ESJ Natural/Life/Medical Sciences Peer-reviewed

A Comprehensive Review of the Three Main Topic Modeling Algorithms and Challenges in Albanian Employability Skills

Milena Shehu, PhD Student
University of Tirana, Albania
Eralda Gjika, Associate Professor
Royal College of Physicians and Surgeons of Canada, Ottawa, Canada

Doi:10.19044/esj.2024.v20n12p31

Submitted: 27 March 2024
Accepted: 09 April 2024
Published: 30 April 2024

Cite As:
Shehu M. & Gjika E. (2024). *A Comprehensive Review of the Three Main Topic Modeling Algorithms and Challenges in Albanian Employability Skills.* European Scientific Journal, ESJ, 20 (12), 31. https://doi.org/10.19044/esj.2024.v20n12p31

Abstract

Today's jobseekers face many obstacles while trying to find a career that aligns with their interests, employability soft skills, and professional experience. In Albania, jobseekers frequently initiate their job search by actively exploring job vacancies listed on various online job portals. The analysis of job vacancies posted online provides an added advantage to the labour market actors compared to traditional survey-based analyses. This is because it enables a faster analytical process, promotes decision-making based on accurate data, and should be carefully considered by every country when formulating their Labor Market Policies. Since the data posted online are unlabelled, it has been proven that the potential of unsupervised learning techniques, more precisely the Topic Modelling algorithms, is outstanding when applied to analysing job vacancies, mainly with regard to assessing employability soft skills. Algorithms in topic modelling are essential for uncovering hidden patterns in texts, facilitating the extraction of important data, generating document summaries, and enhancing content comprehension. This paper analyses and compares the three primary methodologies and algorithms used in topic modelling, which can be applied to analyse employability soft-skills: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and BERTopic. At the end of the paper, conclusions are drawn regarding superior performance and optimal algorithm applicability, challenges, and limitations through a review of studies conducted in the

Albanian job market.

Keywords: Topic modelling, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), BERTopic, Employability Skills

Introduction

In the context of employment and professional growth, skills are fundamental components that define individual capabilities and competencies, which contibutes to social development within career advancement.

International Standard Classification of Occupations 2008 (ISCO-08) and the European Framework of Skills, Competences, Qualifications, and Professions (ESCO) provide comprehensive structures for categorizing skills based on standardized taxonomies. ISCO-08 classifies skills into several main categories, including cognitive skills, technical skills, interpersonal skills, and practical skills. Broadly, skills are further defined to capture the nuances of occupations and different industries. For example, cognitive skills include analytical thinking, decision making, and problem-solving skills, while technical skills may include language programming, machine operation, and data analysis techniques (Tijdens Kea, 2019).

ESCO offers a comprehensive categorization of skills, competencies, qualifications, and professions across European nations, thereby establishing a shared framework for defining and classifying skills.

ESCO identifies skills in different areas such as knowledge, skills, abilities, and competences, which enables efficient communication and alignment of skills profiles within the European market of work. This approach facilitates targeted workforce development initiatives, competency-based education programs, and skills matching services, thus promoting mobility and employment across sectors and regions geographical (Chiarello et al., 2021).

Recently, machine learning (ML) algorithms have revolutionized analysis and assessment of skills on a large scale (Tufail et al., 2023). ML algorithms use clustered data to discover patterns, correlations, and predictive models related to the acquisition, use, and demand of skills. Several ML techniques are typically used to analyze and provide an overview of the most required skills, analyze new skills and job roles, and assist policy makers, educators, and employers in strategic decision making. These techniques include, but are not limited to: Algebraic Models like Latent Semantic Analysis, Probabilistic Models like Latent Dirichlet Allocation, and Neural Models like BERTopic (ElSharkawy et al., 2022; Varavallo et al., 2023). Skills are crucial for the development of human capital, competitiveness, and the achievement of economic prosperity.

ISCO-08 and ESCO provide comprehensive frameworks for the classification and categorization of skills, while the algorithms of Machine learning provide powerful tools for analysis, prediction, and harnessing the dynamics of skills within the evolving landscape of work and learning (Nikolaev, 2023) By maximizing the combined effectiveness and utilizing complementary strengths between skill taxonomies and ML techniques, stakeholders can promote inclusive growth, talent development, and innovation across global economies. Continuous research, collaboration, and investment in skill analytics and ML-driven insights are key to address the challenges and opportunities presented by rapid technological advances, demographic changes, and the evolving dynamics of the labor market (Djumalieva et al., 2018)

Moreover, in countries like Albania, in which occur undergoing transitions in their economic structures and experiencing challenges in integrating into global markets, skill analytics and machine learning-driven insights can play a crucial role in fostering innovation and competitiveness. By identifying emerging skill trends and areas of opportunity, stakeholders can guide strategic investments and policy interventions to drive economic diversification and sustainable development.

