S. Mining topic-level influence in heterogeneous networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Toronto, Canada, **2016**, 199–208.
47. Heimerl, F., Lohmann, S., Lange, S. and Ertl, T. "Word Cloud Explorer: Text Analytics Based on Word Clouds," *2014 47th Hawaii International Conference on System Sciences*, Waikoloa, HI, USA, **2014**, pp. 1833-1842.
48. Newman M. Networks: An Introduction. Oxford University Press: Oxford, UK, **2010**.
49. Park, S., Kang, J., & Chung, S. Skill Network Analysis for the Future Labour Market. *Technological Forecasting and Social Change*, **2021**, 162, 120401.
50. Sievert, C. and Shirley, K. LDAvis: A Method for Visualizing and Interpreting Topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL Maryland, USA, **2014**, pp. 63–70.
51. Skills Forecast. Available online: https://milenashehu.shinyapps.io/SkillsForecastapp/ (Accessed on 20 July 2025).
52. Chuang J., Manning C.D., and Heer J. Termite: Visualization Techniques for Assessing Textual Topic Models. Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI) Umbria, Italy, **2012**, pp. 74–77.
53. Ali, G., Asiku, D., Mijwil, M. M., Adamopoulos, I., & Dudek, M. Fusion of Blockchain, IoT, Artificial Intelligence, and Robotics for Efficient Waste Management in Smart Cities. *International Journal of Innovative Technology and Interdisciplinary Sciences*, 2025, 8(3), 388–495.
54. Acemoglu, D., & Restrepo, P. Automation and New Tasks: How Technology Displaces and Reinstates Labour. *Journal of Economic Perspectives*, **2019**, 33(2), 3–30.