applied sciences

MDPI

Article

Enhancing Skills Demand Understanding through Job Ad Segmentation Using NLP and Clustering Techniques

Mantas Lukauskas [1,*], Viktorija Šarkauskaitė [2], Vaida Pilinkienė [3], Alina Stundžienė [3], Andrius Grybauskas [3] and Jurgita Bruneckienė [3]

[1] Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences, Kaunas University of Technology, 44249 Kaunas, Lithuania
[2] Independent Researcher, 51297 Kaunas, Lithuania
[3] School of Economics and Business, Kaunas University of Technology, 44249 Kaunas, Lithuania
* Correspondence: mantas.lukauskas@ktu.lt

Abstract: The labor market has been significantly impacted by the rapidly evolving global landscape, characterized by increased competition, globalization, demographic shifts, and digitization, leading to a demand for new skills and professions. The rapid pace of technological advancements, economic transformations, and changes in workplace practices necessitate that employees continuously adapt to new skill requirements. A quick assessment of these changes enables the identification of skill profiles and the activities of economic fields. This paper aims to utilize natural language processing technologies and data clustering methods to analyze the skill needs of Lithuanian employees, perform a cluster analysis of these skills, and create automated job profiles. The hypothesis that applying natural language processing and clustering in job profile analyzes can allow the real-time assessment of job skill demand changes was investigated. Over five hundred thousand job postings were analyzed to build job/position profiles for further decision-making. In the first stage, data were extracted from the job requirements of entire job advertisement texts. The regex procedure was found to have demonstrated the best results. Data vectorization for initial feature extraction was performed using BERT structure transformers (sentence transformers). Five dimensionality reduction methods were compared, with the UMAP technique producing the best results. The HDBSCAN method proved to be the most effective for clustering, though RCBMIDE also demonstrated a robust performance. Finally, job profile descriptions were generated using generative artificial intelligence based on the compiled job profile skills. Upon expert assessment of the created job profiles and their descriptions, it was concluded that the automated job advertisement analysis algorithm had shown successful results and could therefore be applied in practice.

Keywords: clustering; natural language processing; NLP; jobs requirements; machine learning; generative AI; GPT

Citation: Lukauskas, M.; Šarkauskaitė, V.; Pilinkienė, V.; Stundžienė, A.; Grybauskas, A.; Bruneckienė, J. Enhancing Skills Demand Understanding through Job Ad Segmentation Using NLP and Clustering Techniques. *Appl. Sci.* **2023**, *13*, 6119. https://doi.org/10.3390/app13106119

Academic Editors: Jerry Chun-Wei Lin, Gautam Srivastava and Stefania Tomasiello

Received: 1 April 2023
Revised: 13 May 2023
Accepted: 14 May 2023
Published: 16 May 2023

1. Introduction

Economic and geopolitical changes lead to increasing competition, globalization, demographic challenges, and digitization in almost all labor market areas. Digitization poses many challenges to businesses and employees regarding tasks previously performed by people, such as data entry, accounting, and work on the conveyor belt, which are now supported by digital technologies and artificial intelligence [1,2]. Digitization is predicted to only increase from here, meaning the challenges will also increase. For example, in the US, 47% of jobs are predicted to be automated in the coming decades [3]. In Lithuania, it is also noticeable that automation is increasing, and 40% of jobs already face significant changes after its introduction [4]. From an economic perspective, automation reduces company costs. It increases production, efficiency, and productivity, resulting in fewer people needed to achieve company goals.

However, due to the increasing digitization, another challenge arises—the demand for new professions and skills [5]. Automating a significant part of the employees' work creates a problem, as companies lack qualified employees with the appropriate skills to implement and maintain the latest technologies in the organization. There has been a significant shortage of skilled cyber security or data analysis workers. With the popularity and development of artificial intelligence, companies are increasingly choosing and using this tool in their business, meaning therefore more new jobs are emerging in IT, robotics, and industry. Technological changes already make it possible to see that some workers in the market have the necessary skills and can adapt more easily to the labor market.

In contrast, others lack the necessary skills [6]. We call skills the knowledge, qualities, and abilities of a person that can be learned. Learned knowledge, abilities, and acquired qualities enable people to successfully and systematically perform assigned tasks or activities [7]. Skills and their development are particularly important in enabling a country and its people to adapt and function successfully in an ever-changing world and labor market. Those who acquire strong, necessary skills are more innovative, efficient, confident, and have a higher quality of life.

In most cases, skill analysis is conducted through regular surveys. However, in this case, they can only be conducted during a certain time period and observing the dynamic changes in skills is difficult. One solution to better understand these required skills is the application of artificial intelligence methods in analyzing these skills. Developing artificial intelligence methods allow for extraction, process, and interpreting dynamic skill needs, whilst also enabling these processes' extremely fast and automated performance.

In recent years, the analysis of job requirements has received much attention—machine learning methods are increasingly used to extract and classify valuable information from job advertisements. Identifying key job requirements such as education, experience, skills, certifications, language proficiency, physical requirements, background checks, availability, personal characteristics, and legal eligibility can provide valuable insights for job seekers, employers, and policymakers [8]. Natural language processing techniques and machine learning algorithms can effectively classify and analyze the various requirements presented in job postings, thereby offering a data-driven approach to understanding the changing needs of the labor market [9]. For example, text mining and machine learning techniques have been used to study the prevalence of specific skills and qualifications in job postings, revealing trends in employer preferences and skill gaps across industries [10]. In addition, job requirement analyzes can help match job seekers with suitable employment opportunities, leading to more effective labor market outcomes and a better workforce development [11].

Despite the growing interest in analyzing job requirements using machine learning and natural language processing techniques, there remains a research gap in understanding the real-time assessment of job skill demand changes in the Lithuanian labor market. Our study aims to use natural language processing technologies and data clustering to analyze the skill needs of Lithuanian employees, perform a cluster analysis of these skills, and create new employee profiles. The primary objectives of this research are to explore the dynamics of specific skills, compare the different clustering methods, and identify the main profiles of employees in the Lithuanian labor market. The contributions of this study include providing valuable insights into the changing needs of the Lithuanian job market and offering a data-driven approach for assessing job skill demand changes in real time.

The remainder of this paper is organized as follows: Section 2 introduces the concept of employee requirements, definitions of different requirements, and their importance for employers, as well as the application of natural language processing in analyzing employee requirements. Section 3 presents the data used in the study, the methods, and the metrics for evaluating the results, including data processing methods, data dimensionality reduction methods, data clustering methods, and metrics for evaluating their results. Section 4 discusses the research findings, including the dynamics of specific skills, data dimensionality reduction outcomes, a comparative analysis of the different clustering

methods, and the main profiles of the employees. Finally, Section 5 concludes the paper, summarizing the main findings and outlining the potential avenues for future research.

2. Job Requirements and Natural Language Processing Application

Significant changes in the labor market due to economic, geopolitical, and digitization effects, as well as the COVID-19 pandemic, have lead to a greater need for work skills and work requirements [12]. Digitization has already changed several areas of the labor market, and it has been noticed that there is an increased demand for employees who have the necessary competencies to work with the latest technologies and specific programming languages. As digitization is rapidly increasing, there is already a perceived need for specialists with technical skills such as cyber security and data analysis. However, soft skills are no less important: critical thinking, communication, and emotional intelligence are all still valuable [13]. COVID-19 has led to a massive shift in the job market and accelerated remote work opportunities, with organizations allowing employees to work from home or abroad to adapt to this change. However, with the possibility of remote work, the need for strong digital literacy skills and effective virtual communication skills rose, and the need for more individual work grew [14]. As the experience of other countries in the world shows, skills and other job requirements are an integral component of a country's success [15]. Therefore, it is important to recognize these changes and conduct research accordingly so that educators, policymakers, and others can appropriately prepare the workforce. Research has shown that countries with a soft skills gap in the labor market are more efficient and innovative, have a better quality of life, attract more foreign investment, and build greater confidence [16]. It has also been noticeable that in countries where the analysis of workers' skills has not been conducted, there is a risk of a mismatch between the available workers' skills and the employers' needs [17]. There are several risks to the country's economy at this point. First, without clear information and understanding of what the exact skills the country's labor market needs are, this could lead to the investment of money and time in education and training programs that will not bring the desired result, as they do not meet the market's needs. Secondly, without clear information about what skills are needed today, employers cannot find suitable employees, which thereby reduces the company's productivity and economic growth [18]. In the last period, it was observed that employers often fill vacant job positions with foreign talent, but foreign talent is hard to come by, and is usually more expensive.

Employers require skills, abilities, and other requirements when looking for specialists for various positions. The labor market is changing rapidly. For example, a few years ago, soft skills such as communication, teamwork, and creativity were emphasized more [19]. However, soft skills and abilities are easier to learn and do not require specific training. In most cases, people have had such skills since childhood, so employers refer to soft skills as abilities in job advertisements [20]. In the labor market, certain skills and requirements are categorized as technical. These skills are acquired through specific learning and can be continually trained, improved, and expanded [21]. One challenge with technical skills is that employees are often reluctant to learn new abilities, leading to a shortage of specialists with up-to-date expertise. In the recruitment process, the requirements and skills mentioned in these job advertisements play a crucial role in determining whether a candidate possesses the necessary qualifications and qualities to be selected as the most suitable applicant. The specifications outlined in job advertisements hold significant importance throughout the selection and recruitment process, impacting both the job seekers and the employers.

Job seekers typically search for employment opportunities on various job posting portals, which often provide guidelines for employers regarding the information they must include about the position and the desired qualifications of potential candidates. However, these guidelines are not strictly enforced, leading to employers presenting job advertisements creatively, sometimes without listing the essential skills. Additionally, employers may not differentiate between the required skills, presenting them as part of the position's responsibilities. These factors all contribute to the job seeker's difficulty in

discerning the skills needed for a particular position. Conversely, employers might also struggle to prepare an effective job advertisement that accurately captures the requirements for a new position. In job advertisements, employers often specify certain requirements for potential candidates. Many employers mandate a minimum level of education, such as a bachelor's or master's degree in a relevant field of study [22].