Literature Review

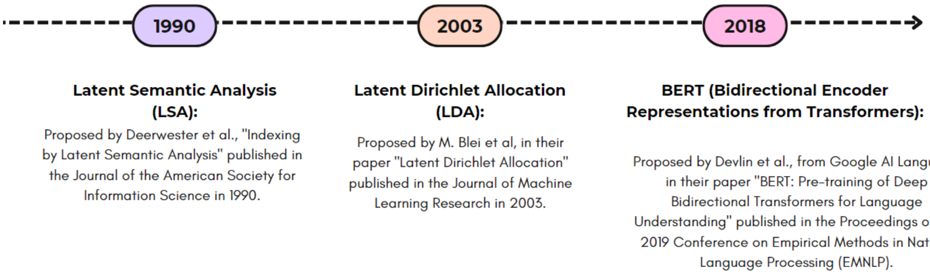
Topic modeling has recently emerged as a key tool in the field of Labor Market Intelligence, revolutionizing the process of finding and recruiting candidates (Abdelrazek et al., 2023). Labor Market Policy makers, eager to increase the efficiency of their employment efforts, have increasingly turned to these algorithms to distill large amounts of textual data into coherent information regarding what is happening in the country's labor market. This analytical approach goes beyond conventional keyword-based searches, providing recruiters with a nuanced understanding of the job requirements, qualifications of candidates, and industry trends. Outlining the semantic landscape of job postings, topic modeling empowers recruiters to identify salient topics, distinguish emerging skill sets, and align organizational needs with the candidates' skills. By discovering themes hidden within job advertisements, this analytical framework facilitates the identification of industry trends, competitive standards, and emerging skill requirements.

The proposal of Latent Semantic Analysis (LSA) was introduced by Deerwester et al. (1990). The objective of LSA is to represent the semantic content of words and documents in a high-dimensional space, enabling the analysis of relationships and similarities between them. It involves applying singular value decomposition (SVD) to a term-document matrix to capture the underlying latent semantic structure. This method enables tasks such as document similarity, information retrieval, and document classification by mapping words and documents to a semantic space where their relationships can be quantified.

Latent Dirichlet Allocation (LDA) was first proposed by Blei et al. (2003). It is a probabilistic generative model used for topic modeling. LDA assumes that documents are probabilistic mixtures of topics, and topics are probability distributions over words.

BERT (Bidirectional Encoder Representations from Transformers) was proposed by Devlin et al. (2019) from Google AI Language. It is a pre-training technique for natural language understanding based on transformers. BERT utilizes bidirectional training on large text corpora to generate deep contextualized word embeddings. Unlike traditional models that process words in isolation or in one direction, BERT captures the meaning of a word by considering its surrounding context from both directions in the text.

Figure 1. Presents the three algorithms in a chronological way, based on the year each of them was proposed.



Latent Semantic Analysis (LSA):
Proposed by Deerwester et al., "Indexing by Latent Semantic Analysis" published in the Journal of the American Society for Information Science in 1990.

Latent Dirichlet Allocation (LDA):
Proposed by M. Blei et al, in their paper "Latent Dirichlet Allocation" published in the Journal of Machine Learning Research in 2003.

BERT (Bidirectional Encoder Representations from Transformers):
Proposed by Devlin et al., from Google AI Language in their paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" published in the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Figure 1. Timeline of the three algorithms
Source: Authors (2024)

Table 1 below provides a summary of the methodologies employed and the findings derived from publications by different authors in the related field, which were published within the last five years. The PRISMA technique, as outlined by Moher et al. (2010), is utilized to draw inferences from the papers included in the table. Initially, a comprehensive search was conducted for academic papers published between 2018 and 2023 that employed LSA, LDA, and BERTopic algorithms across several domains. Based on the initial search, a total of 50 publications was examined. Among these papers, only those that involved these three algorithms in the field of employability and the ones that used job market advertisements as dataset were thoroughly assessed. The resulting papers are listed in the table below.

Table 1. Literature Review

Authors (year)	Title of the paper	Findings
Jyldyz Djumalieva, Antonio Lima, and Cath Sleeman, (Djumalieva et al., 2018)	Classifying occupations according to their skill requirements in job advertisements	The authors propose a methodology for developing an occupational classification. The approach entails the utilization of Natural Language Processing techniques, including document clustering and distributed word representations, on internet job postings in the United Kingdom.
Francis G. Balazon, Albert A. Vinluan, and Shaneth C. Ambat (Balazon et al., 2018)	Job Matching Platform Using Latent Semantic Indexing and Location Mapping Algorithms	This research project presents algorithms that utilize latent semantic indexing and location mapping algorithms to suggest appropriate work opportunities to job seekers. The algorithms are based on the collection and analysis of information, while also allowing employers to find eligible applicants for specific positions. The Latent Semantic Indexing algorithm is utilized to extract and portray the contextual usage and meaning of words over a document list through the application of statistical computation. The location mapping module is linked to the input of employers/recruiters regarding the location of jobs. Similarly, job seekers' input regarding desired work places is linked to "geocodes". According to the findings and examination of the created job matching platform, it has the capability to identify comparable job opportunities from a search query, allowing employers to assess job seekers without requiring human involvement.
Branislava Cvijetic and Zaharije Radivojevic (Cvijetic & Radivojevic, 2020)	Application of machine learning in the process of classification of advertised jobs	Traditional machine learning algorithms, such as Multinomial Naive Bayes and Support Vector Machine, have been used by the authors in order to classify advertised positions in accordance with ISCO-08 data. This classification was performed on a dataset that contained many languages.
Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso (Giabelli et al., 2021)	NEO: A System for Identifying New Emerging Occupation from Job Ads	NEO is a tool that gives authors the opportunity to demonstrate how it can automatically enhance the European Occupation and Skill Taxonomy (ESCO) with terms that reflect new professions that are taken from millions of Online Job Advertisements (OJAs).