There has been a growing demand for candidates with PhD degrees, particularly in sectors such as Fintech, artificial intelligence, and related fields. A candidate's educational background allows the employer to better understand the applicant's profile, and possessing a degree implies that the candidate has acquired the foundational knowledge, and is therefore likely capable of achieving good results in the relevant field [10]. Many employers seek candidates with specific technical skills related to the job, such as laboratory techniques, software expertise, or proficiency in various programming languages and tools. Evaluating technical skills enables the employer to determine whether the candidate can perform the assigned tasks and gauge their ability to learn new programming languages or tools quickly. Experience is often another requirement or advantage for gaining employment in a particular field, such as project management or data analysis projects. A candidate's work experience allows the employer to assess their existing skills and specialized knowledge. Typically, candidates with more work experience better understand the work environment [23]. Job postings often indicate the minimum years of experience required in a specific field. Although technical skills and experience are crucial, soft skills such as communication, collaboration, problem-solving, leadership, cultural fit, and a strong work ethic are also highly valued and sought after across all job fields. Employers actively search for candidates possessing suitable qualities and a cultural and value alignment with the organization, as these factors contribute significantly to the overall success of the company [24]. The above skills enable employers to evaluate a candidate's potential performance in their assigned tasks. While certificates and licenses may not be common requirements in all job advertisements, they are crucial for specific fields. For instance, medical professionals must possess a medical license and the necessary certificates to demonstrate their competencies and eligibility to work in a particular position. Certificates and licenses indicate a candidate's skills, experience, work quality, productivity, and suitability for a specific field [10].

The requirement for publications and research experience is not popular in job advertisements. However, this requirement is mandatory for candidates applying for academic and research-related positions. In order to occupy academic positions, a certain number of scientific research articles are often required. Employers often require knowledge of one or more languages besides their native language, especially if the work involves international clients, or the work environment is multilingual. Knowing multiple languages is a necessary skill in this era of globalization, as it facilitates communication between colleagues, partners, and clients. Knowledge of languages makes it possible to assess employees' ability to find information in another language, making work much easier [25]. One essential requirement for employees looking for work in the industry, construction, or production field is physical capacity, meaning in this case, the physical endurance needed to perform certain tasks, such as lifting weights, standing for prolonged periods, operating various machinery, and performing tasks properly and safely in compliance with all requirements. Some positions also require a background check. These are usually areas where employees can access confidential information or finances, such as public administration or finance. To get a job, a person must pass various background checks, which is important to maintain a safe working environment. Many jobs require the employee to be able to work nights, weekends, shifts, and holidays. Employers value the flexibility of candidates, which is important for the productivity and efficiency of the organization [26], and is especially valuable for areas such as industry, where work takes place around the clock. Moreover, legal eligibility is another important requirement, especially now that there are many workers from other countries in our country. This requirement indicates that the person must have a valid work visa and other necessary documents, according to which the individual has

the official permission to work in the country. It is an important requirement that ensures that employers do not face legal problems after hiring a candidate.

Job requirement analysis is becoming increasingly relevant in the current labor market. Understanding the required qualifications, skills, and attributes is crucial for job seekers, employers, and policymakers [8,27]. These requirements, which include factors such as education, experience, skills, certifications, language skills, physical requirements, background checks, availability, personal qualities, and legal eligibility, all play a critical role in determining the suitability of candidates for specific positions and ensuring the success of organizations [10]. An accurate assessment of job requirements is essential for job seekers to focus on acquiring their relevant qualifications and skills, ultimately enhancing their employment and career prospects [28]. For employers, clearly defined job requirements facilitate efficient recruitment procedures and select candidates with the right skills, thereby reducing turnover and improving overall workforce productivity. In addition, a deeper understanding of the job requirements can help policymakers design targeted education and training programs that address these skills gaps and promote workforce development, contributing to a more efficient and competitive labor market. Advances in machine learning, clustering, and NLP have improved our ability to analyze job postings and identify trends in the job requirements. Machine learning algorithms can be trained on large datasets of job postings to identify job requirements and their patterns. Clustering algorithms can group similar job postings according to their requirements, allowing for a more detailed analysis of the skills and qualifications required in a specific field. NLP techniques can extract information from unstructured text in job postings, such as required skills, education, and years of experience. By combining these methods, students can better understand the labor market and identify new trends and opportunities. Natural language processing (NLP) is a branch of artificial intelligence that develops algorithms and techniques to enable computers to understand, interpret, and generate human speech [29]. NLP has revolutionized text analysis by offering advantages over previous methods, such as increased efficiency, scalability, and the ability to uncover hidden patterns in enormous amounts of unstructured text data. Compared to manual text analyzes, NLP techniques can quickly process enormous quantities of data, making them ideal for sentiment analysis, topic modeling, machine translation, and information extraction applications. In addition, NLP techniques can capture complex linguistic structures and semantic relationships, allowing for a more accurate and detailed analysis of textual data. Despite its advantages, NLP also faces several challenges, including the inherent ambiguity and variability of natural languages, making it difficult for algorithms to accurately interpret the meaning and context of words and phrases [30]. Additionally, training NLP systems often require copious amounts of labeled data, which can be time consuming and expensive. Additionally, NLP models may struggle to accommodate the dynamic nature of language, which evolves and varies across domains, cultures, and communities [31].

3. Materials and Methods

This article subsection provides information about the data used in the study, their acquisition and processing, and the main characteristics of this data. This subsection also provides information about the main methods used in the research to solve the tasks of different stages of the research. Basic data processing methods, data clustering and evaluation metrics, natural language processing methods used in this study, and other possible methods were discussed. The main scheme of the study is shown in the figure below (see Figure 1), which helps to understand the main idea of this study. The following is a detailed description of the stages of the study.

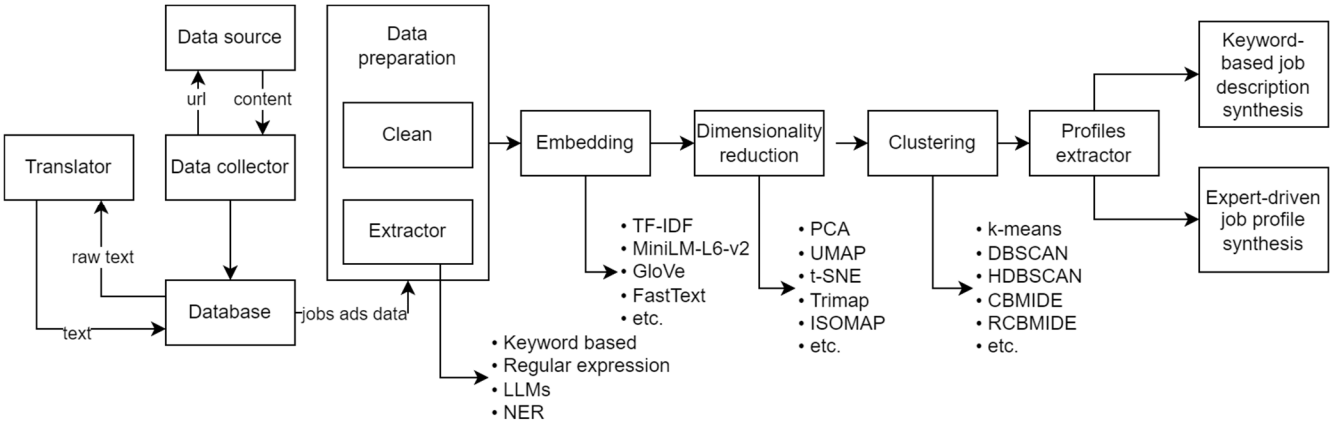

Figure 1. The main research scheme.

3.1. Data Gathering, Processing, Extraction, and Analysis

In order to collect the data required for this study, freely available Lithuanian job advertisements on the largest Lithuanian portals were used. Different Python libraries, including Playwright, BeautifulSoup, and Selenium, were used to collect the data. It is important to understand that the data collected was unstructured, requiring more processing than structured data. Moreover, due to the highly unstructured nature of the data, various data/advertisement structures were also possible, which only further complicates the work of data processing. The figure below shows an example of one of the possible data structures that was collected. The title section describes the name of the job position, and in rare cases, the exact job position was not written. However, a few words typically describe a certain activity, e.g., "Driver required", where it is not specified what driver is required. The location and company section contains information about the company and its location/city. The statistics section provides information on how many people have viewed the advertisement under analysis since it was posted. The nature of the work section describes the activities the employee will perform while working in this position in the company that submitted the advertisement. The requirements section contains the main requirements for an employee seeking to work in the analyzed position. The requirements are the main part of the analysis of this study, as the aim is to determine which skills are the most in demand, to create profiles of the advertisements presented, and to determine opportunities for employees to retrain for other positions. The offer section provides information about the employer's offer to the employee. Depending on the employer, additional benefits may be provided in this section, and a team description may also be provided. The last section would contain information regarding the salary and benefits that the employee will receive during additional work at this workplace. All collected data were stored in the created and protected PostgreSQL database, allowing for data collection and analysis in parallel.

Upon initial observation, the data seems to have a clear structure; however, in this example (see Figure 2), only a perfectly written job advertisement was displayed. In reality, the number of such well-structured job advertisements is quite limited. All components of the job advertisement are free form, meaning they may not always be present, and employers might label these components differently, such as "Requirements", "Required Skills", or "Competencies". Furthermore, these components can be placed in various sections of the job advertisement, potentially describing the required skills while only providing information about the job tasks. Consequently, significant uncertainties in data acquisition makes data extraction particularly challenging. Several different methods for extracting data from such texts are explored below.

Keyword-based search: this method searches for specific words or phrases in the analyzed text. This method can easily be used in common search tools. In this case, the big problem is that an initial set of keywords are required. Only specific words are searched

for, meaning when new requirements appear, adding the dictionary of the searched words is also required, which is unacceptable for a real-working system. One solution, in this case, is a synonyms search using WordNet [32], WordHoard (WordHoard Github repository: https://github.com/johnbumgarner/wordhoard, accessed on 25 March 20233), and others. A similar method is the rule-based method. In this case, the aim is to create rules based on certain linguistic patterns or parts of speech (POS) that reflect certain job posting requirements. For example, it is possible to find noun phrases that follow specific verbs (e.g., "require", "seek", and "need") or adjectives (e.g., "strong", "excellent", and "demonstrated").

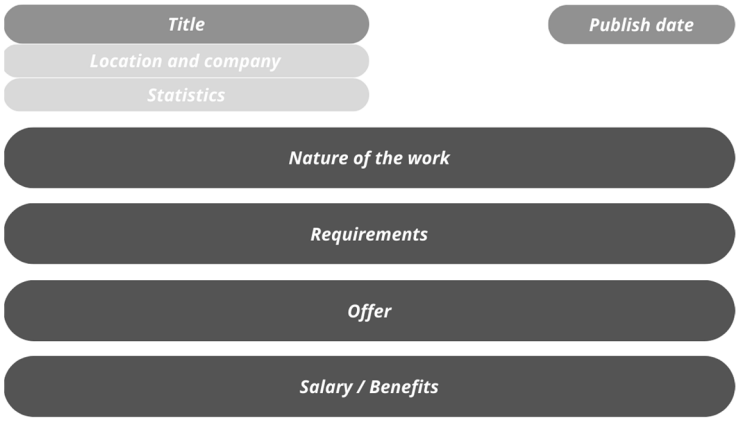

Figure 2. The ideal structure of the unstructured job advertisement text.

Another way to extract the required data from the text is regex. Regular expressions, abbreviated as regex, are powerful for pattern matching and text processing. A regex is a string of characters defining a specific search pattern. Regular expressions provide a concise and flexible way to extract information from unstructured data, such as log files, emails, and web pages. They can be implemented in various applications, including data validation, text extraction, and cleaning. While regular expressions can be complex and challenging to master, they are an essential tool for data scientists and software developers who need to work with copious amounts of text data. In this case, using regex makes it possible to find the requirements in job advertisements using patterns such as "Requirements" and the ending pattern "Company offers". For example:

> *"This is the job description. Requirements: experience, English etc. Company offers: health insurance, flexible work".*

After using regex and the initial pattern "Requirements" and the final pattern "Company offers", the text "experience, English etc." is then extracted.

Another particularly developing way to extract the necessary information, systematize substantial amounts of textual information, or even generate additional textual data is large language models. LLMs are advanced machine learning models that are capable of processing and generating human-like language. These models are trained on massive amounts of textual data, such as books, articles, and websites, using complex algorithms that enable them to identify and learn patterns in language. Some examples of LLMs include OpenAI's GPT-4 (OpenAI GPT-4 information: https://openai.com/product/gpt-4, accessed on 25 March 2023), GPT-3.5-turbo (OpenAI GPT-3.5-turbo information: https://platform.openai.com/docs/models/gpt-3-5, accessed on 25 March 2023), GPTNeoX [33], and Google's BERT [31]. LLMs have many applications, including in natural language processing, machine translation, and chatbot development. They are becoming increasingly popular in artificial intelligence and have the potential to revolutionize how humans interact with computers and machines. Using the above example and extracting the necessary data using this methodology, the request looks like this:

> *"Extract job requirements from the text provided:*

This is the job description requirements: experience, English etc. The company offers health insurance and flexible work".

The results obtained after using OpenAI's GPT-3.5-turbo are:

"Based on the text provided, the job requirements are experience (specifics not mentioned) English proficiency (specifics not mentioned). It is important to note that without further context, it is difficult to determine the level of experience or proficiency required for the job".

At the time of research, the price of GPT-3.5-turbo was $0.002 per 1000 tokens, where 1000 tokens are about 750 words. The results obtained after using Open-AI's GPT-4 look like this: "*From the provided text, the job requirements are 1. Experience 2. English proficiency. Additionally, the company offers: 1. Health insurance, 2. Flexible work*". At the time of research, GPT-4 was $0.03 per 1000 prompt tokens and $0.06 per 1000 completion tokens, making it even about 15 times more expensive.

Given that the collected data contained about 120 million words and the answers' size would also be similar, the required number of tokens would be 320 million. For instant research, it is a more expensive method than the one mentioned above. A considerable amount of prompt engineering was also required for the model to answer what was being asked and to provide the results in the desired format.

Finally, the last method discussed to identify the required information from the text was named entity recognition (NER). NER is a fundamental task NLP that involves identifying and classifying named entities, such as persons, organizations, locations, and other specific terms in the unstructured text [34]. In the context of job advertisement requirement analysis, NER can play a crucial role in extracting the relevant information regarding the desired qualifications, skills, and other attributes employers seek in potential candidates. By automatically recognizing and categorizing the key entities mentioned in the job advertisements, NER can help streamline the process of analyzing large numbers of job postings, thereby enabling researchers, job seekers, and employers to gain insights into the dynamics of the job market and the most in-demand skills and qualifications. Furthermore, NER techniques can be combined with other NLP and machine learning methods, such as topic modeling and clustering, to group job advertisements based on the extracted requirements, as well as identify the common patterns, trends, and skill gaps across various industries and job roles.

In summary, applying named entity recognition to job advertisement requirement analysis can provide valuable information for job seekers, employers, and educational institutions, thereby helping them to make informed decisions and adapt to the ever-changing job market. However, in order to use this methodology in this research, a specially trained model was needed to determine the job advertisement's requirements. To train the model, a large amount of data was required, which was available in this work, but these data were not labeled. For this reason, the application of NER in this study was postponed in further project plans, which are discussed in the subsection future directions.

This work relies on keywords-based (for the visualization of the specific skills) and regex methodology, which was already presented and mentioned earlier, and during the research, these data were prepared for the training of the NER model, which can be applied in further research.

3.2. Text Vectorization

After cleaning the data from unnecessary information, the textual data must be vectorized for analysis, clustering, and other research tasks. Text vectorization converts the text data into numerical data for later utilization in machine-learning techniques [35]. Text vectorization is a particularly crucial step in natural language processing in classification, clustering, information extraction, and sentiment analysis. It allows machine learning algorithms to operate on the text data in a numerical format. The main idea of text vectorization is that each text document (a sentence, paragraph, or entire document) is recorded with

numerical data while preserving the main information of these documents. An essential characteristic of text vectorization is preserving information as much as possible after this process. Various methods were used for text vectorization, such as:

Term frequency-inverse document frequency (TF-IDF) is a popular text vectorization method. TF-IDF computes the importance of each term in a document relative to a collection of documents [36]. It is based on the idea that terms frequently appear in a specific document. However, it is rare to carry more discriminative information across the entire document collection. The resulting term-weighted document vectors can be used for tasks such as document classification and information retrieval. Although TF-IDF might not capture semantic meaning as effectively as sentence transformers, it is computationally efficient. It has been proven useful for various text analytical applications:

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right) \quad (1)$$

where $tf_{i,j}$—frequency of the i in the j, df_i—number of the documents containing i, and N—total number of the documents.

Another important method that has recently received a lot of attention is sentence transformers. Sentence transformers, introduced by Reimers and Gurevych (2019) [37], employ pre-trained transformer models such as BERT [31] to generate dense vector embeddings of sentences. These embeddings capture the semantic meaning of the input text. They can be used for various tasks, including text classification, semantic similarity, and clustering. The advantage of sentence transformers is that they can capture complex linguistic structures and relationships within the text, thereby improving the performance on various NLP tasks. BERT is a pre-trained transformer model that utilizes self-attention mechanisms to capture contextual information in both directions (left-to-right and right-to-left) [31]. Fine-tuning BERT on specific tasks can generate high-quality sentence or document embeddings, which are effective for various NLP tasks such as sentiment analysis, text classification, and semantic similarity. Advantages of these pre-trained transformer models are as follows: captures context and word order at the sentence or document level, can be fine-tuned on specific tasks, yielding high-quality embeddings, and that it can outperform traditional methods such as TF-IDF and Word2Vec in numerous NLP tasks. However, disadvantages of these pre-trained models include that it requires substantial computational resources for training and fine-tuning, generates high-dimensional embeddings, which may necessitate dimensionality reduction techniques, and that these models are characterized with a complex architecture and larger model size compared to other methods. Another method that can be compared with the sentence transformers is OpenAI Ada embedding.

However, although the previously mentioned methods are among the main methods for vectorizing textual data, other methods can be used to implement these tasks. Among the alternative methods, we can mention other, more frequently used methods:

- Word2Vec, proposed by Mikolov et al. (2013) [38], is a popular word embedding technique that generates dense vector representations of words by predicting the context of a given the word using an external neural network. These embeddings capture the semantic and syntactic relationships between words. The word vectors can be averaged, summed, or combined for a sentence or document-level representation using more sophisticated techniques including weighted averages or the smooth inverse frequency method [39]. Word2Vec consists of two primary architectures: Continuous bag of words (CBOW) and Skip-Gram. CBOW predicts a target word based on its surrounding context, while Skip-Gram predicts the context given a target word. There are several limitations to the Word2Vec method: It requires substantial computational resources for training on large corpora. It focuses on word-level embeddings, which may not capture sentence- or document-level semantics. Despite these limitations, Word2Vec does possess several benefits: It captures semantic and syntactic relationships between words. Generates dense continuous vectors, reducing dimensionality

compared to sparse methods such as TF-IDF. Pre-trained models are available for various languages and domains.

- GloVe (Global Vectors for Word Representation) is another word embedding method introduced by Pennington et al. (2014) [40]. It generates word embeddings based on the global co-occurrence statistics of words in a corpus. Similar to Word2Vec, GloVe embeddings can be aggregated to create a sentence or document-level representations similarly.
- The bag-of-words (BoW) model is a simple text vectorization method that represents documents as fixed-size vectors based on the frequency of words they contain [41]. While BoW does not capture the order of words or semantic relationships, it is computationally efficient. It can be effective for certain text analytical tasks.
- Developed by Bojanowski et al. (2017) [42], FastText is an extension of the Word2Vec model that generates embeddings for sub-word units (such as N-grams) instead of entire words. This approach enables capturing of morphological information and better handling of the rare and out-of-vocabulary words. The sentence or document-level embeddings can be obtained by aggregating the sub-word embeddings.
- Doc2Vec, also known as paragraph vectors, is an extension of Word2Vec introduced by Le and Mikolov (2014) [43]. It generates dense vector representations for entire documents by considering both the words and the document as an input during the training process. This method can capture the overall semantic meaning of a document and can be directly used for document-level tasks.

In summary, various text vectorization methods can serve as alternatives to sentence transformers and TF-IDF, each with their strengths and limitations. The choice of the appropriate method depends on factors such as the specific NLP task, data characteristics, and computational resources. In this work, we used sentence transformers models for the initial data vectorization and feature extraction.

3.3. Dimensionality Reduction Methods

Vectorization of textual data often results in extremely large matrices. In the case of sentence transformers, the resulting matrix is Nx364. In contrast, the TF-IDF matrix depends on how many different words are in the data and which N-gram is used. Data dimensionality reduction plays a particularly significant role in data analysis and various machine learning techniques by reducing high-dimensional datasets to low-dimensional datasets that contain only the most valuable information and relationships. Large datasets can be challenging to work with as they often contain noise, redundancy, and sparsity, making it difficult to identify patterns and relationships in the data. When performing data clustering, a particularly enormous number of dimensions is not desirable if these dimensions do not provide additional information. Data dimensionality reduction is an important preprocessing step prior to data clustering, as it helps overcome the challenges many dimensions face. This phenomenon is often known as the "curse of dimensionality" [44]. As sparsity in high-dimensional spaces increases, many dimensions can lead to the overfitting and poor clustering of the results. With the utilization of data dimensionality reduction, it has been noticeable that better data clustering results are obtained, and such results are much easier to interpret [45]. Data dimensionality reduction techniques have also been well used to reduce noise in data [46], or detect outliers in the data [47], and improve the accuracy and generalization of the models being developed [48]. Data dimensionalities reduction methods such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) transform the original high-dimensional data into smaller-dimensional data, while also retaining as much as possible, more similar data structures and relationships between the individual data points [49]. These data dimensionality reduction techniques allow clustering methods such as k-means and DBSCAN to identify clusters in the data more efficiently. Moreover, data dimensionality reduction methods significantly reduce computational complexity and memory requirements for clustering methods, thus enabling

them to perform actions with larger datasets. However, it is important to note that data dimensionality reduction methods also have their disadvantages. One is that a certain amount of information can be lost when the dimensionality is reduced [50].