Nina Golowko (Golowko, 2021)	The Improvement of Sustainable Employability Transfer in Higher Education Institutions Using Large Scale Data Bases and Machine Learning	This work presents and critically examines a proposed solution for enhancing the transfer of sustainable employability in higher education institutions. The proposed solution utilizes large-scale datasets and machine learning techniques. In the context of research on serious games, an AI software has been constructed and subsequently used to a corpus of theses, demonstrating an interdisciplinary approach. The underlying framework of this software is the Latent Dirichlet Allocation (LDA) topic model, which is employed for the purpose of categorizing extensive textual data and extracting themes from the dataset
Aiting Xu, Yuchen Wu, Feina Meng, Shengying Xu, and Yuhan Zhu (Xu et al., 2022)	Knowledge and skill set for big data professions: Analysis of recruitment information based on the Latent Dirichlet Allocation model	This study employed text mining, interviews, questionnaires, and LDA technique to examine the attributes and shortcomings of Chinese universities in the development of Big Data professionals, the expectations of enterprises regarding the professional competence of big data professionals, and the perceptions of students regarding the capabilities of big data professionals. Moreover, the present study employed the plan-action-inspection-action cycle theory to assess the talent development and quality management system of Big Data in China.
Ziqiao Ao, Gergely Horváth, Chunyuan Sheng, Yifan Song, and Yutong Sun (Ao et al., 2023)	Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions	The authors conducted a comparative analysis of the word-counting method using three distinct dictionaries, as well as three unsupervised topic-modeling algorithms, specifically LDA, PLSA, and BERTopic, on a dataset from a job board in the United Kingdom. The authors additionally suggested the utilization of the fraction of compensation variance accounted for by the retrieved abilities as an innovative performance criterion for comparing different techniques.
Giuseppe Varavallo, Giulia Scarpetti, and Filippo Barbera (Varavallo et al., 2023)	The moral economy of the great resignation	The objective of this study is to investigate the socio-psychological factors contributing to the Great Resignation and predict any future changes in individuals' perception of work in their lives. The BERTopic approach was employed to conduct a semantic analysis of 955 posts with high ratings from the r/antiwork subreddit. The purpose of this analysis was to examine identifiable subjects within the posts spanning from February 2020 to February 2022.

Source: Authors (2024)

Analysis Between Main Topic Modelling Algorithms

Topic Modeling is a frequently used approach which discovers hidden patterns semantics portrayed by a given text and automatically identifies the themes that exist within it. It can be defined as a type of statistical modeling that uses unsupervised machine learning to analyze and identify clusters of similar words within a body of text (Abdelrazek et al., 2023). In the section below, all the steps that should be followed when applying Topic Modelling Algorithms are described. The aim is to familiarize with the underlying principles, assumptions, and implementation details of each algorithm, providing insights into their strengths, limitations, and suitability for analyzing soft skills in job vacancies.

Preprocessing of Data

During the data preprocessing phase, various steps are undertaken including normalization, tokenization, removal of punctuation and special characters, elimination of stop words, stemming, and the construction of a corpus. These main preprocessing steps can be described as follows:

- **Normalization:** Transforming the file into a uniform format by converting the characters to lowercase, removing punctuation, and eliminating unnecessary words.
- **Tokenization:** The given text is broken into sentences, which are further broken down into words or tokens. Usually, tokens are bounded in space by each other.
- Removing words that have less than 3 characters with no meaning, punctuation, and special characters (such as eliminating non-alphanumeric characters from the text).
- **Removal of stop words:** In this step, all common words that do not carry significant meaning are removed from the text, such as, conjunctions, pronouns, etc.
- **Stemming (Rooting):** The aim of this phase is to reduce words to their root or base form to capture variations of the same word. The root has semantic meaning, and this process helps in increasing the performance of the model.

Building the Corpus

At this stage, the selection of a method facilitates the extraction of themes and determines their relevance for different objectives. This paper concentrates on three prominent models within this domain: LDA, LSA, and BERTopic. Subsequently, the three methods are briefly described and their respective strengths and applications in topic modelling analysis are discussed.

- **LSA -** Latent Semantic Analysis (LSA) is a method in natural language processing that examines the connections between documents and the vocabulary they encompass (Deerwester et al., 1990). LSA is mainly utilized for concept searching and automatic document classification. It is an unsupervised learning approach. There is a definite purpose to it, yet no tags are assigned. “Latent” itself means hidden. Therefore, it is necessary to hide or encapsulate data within itself (Balazon et al., 2018). Thus, LSA assumes that words with similar meaning appear in similar documents. This method is realized by building a matrix which contains the number of words for each document. Each row represents a separate word, and each column represents each document. Furthermore, a Singular Value Decomposition (SVD) is used to reduce the number of rows while maintaining the similarity structure between columns. SVD is a mathematical method that simplifies the data while preserving its special characteristics. In this technique, SVD is used to maintain relationships between columns and rows. The similarity between documents is determined by cosine similarity, where the cosines of the angles between the two vectors, which in this case represent the documents, are calculated.

a) If the value approaches 1, it indicates that the documents share a high degree of similarity based on their contained words.
b) If the value is close to 0, it suggests that the documents are highly dissimilar.