The main data dimensionality reduction techniques used in this study are briefly discussed below. One of the most widely used and widely known data dimensionality reduction techniques is principal component analysis (PCA). The principal component analysis is a linear method that projects data into a space of smaller dimensions to maximize variance and, at the same time, minimize information loss. PCA identifies and extracts the principal components, which are orthogonal linear combinations of the original variables, capturing the maximum variance in the data while minimizing information loss [51]. The principal components are ranked according to their variance. Hence, the first principal component captures the largest variance, the second principal component the second largest variance, and so on. By retaining only a few principal components with a higher variance, PCA can effectively reduce the dimensionality of the data while retaining most of the information. PCA has been successfully applied in various domains, including image processing, bioinformatics, and finance, for feature extraction, data compression, visualization, and noise reduction [52,53]. Recent advances in PCA include incorporating regularization techniques, such as sparse PCA and elastic net PCA, which add constraints to improve the interpretability and robustness of the model, particularly in high-dimensional data with multicollinearity issues [54]. Despite its advantages, PCA has certain limitations, such that it assumes linearity in the data, which may not always be appropriate, especially when the data exhibits complex nonlinear structures.

Another data dimensionality reduction method that, in contrast to principal component analysis, is not linear, is t-distributed stochastic neighbor embedding (t-SNE). This method has been widely used due to its ability to maintain local and global relationships in high-dimensionality reductions, thus maintaining exceptionally good data visualization and interpretation [55]. This method has gained much attention recently due to its ability to effectively visualize large-dimensional data by representing it in smaller dimensions [56]. The t-SNE algorithm achieves this by converting high-dimensional Euclidean distances into conditional probabilities, representing similarities between data points. Then, a gradient descent optimization method minimizes the divergence between these probabilities across dimensions [57]. Researchers have proposed efficient approximations, such as the Barnes-Hut t-SNE, which allows the algorithm to manage large datasets while maintaining its effectiveness in preserving local structures. T-SNE has been widely applied across various scientific domains, including single-cell analysis, bioinformatics, and computer vision. Despite its success, certain limitations of t-SNE have been identified, such as sensitivity to initial conditions and hyperparameters, slow convergence, and the presence of local optima [56].

Uniform manifold approximation and projection (UMAP) is a newer data dimensionality reduction method. Similar to t-SNE, it is nonlinear, and can maintain local and global relationships. Data dimensionality reduction improves data visualization and interpretation [49]. UMAP is built on the foundations of multivariate learning and topological data analysis, using concepts such as simple fuzzy sets to model the underlying geometry of the data [58]. UMAP creates fuzzy topological representations of high-dimensional data and projects them into a lower-dimensional space. The resulting projection preserves the local structure of the data, meaning that nearby points in high-dimensional space are also close to each other in the low-dimensional projection. UMAP is similar to other dimensionality reduction methods, such as t-SNE and PCA, but has several advantages over these methods. For example, UMAP is generally faster and more scalable than t-SNE and can handle larger datasets. In addition, UMAP can be used with various distance metrics, including non-Euclidean metrics that allow for capturing complex data relationships. UMAP's preservation of local and global structures facilitates data visualization and interpretation. It is suitable for various applications, including single-cell RNA-seq data analysis [58], image classification, and natural language processing tasks [59]. In addition,

UMAP has shown faster execution times than t-SNE, making it more suitable for large-scale datasets [49]. Recent advances in UMAP include the development of supervised and semi-supervised variants that incorporate label information into the dimensionality reduction process, resulting in improved class separation and more meaningful embeddings. Despite its advantages, UMAP can exhibit some limitations, especially in cases where assumptions about the underlying data collector do not hold, or when the data show high noise levels. In such situations, alternative dimensionality reduction methods may be better.

In this work, we assessed different dimensionality reduction methods listed in this section and different method performances presented in the results sections.

3.4. Clustering Methods Used in the Research

The following section provides an in-depth overview of the clustering algorithms that were employed to analyze the dataset under investigation. Clustering is a fundamental technique in unsupervised machine learning. It offers valuable insights into data's inherent structure and relationships by grouping similar objects into clusters based on their features. Selecting appropriate clustering algorithms is crucial for obtaining accurate and meaningful results. Each method has its strengths and weaknesses depending on the data's specific characteristics and the analysis's objectives.

In this research, we have carefully chosen a diverse set of clustering algorithms, including hierarchical, partitioning, density-based, and model-based methods, to provide a comprehensive understanding of the underlying patterns in the data. Each clustering method offers a unique perspective on the data organization. By comparing their results, we aimed to derive robust conclusions and minimize potential biases associated with any algorithm [60]. This section will outline the principles and rationale behind each clustering method used in the study, discuss their respective strengths and limitations, and provide a detailed explanation of their implementation within the context of this research. Furthermore, we will describe the process of selecting the optimal number of clusters and present the unsupervised evaluation metrics employed to assess the quality of the clustering results obtained from each algorithm.

The main data clustering techniques used in this study are discussed below. K-means clustering is a widely used unsupervised learning technique that aims to partition unlabeled data into distinct clusters based on their inherent features [61]. K-means clustering operates by iteratively assigning data points to a predefined number of cluster centroids based on the minimization of within-cluster distances, typically using Euclidean distance as the similarity measure. The algorithm initializes by selecting random or strategically placed centroids. It then iteratively refines their positions until convergence, resulting in compact and well-separated clusters. One of the main advantages of K-means is its simplicity and ease of implementation, which makes it computationally efficient and suitable for large datasets. However, the algorithm has limitations, such as sensitivity to the initial centroid positions and the requirement to specify the number of clusters a priori, which may not always be known or easily determined [62].

Additionally, K-means is predominantly suited for detecting spherical and equally sized clusters and may struggle with more complex data. Recent advancements in this field have expanded its applicability to various domains, including text analytics. For instance, K-means has been employed in document clustering, grouping similar textual information, and enhancing information retrieval efficiency.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that identifies clusters by grouping data points that are tightly packed based on a specified distance metric and density threshold [63] (Ester et al., 1996). This approach allows DBSCAN to effectively manage clusters of arbitrary shapes and noise in the data, which is a significant advantage over traditional partitioning methods such as K-means. However, DBSCAN is sensitive to its hyperparameters, namely the neighborhood radius (Eps), and the minimum number of points required to form a dense region (MinPts), which can be challenging to determine a priori [63]. HDBSCAN (hierarchical density-based

spatial clustering of applications with noise) is an extension of DBSCAN that addresses some of its limitations by incorporating a hierarchical clustering approach [64]. HDBSCAN does not require the specification of a global density threshold; instead, it automatically identifies clusters at varying densities by constructing a dendrogram and applying a cluster extraction method based on the stability of clusters over different density levels. This results in a more robust and flexible clustering algorithm adapting to varying data distributions and densities.

Nevertheless, both DBSCAN and HDBSCAN may suffer from high computational complexity, especially for large datasets, and are sensitive to the choice of a distance metric, which can significantly impact clustering performance. DBSCAN and HDBSCAN have been applied to various text analytical tasks, leveraging their ability to detect clusters of arbitrary shapes and varying densities. For instance, DBSCAN has been employed in topic modeling. It can identify coherent thematic groups within large collections of documents, improving the organization and retrieval of textual information. Similarly, HDBSCAN has been proven valuable in text summarization, enabling the extraction of representative sentences from a given document while maintaining the diversity of information and covering various aspects of the content. These applications demonstrate the versatility of density-based clustering methods in text analysis, offering unique advantages in handling complex data distributions and capturing nuanced relationships within textual data.

BIRCH (balanced iterative reducing and clustering using hierarchies) is a hierarchical clustering method designed to efficiently process large datasets by constructing a tree structure called the clustering feature tree (CF-Tree) that captures the essential attributes of data points, such as their linear sum and squared sum [65]. The algorithm can manage large datasets by incrementally processing data points and adjusting the tree structure dynamically, significantly reducing computational complexity and memory requirements compared to the traditional hierarchical clustering methods. Recent applications of BIRCH in text analysis include document clustering. The method can group similar texts based on their feature representations, such as term frequency-inverse document frequency (TF-IDF) vectors. BIRCH has also been employed in analyzing customer reviews, enabling the identification of patterns and trends in customers' opinions, which can inform businesses about their strengths and areas for improvement.

Furthermore, BIRCH has been utilized in social media analytics for event detection. The algorithm can cluster textual data from social media platforms to identify significant events or emerging discussion topics. These applications demonstrate the potential of BIRCH as an efficient and scalable clustering method for text analysis tasks, particularly in handling large-scale textual data.

Affinity propagation is a clustering algorithm based on message passing that identifies a set of exemplar data points, which best represent the clusters, and groups similar data points around them [66]. Unlike many other clustering methods, affinity propagation does not require the number of clusters to be specified a priori, making it more adaptive to various data distributions. The algorithm works by iteratively exchanging real-valued messages between data points until their convergence, reflecting the suitability of each point to serve as an exemplar, and the preference of each point to select a specific exemplar. In the context of text analysis, affinity propagation has been applied to various tasks, such as document clustering, where it can effectively group similar documents based on their content or feature representations (e.g., TF-IDF vectors) [67]. Additionally, affinity propagation has been employed in analyzing social media data, such as community detection in social networks based on user-generated content, enabling the discovery of meaningful user relationships [68]. Despite its adaptability, affinity propagation may suffer from high computational complexity, particularly for large datasets, and sensitivity to the choice of preference parameter, which influences the number of exemplars. Nevertheless, the algorithm's ability to automatically determine the number of clusters and its robustness to noise make it a valuable method for various text analysis applications.

Spectral clustering is a technique that leverages the spectral properties of the data's similarity matrix to perform clustering in a lower-dimensional space, thus enabling the detection of complex structures and non-linear relationships within the data [69]. The algorithm first constructs a similarity graph, where nodes represent data points, and edges represent pairwise similarities, typically using Gaussian kernel or k-nearest neighbors. It then computes the eigenvalue decomposition of the graph's Laplacian matrix, clustering the eigenvectors corresponding to the k smallest eigenvalues using traditional clustering methods, such as K-means. Spectral clustering has been applied in text analysis tasks to various scenarios, including document clustering. It can efficiently group similar documents based on their feature representations, such as term frequency-inverse document frequency (TF-IDF) vectors [70]. However, Spectral clustering has some limitations, such as sensitivity to the choice of similarity measure and the requirement to specify the number of clusters. The method can also be computationally expensive, especially for large datasets, due to the eigenvalue decomposition step.