- **LDA** – Latent Dirichlet Allocation is a probabilistic model where the document contains hidden topics and each topic has a distribution of words (Blei et al., 2003). LDA represents topics with word probabilities. This unsupervised algorithm is the most widely used and assumes that each document is represented as a probabilistic distribution of hidden topics. In LDA, the distribution of topics in the document and the distribution of words in the topic are independent of each other (ElSharkawy et al., 2022). Therefore, the same words can appear with different frequencies in different topics, or the same topics can appear in different documents. This assumption is based on the probabilistic Bayesian model (Tufail et al., 2023).
- **BERTopic** – This is a neural model with the inclusion of transformers (Devlin et al., 2019). Models like LDA or NMF (Non-Negative Matrix Factorization) ignore the semantic connections that may exist between words. In response to this problem, new techniques have been developed, such as Bidirectional Encoder Representations from Transformers (BERT) (Sawant et al., 2022). These techniques have been used for classification or neural search engines and are also applied in Topic Modelling (Devlin et al., 2019). This framework was created by Google in 2018 and is open-source. BERT is based on the transformer, a "deep learning" model where each output element is connected to each input element, and the weights

of the connections are calculated and dynamically changed. The transformer processes the word in relation to all the other words in the sentence. This increases the capacity to understand the context of the word. This approach has led to significant advancements in various natural language processing tasks, such as question answering, sentiment analysis, and named entity recognition. BERT's pre-trained models are widely used and fine-tuned for specific downstream tasks, demonstrating state-of-the-art performance in many benchmarks (Zhang et al., 2019).

Model Evaluation

At this stage, statistics is generated to determine the best model. Various packages and libraries in different coding systems and program development environments provide functions and built-in methods that facilitate the extraction of these statistics (Xu et al., 2022). Once the statistics are calculated, additional techniques such as cross-validation or hyperparameter tuning may be employed to further assess and optimize the performance of the models. There are several measures that can be employed to assess the performance and effectiveness of each algorithm, even though different methods perform better in specific situations. Various evaluation metrics can be categorized into different dimensions, including measures of quality, interpretability, stability, diversity, efficiency, and flexibility, thus providing comprehensive insights into the performance of models across multiple aspects.

- Perplexity: This is commonly used in LDA and measures how well the model predicts a held-out set of documents. Lower perplexity values indicate better performance (Abdelrazek et al., 2023).
- Coherence Score: This measure assesses the interpretability of topics generated by the model. Higher coherence scores suggest more coherent and semantically meaningful topics. It is commonly used in topic modeling algorithms, such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) (Kherwa & Bansal, 2018).
- Semantic Similarity: This is particularly relevant for BERT. Semantic similarity metrics, such as cosine similarity or Euclidean distance, can be used to evaluate the similarity between documents or sentences. Higher similarity scores indicate better semantic representation (Alcoforado et al., 2022).
- Reconstruction Error: In LSA, reconstruction error measures the difference between the original input data and the data reconstructed by the model. Lower reconstruction error values signify better reconstruction accuracy (Kherwa & Bansal, 2018).
- Classification Accuracy: In supervised tasks, such as document classification, accuracy metrics can be used to evaluate the model's

performance in correctly classifying documents into predefined categories. Classification accuracy is commonly used in algorithms, such as support vector machines (SVM), decision trees, random forests, neural networks, and other supervised learning models. It is not directly applicable to unsupervised algorithms like Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), which do not involve explicit classification tasks. However, in some cases, LDA topics can be used as features for classification tasks, and in such scenarios, classification accuracy could be applied (Abdelrazek et al., 2023).

- Word Embedding Evaluation: For BERT, LSA, and similar models, word embedding evaluation techniques, such as word analogy tasks or word similarity tasks, can assess the quality of word representations learned by the model (Zhang et al., 2019).

These measures provide insights into different aspects of the model's performance, including its ability to generate coherent topics, capture semantic relationships, and accurately represent the underlying data distribution. There has been significant development in the discussion of performance evaluation measures and the exploration of their pros and cons within the context of algorithms such as LSA, LDA, and BERT (Koehn & Knowles, 2017).

A. Comparative Review of Three Algorithms

In this section, authors aim to provide a comprehensive review of the advantages and disadvantages associated with the application of these methods in analyzing soft skills within job vacancies. By examining the strengths and limitations of each approach, this study seeks to offer insights into the effectiveness and suitability of these methods to extract and understand soft skills requirements in job postings.

Advantages and Disadvantages

Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and BERTopic are the most widely used algorithms in natural language processing (NLP) (Mankolli & Bushati, 2023) for topic modeling and document analysis. Despite sharing the common goal of discovering latent structures within textual data, they use different techniques and assumptions. In this comparative analysis, authors explore the operation of each, highlighting their strengths and weaknesses to determine which algorithm might be best suited for the specific task, such as analyzing soft skills from a dataset containing job vacancies. Generally, the advantages and disadvantages of the three algorithms are highlighted in the table below (Tufail et al., 2023; Kalepalli et al., 2020)

Table 2. Advantages and Disadvantages of LDA, LSA, and BERTopic

Algorithm	Advantages	Disadvantages
LSA (1990)	- It's intuitive - Can be applied to both short and long documents - Through the V matrix, topics are open to human interpretation - pLSA is more performant, compared to LSA	-The Document-Term matrix ignores the semantic representation of words and treats similar words as different elements of the matrix. -Preprocessing techniques can be helpful, but only to some extent. For example, "Albania" and "Albanian" would be considered similar, but not "money" and "cash". - LSA requires an extended preprocessing stage to obtain meaningful representations from textual inputs. - The number of topics must be known in advance
LDA (2003)	-Coherence between topics is found, and hyperparameter tune is applied - Number of topics generated is easily interpretable - There is mixed membership of data where one document can contain different topics - Coherence and perplexity is better in LDA than in LSA - Data can be analyzed by using R, and it does not require Python programming language.	- Assumptions must be detailed beforehand. Care must be taken when determining hyperparameter tunning - Output may show overlap of topics, since that same topic may be part of different documents - Number of topics needs to be pre-defined - Reliability and validity are not automatically assured
BERTopic (2019)	-The generated model is more stable - Multi-lingual analysis is supported -Number of topics is automatically generated and not pre-defined as in LDA - No pre-processing of the original data is needed because embeddings are used - Requires Python programming language from the data scientist	-Requires inspection of each generated topic as the embedding process can produce too many topics. - Many outliers are generated - It is not possible to generate objective evaluation metrics.

Source: Authors (2024)

A major difference between LDA and LSA lies in their interpretability. Both LDA and BERTopic produce a clear topic distribution for each document (in this case, for each job vacancy), enabling users to understand the thematic composition of textual data. On the other hand, LSA lacks a clear notion of topics and may not provide intuitive insight into the underlying structure of documents. However, LSA's ability to capture semantic similarities between words and documents makes it valuable for tasks, such as document clustering and query expansion (Koehn & Knowles, 2017)

When compared to LDA and LSA, BERTopic is distinguished by the fact that it offers continuous topic modeling as opposed to discrete topic modeling (Alcoforado et al., 2022). As a result of the stochastic character of the model, findings that are obtained through repeated modeling are varied. After the model has been computed, it is able to retrieve the most significant subjects. In addition, Topic 0, with a count of -1, will always be considered outliers and should not be taken into consideration any further. Also, BERTopic have the ability to search for a keyword and receive the topics that are the most significant based on the similarity score between them. Additionally, it is possible to examine specific subjects depending on the keywords they include. In the end, BERTopic provides an interactive inter-topic distance map for the purpose of evaluating individual topics (Sawant et al., 2022) . This is done in order to improve the analysis of the potentially vast diversity of topics.

In terms of scalability, LDA tends to perform better with large corpora due to its parallelizable nature and efficient inference algorithms. LSA, based on SVD, can encounter computational challenges when dealing with massive data sets or when trying to process in real time. On the other hand, transformer-based language models used by BERTopic enables the creation of topic representations that are more accurate and coherent (Zhang et al., 2019). These representations are based on the semantic similarity of words and phrases. These results improve subject coherence and diversity, particularly when dealing with documents that span multiple domains and are relatively brief.

Comparison in Analysing Employability Soft Skills

When applied to analyze employability soft skills, the ability of LDA and BERTopic to detect topics within job descriptions can help recruiters and job seekers quickly identify key topics or areas of interest required for specific roles (Ao et al., 2023). By aggregating job postings based on their thematic distributions, LDA and BERTopic can help organize and categorize job listings, making it easier to navigate through a large volume of job vacancies (Golowko, 2021).

When considering the effectiveness between these three algorithms, the choice depends on the specific objectives of the NLP (Natural Language

Processing) task (Tufail et al., 2023). If the goal is to analyze and interpret soft skills captured in a dataset of job vacancies, a task that require interpretable topic modeling and nuanced understanding of document topics, LDA and BERTopic provide a more appropriate framework (ElSharkawy et al., 2022). Conversely, LSA excels in applications focused on semantic analysis, document similarity, and information retrieval, where the emphasis is on capturing semantic relationships between words and documents (Deerwester et al., 1990). As such, because LSA's focus is on capturing semantic relationships between words and documents, it can help match job descriptions with candidate CVs and would be used to identify relevant CVs by measuring the similarity between the job description and candidate profiles based on their semantic representations.

In conclusion, while these three algorithms provide valuable insights into textual data, their divergent approaches and capabilities meet different analytical needs at different times.

Also, the choice between LSA, LDA, and BERTopic depends on the specific requirements of the NLP task at hand, with neither algorithm universally superior, but rather complementing each other in the broader landscape of text analysis and understanding. In recent times, there has been a growing emphasis on big data and the availability of work opportunities on online job portals. As a result, academics are increasingly using LDA and BERTopic, which are advanced Natural Language Processing algorithms, to make valuable contributions to the subject of employability skills. Conversely, this will produce a more accurate and interpretable output (Tufail et al., 2023)

Applicability in the Albanian Job Market

Review of Literature

In recent years, researchers and labor market specialists in Albania have launched numerous initiatives to study employability trends and address the needs of the labor market. Additionally, public employment institutions in Albania have implemented various policies, programs, and projects aimed at enhancing youth employability (Minister of State for Youth and Children in Albania, 2022). The studies presented in the same period as the other international studies were selected to maintain consistency in applications of the methods.

An extensive literature review was conducted to analyze the different factors and reasons behind youth unemployment (Fejzulla, 2021). Additionally, the paper aimed to analyze the vocational education and training system in Albania through statistical analysis. Moreover, this article proposes a discussion on strategies to enhance youth employment by identifying the specific skills demanded by the private sector labor market and exploring ways to enhance them through vocational education and training.