CBMIDE (clustering based on the modified inversion formula density estimation) [71] and RCBMIDE (reduced clustering based on the modified inversion formula density estimation) [72] are novel clustering methods that leverage the modified inversion density estimation to effectively capture the underlying structure of the data. In contrast to the traditional methods such as the Gaussian mixtures models, CBMIDE focuses on the reciprocal distance between the data points and the cluster centers, thereby providing a more robust estimation of the density structure. By utilizing this alternative approach to density estimation, CBMIDE has the potential to reveal complex relationships and structures within the data that conventional methods may overlook. CBMIDE can be applied in text analysis to various tasks, such as document clustering. It can efficiently group similar documents based on their feature representations, such as the term frequency-inverse document frequency (TF-IDF) vectors. It is also worth noting that this clustering method exhibits robustness, ensuring that only legitimate data clusters are identified and analyzed, thereby minimizing the influence of noise and outliers on the overall results, and enhancing the accuracy and reliability of the findings.

In this work, we evaluated all the abovementioned clustering methods and compared these methods' performances based on the different metrics. All results are presented in the results sections.

4. Results

This subsection presents the main results obtained during the study. First, the data collected during the research and the dynamics of the data collected during the research were reviewed. The second subsection of this chapter presents the data preparation and feature extraction used in this research. The third subsection of this chapter presents the results of different methods of data dimensionality reduction and data clustering, allowing us to decide on further methods in this study. The fourth subsection of this chapter presents the data clustering results and the interpretation of these results, which are the main insights.

4.1. Dynamic of the Specific Requirements/Keywords in Lithuania

The rapid evolution of technology and industries has resulted in the job market's continuous shift of skill requirements. It is therefore essential to understand these changing dynamics to help job seekers, employers, and educational institutions adapt to the evolving landscape. In this section, we present an exploratory data analysis (EDA) of job advertisements over time, focusing on the prevalence of multiple keywords that represent specific job requirements. By extracting these keywords from job adverts and tracking their frequency over time, we aimed to uncover the trends and patterns that reflect the growing or declining importance of various skills and qualifications across different industries and job roles.

The analysis includes a series of graphs for each keyword, illustrating its frequency in job adverts over time. These visualizations enabled us to identify emerging trends, such as

the increasing demand for certain programming languages, data analysis tools, soft skills, and shifts in industry-specific requirements. Comparing these trends across multiple keywords also revealed the relative importance of various skills and qualifications in the job market, thereby providing valuable insights for job seekers, employers, and educational institutions.

By understanding the dynamic nature of job requirements, job seekers can make informed decisions regarding the skills they should acquire or further enhance. At the same time, employers can design more targeted job adverts and recruitment strategies. Moreover, educational institutions can adjust curricula to better prepare students for the evolving job market, ensuring that graduates have the most relevant and in-demand skills.

$$F_{k,t} = \frac{\sum_{i=1}^{n} D_{k,t,i}^{T}}{n} \quad (2)$$

where $F_{k,t}$—frequency of the specific keyword k in the time moment t, n—total number of the documents, and $D_{k,t,i}^{T}$—documents that contain specific keyword k at the moment t.

The graph below (see Figure 3) provides information about the demand for selected programming languages in job advertisements by analyzing the requirements of job advertisements. Based on the presented graph, the SQL programming language is in the greatest demand. However, at the same time, it was also noticeable that in the last period, the demand for this programming language, similar to the other programming languages, has decreased quite strongly.

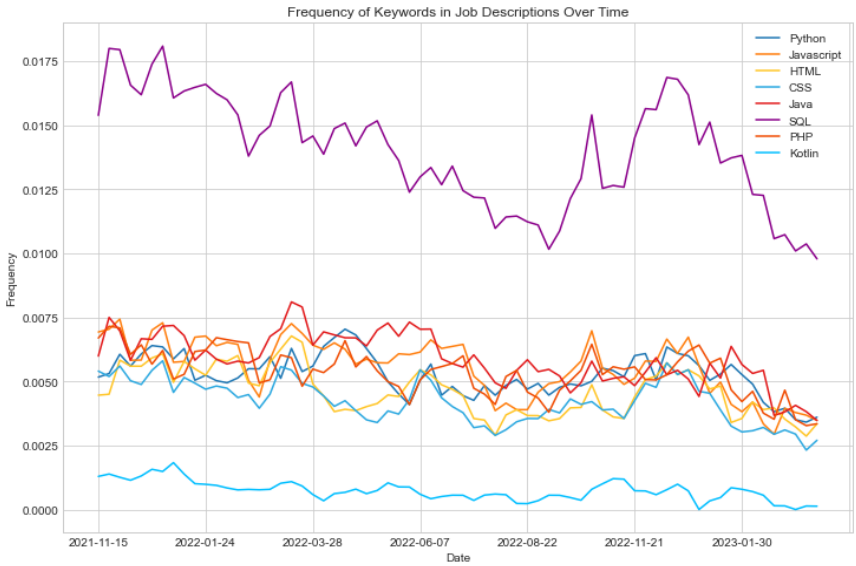

Figure 3. Different selected programming language demand changes over time.

4.2. Feature Extraction of the Job Adverts

After collecting the information, the relevant information was extracted from these job adverts—the job requirements, the publication date necessary to determine dynamic changes, and the name. After comparing the methods discussed earlier, the best results and the simplest implementation were observed using the regex procedure and distinguishing certain job requirements section names. It is well known that using method-generated synonyms makes this implementation simpler, as it reduces possible different combinations. After extracting the information significant for the study from the data, feature extraction was performed. The previously mentioned sentence transformers (BERT structure) method, whose dimensions are 384, was found to have performed this function best.

4.3. Comparative Analysis of Dimensionality Reduction Results

After performing feature selection, the size of the data dimensions obtained, as mentioned earlier, was 384. With such large data dimensions, data clustering required much

larger resources. Several methods, such as CBMIDE, are not adapted to such dimensions. The following work uses data dimensionality reduction methods based on these insights. Different data dimensionality reduction methods have different properties and perform differently depending on the dataset. In order to properly evaluate data dimensionality reduction, metrics were needed to enable this. Estimating the dimensions of the data allow for determining which new dataset was the best, and how many new dimensions were needed to retain the maximum amount of information possible. One of the possible evaluation metrics is the trustworthiness metric [73]. This metric assesses how well the information and relationships between observations in the original and reduced dimensions are preserved. The computation of this metric was accomplished by utilizing the following formula:

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in N_i^k} \max(0, (r(i,j) - k)) \quad (3)$$

where for each observation i, N_i^k are its k nearest neighbors (the output space obtained after applying the methods), and for each observation j, $r(i,j)$ was its original data space. In the event of a random observation appearing in the output space, it incurred a penalty. At the same time, the study employed a set of five neighbors.

During data dimensionality reduction in this work, different data dimensionality reduction methods were used, and their hyperparameters were changed. For all data dimensionality reduction methods, the number of new dimensions varied from two to fifty. In the case of PCA, only the number of dimensions was varied. In the case of UMAP, the parameters and their values were changed: n neighbors hyperparameter set {5, 10, 20, 30, 50, 100}, minimum distance set {0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.75, 1}, and distance metrics {euclidean, cosine, and manhattan}. ISOMAP parameters and their values: n neighbors hyperparameter set {5, 10, 20, 30, 50, 100}, and the distance metric hyperparameter set {manhattan, cosine, manhattan, and minkowski}. T-SNE parameters and their values: learning rate {10, 50, 100, 500, and 1000}, and perplexi-ty {5, 10, 20, 30, 40, 50}. Trimap parameters and their values: number of inliers hyperparameter set {5, 10, 12, 20, 30, 50}, and the number of outliers hyperparameter set {2, 3, 4, 5, 10}. Below is a table of the best model results (see Table 1). More than 11,000 models were created to evaluate the best dimensionality reduction method for the current task.

Table 1. Trustworthiness metric values for the best dimensional reduction model of all the tested methods in specific dimensions.

Components	PCA	UMAP	Trimap	t-SNE	ISOMAP
2	0.756	0.933	0.831	0.871	0.748
3	0.805	0.950	0.882	0.881	0.821
4	0.844	0.955	0.908	0.883	0.859
5	0.876	0.961	0.931	0.886	0.896
6	0.898	0.964	0.942	0.902	0.908
7	0.915	0.966	0.950	0.912	0.920
8	0.928	0.969	0.955	0.928	0.938
9	0.940	0.971	0.961	0.937	0.949
10	0.948	0.973	0.964	0.954	0.965
15	0.968	0.974	0.971	0.966	0.977
20	0.980	0.975	0.974	0.975	0.989
25	0.986	0.978	0.976	0.978	0.989
30	0.990	0.980	0.977	0.979	0.990
35	0.992	0.983	0.978	0.981	0.991
40	0.993	0.984	0.978	0.982	0.991
50	0.996	**0.986**	0.978	0.984	0.992

Bold underlined value indicates the selected model used in the further research.

Based on the results of data reduction, it can be observed that all the methods performed quite similarly for a larger number of components. All studied methods became fairly stable from 15 dimensions. A higher number of dimensions was found to not improve the results of the models. Given that the results are important to consider the training time of the models, it was noticeable that the t-SNE models were trained longer compared to others as the Barnes-Hut algorithm was not used. The ISOMAP method also had quite a higher computational time compared with the other methods. It is important to note that the PCA method has a high trustworthiness value. However, after further analysis, it was noticed that the work profiles created using the PCA method were more complicated to interpret, and therefore this method was abandoned in further work. Based on these results, the UMAP dimensionality reduction method with fifty new output dimensions was further used in this work due to a faster computation time.

4.4. Clustering Results

This subsection provides information about the different clustering methods used during the study and the obtained results. A key point to evaluate clustering, in this case, is that data clustering was performed without prior knowledge of the actual data classes/clusters. It was impossible to use metrics such as accuracy, NMI, or other metrics requiring real clusters. Therefore, this paper used only the metrics that do not require real clusters and can be applied to solve real problems. Commonly used evaluation metrics for clustering without prior knowledge of labels include the silhouette coefficient, Davies–Bouldin index, and the Calinski–Harabasz index (CH Index). The silhouette coefficient measures the cohesion and separation of the clusters by comparing the average distance between data points within the same cluster to the average distance between data points in the nearest different cluster [74]. Higher silhouette coefficient values indicate better clustering quality, ranging from -1 to 1. The Davies–Bouldin index evaluates the clustering quality by combining intra-cluster similarity and inter-cluster dissimilarity [75]. Lower Davies–Bouldin index values signify better clustering with compact and well-separated clusters. The Calinski–Harabasz index (CH Index) assesses clustering quality by comparing the ratio of the between-cluster dispersion to the within-cluster dispersion [76]. A higher CH index value indicates better clustering, as it signifies a greater separation between clusters relative to the dispersion within clusters.