The study of Kraja and Boriçi (2021) aimed to identify the impact of hard skills and soft skills on employability. In this paper, a questionnaire was administered to different employers, both private and state employers in Albania, and analyses were perfomed not through the use of topic modelling algorithms, but through statistical processing.

Another approach presented by Shehu and Stringa (2024) aimed to emphasize the present national policies and programs aimed at fostering youth employment, as well as initiatives for training and coaching programs for unskilled young individuals. The paper highlighted the effects of national measures on guiding jobseekers toward improving their employability skills and analyzing effects on reducing unemployment and migration rates in Albania. However, this paper was solely theoretical and did not provide any further analysis using statistics or machine learning algorithms.

Moreover, Fetahu and Lekli (2023) conducted a study with the objective of providing an extensive viewpoint on the field of entrepreneurship in Elbasan. Their research aimed to enhance comprehension of particular skills and training requirements for jobseekers in this domain. The study included a comprehensive analysis of the data obtained from a thorough and well-designed questionnaire, which was administered to a sample of 39 companies across 5 important sectors. In this paper, analysis was not conducted through machine learning algorithms.

The techniques of NLP to forecast the likelihood of a candidate's success in a job vacancy were applied by Mankolli and Bushati (2023). The dataset used in this article includes 648 curriculum vitae of jobseekers. To forecast the likelihood of a job posting being successful, the researchers applied the XGBoost Classifier model, which is a form of gradient-boosting decision trees commonly used in Supervised Learning methods. The primary finding of the paper indicates that employing NLP approaches to analyze text and speech data improves the ability to distinguish between candidates and further enhances the accuracy of predicting job success.

In reference to the article of Çano and Lamaj (2024), the scarcity of text corpora for low-resource languages like Albanian poses a significant challenge for natural language processing research. However, the introduction of AlbNews, comprising 600 topically labeled news headlines and 2600 unlabeled ones in Albanian, offers a valuable resource for conducting topic modeling research. This dataset can be freely utilized to explore various natural language processing tasks. Initial classification scores of traditional machine learning classifiers trained with AlbNews samples are reported in the article. Interestingly, the results indicate that basic models outperform ensemble learning methods, thus establishing a baseline for future experiments in this domain.

Overall, it can be concluded that there are many initiatives to analyze employability skills in Albania. Nonetheless, Mankolli and Bushati (2023) employ Topic Modelling techniques, which are the main algorithms of the Unsupervised Machine Learning used when dealing with Natural Language Processing Data (NLP).

Challenges of Using LSA, LDA, and BERT in Albanian Language

Despite the initiatives to analyze employability skills in Albania, the utilization of advanced topic modeling algorithms like LSA, LDA, and BERT in Albanian language poses several challenges. While these algorithms hold promise for extracting insights from Albanian text data, addressing these challenges will be crucial for their effective application in understanding and addressing employability trends in the region.

Below, several potential challenges encountered when employing these methodologies (Koehn & Knowles, 2017) in the analysis of Albanian text are outlined:

Lack of Training Data: One of the primary challenges is the scarcity of annotated training data in Albanian language. Building robust models like LSA, LDA, and BERT requires large amounts of labeled data, which may be limited for Albanian

Language Complexity: Albanian language exhibits its own linguistic complexities, including morphology, syntax, and semantics, which may pose challenges for models trained on languages with different linguistic structures.

Model Adaptation: Pre-trained models like BERT are typically trained on large corpora of English text. Adapting these models to Albanian language may require extensive fine-tuning on Albanian-specific data, which can be resource-intensive and time-consuming. Packages like SentiBert and KeyBERTInspired will be incorporated in order to provide better extraction feature and fine-tunning process. When using LDA algorithm for the hyperparameter tunning, attention should be given to the adjustment of the Number of Topics (K), Dirichlet hyperparameter alpha: Document-Topic Density and the Dirichlet hyperparameter beta: Word-Topic Density, accordingly. Since the Albanian Language is complex, it is expected that the accuracy and reliability of the results will be affected. For this reason, the approach of translating the dataset into english language is also considered, aiming to obtain more accurate results.

Domain Specificity: Employability skills and job vacancies in Albania may have specific linguistic nuances and terminologies that differ from those in English or other languages. Adapting topic modeling algorithms to capture these nuances accurately can be challenging.

Evaluation and Validation: Assessing the performance of LSA, LDA, and BERT models in the context of Albanian language may require developing custom evaluation metrics and validation procedures tailored to the characteristics of Albanian text data. As a first approach, using different accuracy measures will enable a more accurate overview of the complexity of the Albanian language and the performance of the proposed methods. Some of the measures include visualization of inter-topic distance map and calculation of the coherence score in order to evaluate and validate the model. When applying LDA algorithm, an important measure to consider is the perplexity score based on the performance and accuracy of the model.

As an Indo-European Language, Albanian language differs from many other languages. Therefore, it may be considered as a Low Resource Language, which is difficult to analyze through Natural Language Processing techniques. When applying LDA and BERTopic algorithms with data in Albanian Language, especially in the pre-processing phase, the aim is to apply packages that focus on sentiment analysis and the narrative morphology of the text, enabling the study of emotional and thematic structures. If the outcome of using these packages in Albanian language data is not satisfactory, then consideration will be given to translating web-scrapped data into English language, in order to provide better outcomes and more interpretable analysis regarding employability soft skills.