During data dimension reduction in this work, different data clustering methods were used, and their hyperparameters were changed, thus determining the most suitable data clustering method for the data. K-means variable parameter k is the number of clusters whose parameter set is {2, 3, 4, 5, 10, 20, 30}. DBSCAN changeable parameters: eps with set {0.1, 0.2, 0.3, 0.4, 0.5, 0.75}, and minimum samples {5, 10, 20, 30, 50, 100}. HDBSCAN changeable parameters: minimum cluster size set {10, 20, 30, 40, 50, 100}, minimum samples set {5, 10, 20, 50}, and epsilon set {0.1, 0.2, 0.3, 0.4, 0.5, 1}. BIRCH changeable parameters: number of clusters set {2, 3, 4, 5, 10, 20, 50}, and branching factor set {10, 20, 50, 100}. Affinity propagation damping factor set {0.5, 0.6, 0.7, 0.8, 0.9}. CBMIDE and RCBMIDE parameters: number of the cluster as earlier methods, number of projections directions T set {5, 10, 20, 50, 100}, smoothing parameter h {0.05, 0.1, 0.2, 0.3, 0.4, 0.5}, and probability of noise cluster set {0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5}. Below is a table of the best model results (see Table 2). More than 3000 models were created to evaluate the best available model for this task.

As mentioned earlier, more than 3000 data clustering models were created based on the reduced dataset. The reduced dataset made it possible to perform calculations faster. The results presented in the previous subsection prove that it is possible to do this with almost no loss of information. Data clustering assessment was performed based on Davies–Bouldin and Calinski–Harabasz metrics. In this case, the main metric was time Davies–Bouldin. In the case of extremely similar values, the Calinski–Harabasz metric was considered. Based on the obtained results, it was found that the HDBSCAN method displayed the best results, with a Davies–Bouldin value of 0.4475. HDBSCAN has emerged as the best-performing method in this scenario due to its unique ability to manage clusters of varying densities and

shapes, and its robustness in identifying noise points. Unlike other clustering algorithms that require a predefined number of clusters, HDBSCAN automatically detects the optimal cluster structure based on the data's underlying density distribution. This adaptability makes it particularly suitable for complex datasets with uneven densities and non-spherical clusters. Furthermore, it is important to note that good results were also obtained using the RCBMIDE clustering method. This robust method eliminates the outliers' impact. Based on these results, the HDBSCAN method was chosen to be used in further research.

Table 2. Davies–Boulding and Calinski–Harabasz metrics values for the top three models of each clustering model evaluated in the research.

Method	Parameters	Davies–Bouldin	Calinski–Harabasz
K-means	{'n_clusters': 5}	0.9143	**3386**
	{'n_clusters': 10}	1.0041	2995
	{'n_clusters': 20}	1.0487	2551
DBSCAN	{'eps': 0.2, 'min_samples': 30}	1.1245	1352
	{'eps': 0.3, 'min_samples': 20}	1.1568	1458
	{'eps': 0.3, 'min_samples': 50}	1.2658	1589
HDBSCAN	{'cluster_selection_epsilon': 0.3, 'min_cluster_size': 50, 'min_samples': 20}	**0.4475**	2698
	{'cluster_selection_epsilon': 0.2, 'min_cluster_size': 20, 'min_samples': 5}	0.7968	1398
	{'cluster_selection_epsilon': 0.3, 'min_cluster_size': 30, 'min_samples': 5}	0.9033	1548
BIRCH	{'branching_factor': 100, 'n_clusters': 5, 'threshold': 0.4}	1.1823	3216
	{'branching_factor': 10, 'n_clusters': 4, 'threshold': 0.3}	1.2641	2515
	{'branching_factor': 20, 'n_clusters': 30, 'threshold': 0.4}	1.2951	1927
Affinity propagation	{'damping': 0.5}	1.1374	1011
	{'damping': 0.8}	1.2493	1265
	{'damping': 0.7}	1.2623	1255
CBMIDE	{'n_components': 2}	1.1731	1689
	{'n_components': 30}	1.1875	1456
	{'n_components': 20}	1.2041	1265
RCBMIDE	{'n_components': 4}	1.0931	2035
	{'n_components': 20}	1.1175	1689
	{'n_components': 10}	1.1540	1356

Bold underlined: Best overall model metrics values.

4.5. Jobs Advertisement Requirements Cluster Analysis for Demand Understanding

In this section, we present the results of our comprehensive analysis of natural language processing (NLP) and clustering-based job profile extraction from a large dataset of over half a million job advertisements. The primary objective of this study was to synthesize meaningful and coherent job profile descriptions by employing advanced clustering techniques and NLP algorithms, thereby facilitating a deeper understanding of the labor market dynamics and the evolving nature of these job requirements. The results of our analysis not only provide valuable insights into the underlying structure and patterns of job profiles, but also highlight the frequency changes of extracted job profiles over time. In doing so, we aimed to contribute to this scientific field and human resource management by offering a data-driven, robust, and scalable approach to identifying the key trends and shifts in the job market. In the visualization presented below (see Figure 4), the automatically extracted job profiles have been depicted, which were generated utilizing clustering techniques and natural language processing methodologies. The results obtained through these methods demonstrated the successful creation of distinct job profiles. A closer examination of the skills identified in the first job profile reveals competencies such as "accounting", "finance",

"economics", and other related areas, which can be reasonably interpreted as describing a job profile within the finance/economics domain.

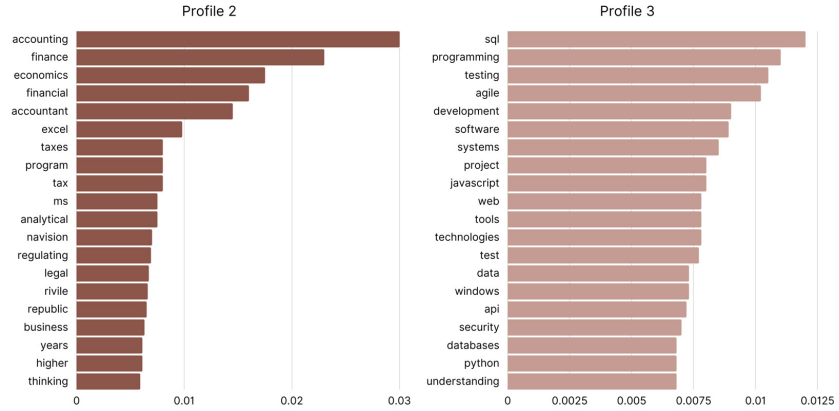

Figure 4. Requirements for extracted profile 2 (finance/accounting specialist) and profile 3 (programmer).

In contrast, the second job profile illustrated skills such as "SQL", "programming", "testing", and "agile", which are indicative of a programmer's profile. Further elucidation of these job profiles, including additional examples, can be found in Appendix A Figure A1. Based on the compiled profiles, it can be seen that several profiles included skills that were more difficult to understand, such as German, cat, and others (see Profile 8 in Appendix A). Such results show that this method does have limitations when the obtained skills are more difficult to interpret. In this case, it was noticeable that we are talking about knowing the German language in the case of the "German" skill. At the same time, "Cat" describes job advertisements related to animal care.

Applying these advanced techniques in clustering and natural language processing has allowed the efficient and accurate extraction of salient job-related skill sets, which can be used to better understand and categorize various roles within different industries, and which could thereby facilitate improved job matching and skill development opportunities for both employers and job seekers.

Two distinct approaches can be employed to automatically generate job profiles using the data acquired through clustering and natural language processing techniques: expert evaluation and profiling, or automated profiling with the assistance of generative artificial intelligence (AI). The following table (see Table 3) offers detailed information regarding the synthesized job profile descriptions based on the previously extracted keywords. These synthesized job descriptions were created using the GPT-4 API version that used the prompt "Generate job profile based on the keywords listed: {profile keywords}". Upon examination, it was apparent that the job profiles were accurately and comprehensively portrayed, capturing not only the primary attributes of the job profiles but also elaborating on the associated information. Utilizing generative AI, these job descriptions can be further expanded and refined, providing a more in-depth understanding of the roles and responsibilities involved in each profile. This approach of combining expert evaluation and profiling, along with generative AI, has the potential to yield highly accurate and detailed job profiles that can better facilitate matching candidates to relevant job opportunities and inform career development and training programs. Additionally, these synthesized job profile descriptions can serve as a valuable resource for human resources professionals, recruiters, and job seekers, enabling them to better understand the various positions available across different industries and the corresponding skill sets required for success in those roles. For full results of the generated profiles see Appendix B Table A1.

The analysis of job profiles and the identification of relevant skills not only involves the instantaneous recognition of specific profiles, but also entails examining the temporal dynamics of these profiles. The graph below illustrates the fluctuations in demand for profile 3 over time (see Figure 5). A significant surge in demand for workers possessing

these particular skills was observed at the beginning of 2022. However, following the outbreak of the war in Ukraine in March 2022, the demand for these specialists sharply declined. By November 2022, there was no noticeable increase in the worker requirement within this profile. A modest uptick in demand for these professionals has been observed since November 2022, and the need for workers with these skills has been anticipated to grow, with the recent upward trend in the latest available data supporting this projection. The utility of this data lies in its capacity to inform various stakeholders, such as employers, job seekers, and policymakers, about the shifting demand for particular skill sets in the job market. By understanding these fluctuations, stakeholders can make informed decisions about workforce development, investment in training programs, and recruitment strategies that align with the evolving needs of the industry. Furthermore, this temporal analysis provides valuable insights into the potential impact of external factors, such as geopolitical events and on-demand for specific job profiles, enabling more adaptive and resilient planning for the future.

Table 3. Synthesized keyword-based job profiles for two job profiles requirements.

Profile	Synthesized Keyword-Based Job Profile
2	This profile is about a finance and accounting professional with a strong background in economics and financial management. They have experience working as an accountant and are skilled in using Excel and other MS Office programs, as well as accounting software such as Navision and Rivilė. Their expertise includes tax preparation, regulation compliance, and legal aspects related to financial operations. They possess analytical and critical thinking abilities, which contribute to their effectiveness in the financial field. The individual has a higher level of education, potentially a degree in business or economics and has several years of experience in industry. They may also have knowledge of the financial regulations specific to a certain republic or region.
3	This profile is about a software development and testing professional with expertise in various programming languages, tools, and technologies. They have experience working with databases, APIs, security, and data management in both web and Windows environments. They are also knowledgeable in Agile project management methodologies and have a strong understanding of software systems and development processes. Their skills include SQL, Python, and JavaScript programming, as well as using various testing and development tools.

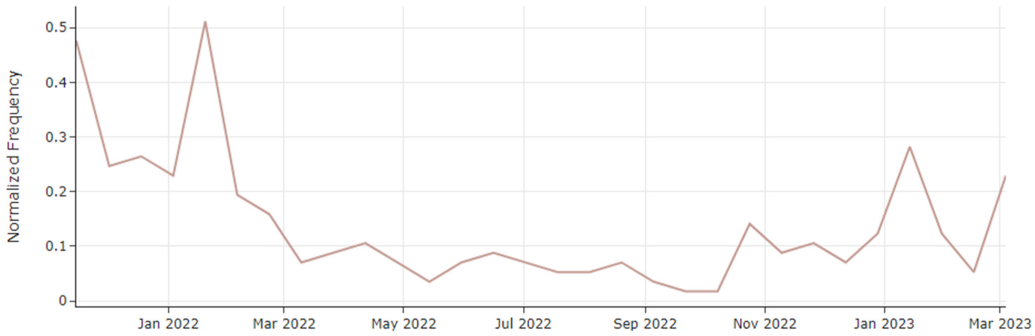

Figure 5. Job requirements profile 3 (programmer) over time.

5. Discussion

The primary objective of this article was to explore the application of natural language processing, data clustering, and other machine learning techniques in determining job profile requirements within the Lithuanian context. The discussion focuses on the key methods to extract data from publicly available unstructured sources. Moreover, the paper delves into the subsequent processing of these data using various vectorization techniques, examining the advantages and disadvantages of each method.

During the research, it was observed that the regex method yielded the most favorable results for data extraction in the case of the Lithuanian job advertisements. Regarding data vectorization and feature extraction, the study utilized two primary techniques—TF-IDF

and sentence transformers (BERT model). Due to the high dimensionality of the feature extracted data (384 dimensions), dimensionality reduction methods were employed, with the ISOMAP approach achieving the best results. A total of 11,000 models for dimensionality reduction were created to assess the different dimensionality reduction methods and their parameters, as well as the optimal parameters set for each method. Nonetheless, other methods, such as UMAP, also demonstrated a satisfactory performance. Speed was the main parameter in the automated job advertisement analyzes, as the UMAP method was selected as the research method. The UMAP method, compared with the other methods such as t-SNE or ISOMAP, demonstrated a better speed, and as mentioned earlier, is important in the daily job advertisement analysis. To have the best speed, the PCA method could have been used, but its results compared with the UMAP would have been of lower quality. Notably, increasing the latent dimensions did not yield improved outcomes in the context of data dimensionality reduction. From 20 latent dimensions, the applied data dimensionality reduction methods did not show better results.

For this reason, 20 latent dimensions could also be used in case of larger system limitations. This study used the 50 latent dimensions obtained with the previously mentioned UMAP method for dimensionality reduction. It was also observed that dimensionality reduction for this specific data, even with an extremely small number of latent dimensions, e.g., 2, 5, that the value of the trustworthiness metric was more than 0.9.

The paper reviewed many different data clustering methods. More than 3000 different models were created for data clustering to optimize the parameters of different methods. The HDBSCAN method proved to be the most effective in data clustering, largely attributable to its hierarchical structure. Additionally, the RCBMIDE method also exhibited relatively strong results, as per the metrics employed in the study, by enabling the elimination of the outliers and the inclusion of data that best represented the information. The obtained job profiles based on both methods were quite similar. However, judging by the metrics, HDBSCAN was used as the main data clustering method. It was also noticeable that when using the most used k-means method, the best clustering result with this method was obtained when the number of clusters was only 5, which was deemed to be extremely small when compiling the number of job profiles of the country. The number of formed profiles must therefore be interpreted.

Upon evaluating the job profiles, it was evident that the extraction quality was high, thereby supporting the utilization of these methods for automated profile research. Based on the compiled profiles, it was observed that some profiles included skills that were more difficult to understand, such as German, cat, and others. Such results demonstrate that this method has limitations when the obtained skills are more difficult to interpret. In this case, it was noticeable that they are talking about knowing the German language in the case of the "German" skill. At the same time, "Cat" describes job advertisements related to animal care. Based on the received job profiles, these job profiles were described using an automated generative artificial intelligence algorithm. Different models of generative artificial intelligence were used to describe these job profiles. The best results were seen with the current latest GPT-4 language model, although the GPT-3.5 model also performed well. Meanwhile, older models of generative artificial intelligence did not have particularly good results, and the created job profile descriptions could be only applied in practice through their additional analysis and improvement.

5.1. Extrapolating this Study to Other Countries

Several considerations should be considered regarding the extrapolation of this study to other countries. Firstly, differences in the language and terminologies used in job advertisements may require adaptations to natural language processing techniques and data extraction methods. For instance, cultural nuances and terminologies specific to a particular country may necessitate adjustments to the regex method to maintain its effectiveness. Secondly, variations in labor market structure and industry composition across countries could affect the applicability of the chosen clustering and dimensionality reduction methods. It would therefore be essential to assess the effectiveness of these

methods in the context of the target country's labor market characteristics to ensure that the extracted job profiles are both accurate and reliable. Lastly, legal, and ethical considerations, such as data privacy regulations and consent requirements for using publicly available data, may differ across countries. Researchers must consider these factors when applying the methodology used in this study in other contexts. In conclusion, although this study successfully extracted job profiles within the Lithuanian context, researchers should be aware of the potential challenges and limitations when attempting to extrapolate the findings to other countries. This methodology may provide valuable insights into job profile requirements across various international contexts by addressing the language, labor market, and legal considerations.

5.2. Future Directions

As we look towards future research, several promising avenues exist for extending the current study on job advertisements and job profiles within a multi-country European context. The following areas should be investigated in future work, allowing for a more comprehensive understanding of the European labor market dynamics. Standardization and harmonization: Developing a standardized framework for job titles and classifications across European countries will facilitate comparative analyzes and enhance our results' generalizability. Adopting an existing system, such as the International Standard Classification of Occupations (ISCO), can provide a foundation for harmonizing job roles and enable more efficient cross-country comparisons. Adaptation to diverse contexts: With the diversity in economic conditions, labor regulations, and employment practices across the European countries, future research should aim to adapt data extraction and processing techniques accordingly. It may involve employing supervised and unsupervised machine learning algorithms to capture the nuances specific to each country and ensure the accuracy of the extracted job profiles. Cross-cultural analysis: Investigating cultural differences in job advertisement language and presentation can provide valuable insights into the variation in job requirements and expectations across the European countries. By examining these cultural aspects, we can better understand the dynamics of the European labor market and develop more tailored strategies for workforce development and talent acquisition. Longitudinal analysis: Conducting a longitudinal analysis of the job advertisements and job profiles can help to identify the trends and shifts in labor market demands over time. This temporal perspective would enable policymakers, employers, and job seekers to anticipate and respond to emerging needs and opportunities more effectively. Integration with labor market indicators: Combining the analysis of job advertisements and job profiles with other labor market indicators, such as unemployment rates, wage levels, and skill shortages, can provide a more holistic view of the European labor market landscape. This integrated approach can support evidence-based decision-making for education, training, and employment policies at both national and regional levels. By pursuing these future research directions, we can further advance the understanding of job profile requirements across various European contexts and contribute to more informed policymaking and workforce development strategies.

Author Contributions: Conceptualization: M.L., V.Š., V.P., J.B., A.S. and A.G; methodology: M.L. and V.Š.; software: M.L.; validation: V.P., J.B., A.S. and A.G.; formal analysis: M.L. and V.Š.; investigation: M.L., V.Š., V.P. and A.G.; resources: M.L., V.Š., V.P., J.B., A.S. and A.G.; data curation: M.L.; writing—original draft preparation: M.L. and V.Š.; writing—review and editing: M.L., V.Š., V.P., J.B., A.S. and A.G.; visualization: M.L. and V.Š.; supervision: V.P., J.B. and A.S.; project administration: V.P., J.B. and A.S.; funding acquisition: V.P., J.B. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This project has received funding from the European Regional Development Fund (project No 13.1.1-LMT-K-718-05-0012) under a grant agreement with the Research Council of Lithuania (LMTLT). Funded as the European Union's measure in response to the Cov-19 pandemic.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the area editor and the reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

This appendix contains information about the skills prevalent in different profiles, i.e., what skills are specific to a certain profile. It is important to note that the top 20 skills and their repetition are presented in the job postings.