This research paper will be followed by a study which will apply LDA and BERTopic algorithms in a dataset with job vacancies published in Albanian companies or institutions job portals. The data are gathered using web scraping, and the pre-processing and cleaning process will follow before the data analysis and accuracy of the methods applied. However, one challenge is the application of the methods in situations where the Albanian language is used for job posting and no translating option is offered by the webpage. In these cases, possible applications in non-English language are reviewed and their efficiency is monitored.

Employability soft skill in Albania is a new field of study due to recent developments that have affected the labor market and new professions that have appeared in recent years, not only in Albania but internationally. These professions that have been embraced by the labor market require adaptation and clarity in the description of the work and tasks that professionals in the fields must cover. Therefore, analyzing soft skills employability in a small country like Albania, faced with many new professions in the last years and ongoing improvements in online job posting tools, poses multifaceted challenges. There is also a limited historical amount of data on soft skills demand that affects the assessment and match of skills to their corresponding job roles. Human Resources teams within companies or large public institutions must navigate into the dynamics of soft skills job market, ensuring that the skills align with the rapid change of recruitment needs and technological advancements. Development of comprehensive competency frameworks and standardization of evaluation methods becomes crucial for recruitment agencies in Albania. Therefore, this ongoing study, which will use real data obtained from the webpage of companies or institutions through webscraping, will enable the creation of a general overview of soft skills employability topic modeling that can be used by employment institutions in Albania. The output will support the revision of the National Employment and Skills Strategy, as well as other national initiatives for skills development that will be undertaken by the government of Albania in the future.

Conclusion

This paper aimed to compare Topic Modelling Algorithms, which can be used to analyze employability soft skills – a consideration crucial for every country when formulating their Labor Market Policies as it helps to increase the efficiency of recruitment and selection of candidates. There are three main topic modelling algorithms that can be used in these types of analyses, which produce accurate and interpretable information about employability soft skills. Each of these algorithms clearly presents challenges in the process of analyzing soft skills data. Although LSA (Latent Semantic Analysis) models are simple, they lack a strong statistical basis and do not define a proper model. However, they are relatively efficient.

Models created through LDA (Latent Dirichlet Allocation) are intuitive and improvable. They define a generative process related to the documents and attempt to determine the topics. Parameters in probabilistic models tend to be more understandable and easily interpretable. In other words, if the generated results is interpreted easily, more model errors can be easily found.

Neural models, like BERTopic, are quite flexible, but the parameters can have minor problems in interpretation. It is usually difficult to investigate why the model works or not. BERT is a better model because it takes into account the context in which the word is located and considers words that are similar to generate topics, unlike LSA and LDA.

The utilization of these methodologies in analyzing soft skills within the Albanian language job market presents both opportunities and challenges. While these methods offer powerful tools for extracting insights from text data, their application in the context of Albanian language text presents challenges, such as the scarcity of annotated data, linguistic complexities, and the need for model adaptation. However, the potential benefits of employing these methodologies, including enhanced understanding of employability trends and improved decision-making in workforce development, make overcoming these challenges worthwhile.

Moving forward, the next step in this article involves the extraction of information from web sources, such as job vacancy postings on online portals in Albania. Web scraping will be used to extract information from companies or institutions webpage and the cleaning process will purposefully increase the quality and chances of applying the methodologies to Albanian data. Additionally, efforts will be made to modify these methodologies to better support the Albanian lexicon, including fine-tuning models on Albanian-specific data and developing custom evaluation metrics. Furthermore, exploring translation methods that enable the direct use of these methodologies in popular software like R, Python, and other open-source tools will be pursued to facilitate wider adoption and accessibility. These advancements aim to enhance the applicability and effectiveness of text analysis techniques in supporting labor market research and policy development efforts in Albania.

Conflicts of Interests: The authors declare that they have no competing financial, professional, or personal interests that could have influenced the conduct or reporting of this research.

Data Availability: All data are included in the content of the paper.

Funding: The research reported in this article was conducted without external funding.

Author Contributions

Milena Shehu conceptualized the study and drafted the initial manuscript. Shehu and Eralda Gjika contributed to the methodology design. Shehu conducted data collection and analysis, while Gjika provided supervision. All authors reviewed and edited the manuscript.

References:

1. Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. Information Systems, 112, 102131. https://doi.org/10.1016/j.is.2022.102131
2. Alcoforado, A., Ferraz, T. P., Gerber, R., Bustos, E., Oliveira, A. S., Veloso, B. M., Siqueira, F. L., & Costa, A. H. R. (2022). ZeroBERTo: Leveraging Zero-Shot Text Classification by Topic Modeling (pp. 125–136). https://doi.org/10.1007/978-3-030-98305-5_12
3. Ao, Z., Horváth, G., Sheng, C., Song, Y., & Sun, Y. (2023). Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions. Information Processing & Management, 60(2), 103185. https://doi.org/10.1016/j.ipm.2022.103185

4. Balazon, F. G., Vinluan, A. A., & Ambat, S. C. (2018). Job Matching Platform Using Latent Semantic Indexing and Location Mapping Algorithms. Asia Pacific Journal of Multidisciplinary Research, 6(4). www.apjmr.com
5. Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In Journal of Machine Learning Research (Vol. 3).
6. Boriçi Kraja, Y. & Albana Begani Boriçi, A. (2021). Enhancing employability skills valued by employers-Case of Albania. Academic Journal of Business, 7(3). www.iipccl.org
7. Çano, E. & Lamaj, D. (2024). AlbNews: A Corpus of Headlines for Topic Modeling in Albanian. http://arxiv.org/abs/2402.04028
8. Chiarello, F., Fantoni, G., Hogarth, T., Giordano, V., Baltina, L., & Spada, I. (2021). Towards ESCO 4.0 – Is the European classification of skills in line with Industry 4.0? A text mining approach. Technological Forecasting and Social Change, 173, 121177. https://doi.org/10.1016/j.techfore.2021.121177
9. Cvijetic, B. & Radivojevic, Z. (2020). Application of machine learning in the process of classification of advertised jobs. IJEEC - INTERNATIONAL JOURNAL OF ELECTRICAL ENGINEERING AND COMPUTING, 4(2). https://doi.org/10.7251/IJEEC2002093C
10. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990a). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
11. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990b). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
12. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North, 4171–4186. https://doi.org/10.18653/v1/N19-1423
13. Djumalieva, J., Lima, A., & Sleeman, C. (2018). Classifying Occupations According to Their Skill Requirements in Job Advertisements. www.escoe.ac.uk.
14. ElSharkawy, G., Helmy, Y., & Yehia, E. (2022). Employability Prediction of Information Technology Graduates using Machine Learning Algorithms. International Journal of Advanced Computer Science and Applications, 13(10). https://doi.org/10.14569/IJACSA.2022.0131043

15. Fejzulla, P. E. (2021). Increasing Youth Employability in Albania by Enhancing Skills through Vocational Education. European Journal of Economics and Business Studies, 7(2), 12. https://doi.org/10.26417/685lur76k
16. Fetahu, E. & Lekli, L. (2023). Developing Soft Skills, the Intangible Qualities Empowering Competitiveness and Success in the Labor Market, Case Study, Elbasan, Albania. WSEAS Transactions on Business and Economics, 20, 965–976. https://doi.org/10.37394/23207.2023.20.89
17. Giabelli, A., Malandri, L., Mercorio, F., Mezzanzanica, M., & Seveso, A. (2021). NEO: A System for Identifying New Emerging Occupation from Job Ads. Proceedings of the AAAI Conference on Artificial Intelligence, 35(18), 16035–16037. https://doi.org/10.1609/aaai.v35i18.18004
18. Golowko, N. (2021). The Improvement of Sustainable Employability Transfer in Higher Education Institutions Using Large Scale Data Bases and Machine Learning (pp. 165–185). https://doi.org/10.1007/978-3-658-33997-5_6
19. Kherwa, P. & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. ICST Transactions on Scalable Information Systems, 0(0), 159623. https://doi.org/10.4108/eai.13-7-2018.159623
20. Koehn, P. & Knowles, R. (2017). Six Challenges for Neural Machine Translation. Proceedings of the First Workshop on Neural Machine Translation, 28–39. https://doi.org/10.18653/v1/W17-3204
21. Mankolli, E. & Bushati, S. (2023). Candidate Engagement Success Prediction Using Machine Learning and Natural Language Processing Techniques. 2023 24th International Conference on Control Systems and Computer Science (CSCS), 431–435. https://doi.org/10.1109/CSCS59211.2023.00074
22. Minister of State for Youth and Children in Albania (2022). National Youth Strategy and Action Plan 2022-2029.
23. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. International Journal of Surgery, 8(5), 336–341. https://doi.org/10.1016/j.ijsu.2010.02.007
24. Nikolaev, I. (2023). An intelligent method for generating a list of job profile requirements based on neural network language models using ESCO taxonomy and online job corpus. Business Informatics, 17(2), 71–84. https://doi.org/10.17323/2587-814X.2023.2.71.84
25. Sawant, S., Yu, J., Pandya, K., Ngan, C.-K., & Bardeli, R. (2022). An Enhanced BERTopic Framework and Algorithm for Improving Topic Coherence and Diversity. 2022 IEEE 24th Int Conf on High

Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), 2251–2257. https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys57074.2022.00332

26. Shehu Milena & Stringa Areti (2024). National measures undertaken to improve youth employability and further develop employability skills in Albania. CIDE Conference 14-21. https://upg-elearning.ro/cide23/about-the-conference/conference-proceedings/
27. Tijdens Kea (2019). Measuring job tasks by ISCO-08 occupational group.
28. Tufail, S., Riggs, H., Tariq, M., & Sarwat, A. I. (2023). Advancements and Challenges in Machine Learning: A Comprehensive Review of Models, Libraries, Applications, and Algorithms. Electronics, 12(8), 1789. https://doi.org/10.3390/electronics12081789
29. Varavallo, G., Scarpetti, G., & Barbera, F. (2023). The moral economy of the great resignation. Humanities and Social Sciences Communications, 10(1), 587. https://doi.org/10.1057/s41599-023-02087-x
30. Xu, A., Wu, Y., Meng, F., Xu, S., & Zhu, Y. (2022). Knowledge and Skill Sets for Big Data Professions: Analysis of Recruitment Information Based on The Latent Dirichlet Allocation Model. Www.Amfiteatrueconomic.Ro, 24(60), 464. https://doi.org/10.24818/EA/2022/60/464
31. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. http://arxiv.org/abs/1904.09675