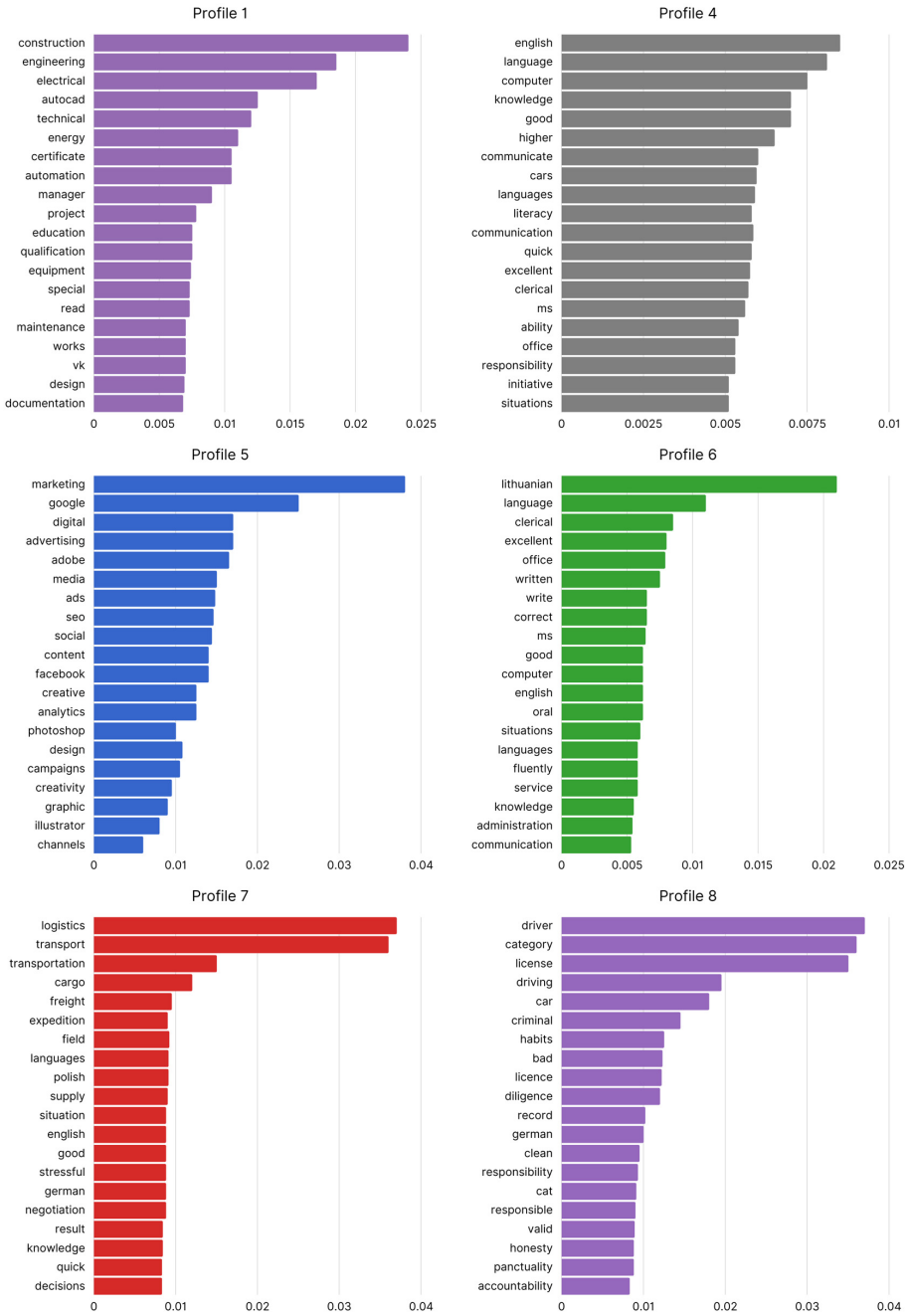

Figure A1. *Cont.*

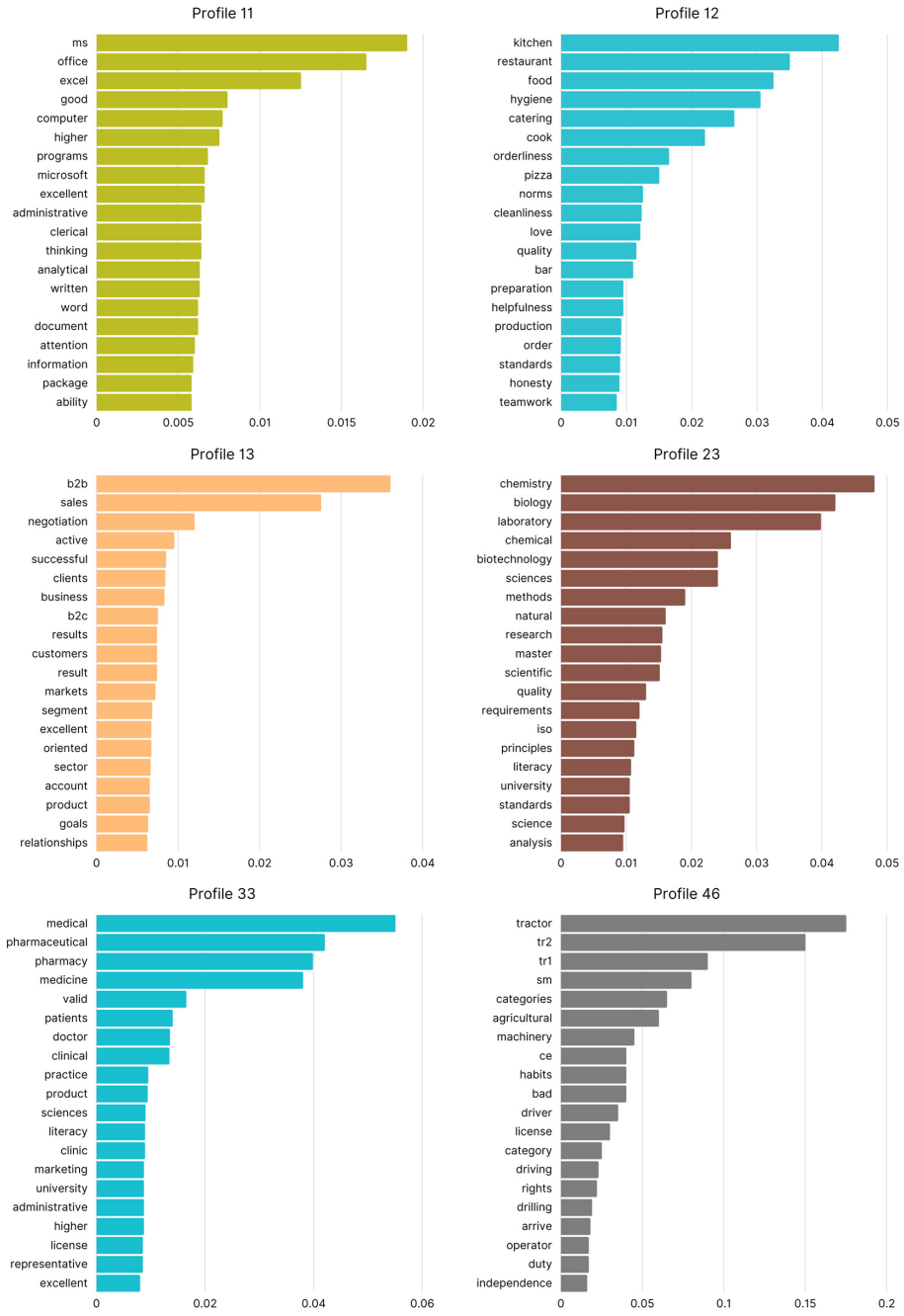

Figure A1. Requirements for extracted profiles.

Appendix B

This appendix describes the job profiles that were created using generative artificial intelligence based on the individual profile skills discussed earlier.

Table A1. Syntesized job profiles descriptions using generative artificial intelligence.

Profile	Synthesized Keyword-Based Job Profile
1	This profile is about a construction and engineering professional with expertise in electrical systems, automation, and energy management. They have experience in project management and are skilled in using AutoCAD for technical design and documentation. This individual holds the relevant education and qualifications, including a certificate in a specialized area. They are knowledgeable in equipment maintenance, handling works related to electrical engineering, and can read and understand technical documentation. Their background also includes managing and designing various construction and engineering projects.

Table A1. *Cont.*

Profile	Synthesized Keyword-Based Job Profile
4	This profile is about a professional with strong English language and communication skills, proficiency in computer literacy, and experienced in clerical and office-related tasks. They have a good knowledge of MS Office tools and can effectively communicate in various situations. The individual is also responsible, takes initiative, and can quickly adapt to new environments. They have a higher level of education and might have an interest in or experience with cars. Their language skills may extend to other languages as well, highlighting their overall linguistic abilities.
5	This profile is about a digital marketing and advertising professional with expertise in various aspects of online marketing, such as Google Ads, SEO, and social media management. They have experience in content creation, campaign management, and analytics, utilizing tools such as Adobe Photoshop and Illustrator for graphic design and creative purposes. Their skills include managing and optimizing advertising campaigns on platforms such as Facebook and other digital channels. They possess a strong creativity and are proficient in using marketing analytics tools to measure the success of their campaigns. Overall, this individual is well-versed in the digital media landscape, and has a deep understanding of how to leverage various platforms and tools to achieve marketing objectives.
6	This profile is about a bilingual professional with fluency in both the Lithuanian and English languages, possessing excellent clerical and administrative skills. They have experience in office administration and are proficient in using MS Office tools and other computer applications. Their strong written and oral communication abilities allow them to excel in various situations, providing excellent service in both languages. They can write and speak fluently and correctly in Lithuanian and English, demonstrating their adaptability in diverse environments. This individual's background includes a good knowledge of administrative tasks and effective communication in multiple languages, making them a valuable asset in any organization requiring multilingual support.
7	This profile is about a logistics and transportation professional with experience in cargo and freight expeditions. They have a strong background in the field of transport and supply chain management, with the ability to handle various situations, including stressful ones. They possess excellent negotiation skills and can make quick, result-oriented decisions. This individual is also proficient in multiple languages, including English, Polish, and German, which allows them to effectively communicate and coordinate in diverse environments. Their knowledge of the logistics sector and expertise in transport make them a valuable asset to any organization involved in the movement of goods and freight.
8	This profile is about a responsible and diligent driver with a valid license for a specific category of vehicles. They have experience driving cars and maintaining a clean criminal record, as well as a good driving record without any bad habits. This individual demonstrates responsibility, honesty, punctuality, and accountability in their work. They may also have knowledge of the German language, which could be beneficial in certain driving situations or locations. Their strong sense of diligence and commitment to safe driving practices make them a reliable and trustworthy candidate for any driving-related job.
12	This profile is about a professional in the kitchen, restaurant, and catering industry who is enthusiastic about food and its preparation. They have experience as a cook, possibly specializing in pizza and other culinary delights. They are committed to maintaining high standards of hygiene, cleanliness, and orderliness in their work environment, adhering to established norms and regulations. This individual values quality in food production and preparation while also demonstrating helpfulness, honesty, and teamwork. Their love for the culinary arts, combined with their dedication to maintaining high standards in the kitchen, makes them an excellent candidate for roles in the food and restaurant industry.
23	This profile is about a professional in the fields of chemistry and biology, with a strong background in laboratory work, biotechnology, and natural sciences research. They have a master's degree from a university, showcasing their expertise in scientific principles and methods. Their experience includes working with chemical analysis, quality control, and adhering to ISO standards and other requirements. This individual is knowledgeable in various scientific techniques and possesses strong literacy in the sciences. Their dedication to maintaining high-quality research and understanding of both chemistry and biology make them an excellent candidate for roles in the scientific and biotechnology industries.
32	This profile is about a professional in the medical and pharmaceutical fields, with experience in pharmacy, clinical practice, and medicine. They have a valid license and a higher degree in the sciences from a university, demonstrating their expertise in the field. This individual possesses excellent literacy in medical and pharmaceutical topics, and has experience working with patients, doctors, and other healthcare professionals. They may also have experience in product marketing and administrative tasks within a clinic or healthcare setting. Their strong background in medicine and pharmaceuticals, along with their dedication to patient care and professionalism, make them an ideal candidate for roles in the healthcare and pharmaceutical industries.

References

1. Nielsen, P.; Holm, J.R.; Lorenz, E. Work policy and automation in the fourth industrial revolution. In *Globalisation, New and Emerging Technologies, and Sustainable Development*; Routledge: Abingdon, UK, 2021; pp. 189–207.
2. Lloyd, C.; Payne, J. Rethinking country effects: Robotics, AI and work futures in Norway and the UK. *New Technol. Work. Employ.* **2019**, *34*, 208–225. [CrossRef]
3. Frey, C.B.; Osborne, M.A. The future of employment: How susceptible are jobs to computerisation? *Technol. Forecast. Soc. Chang.* **2017**, *114*, 254–280. [CrossRef]
4. Quintini, G. *Automation, Skills Use and Training*; Technical Report; OECD Publishing: Paris, France, 2018.
5. Bacher, J.; Tamesberger, D. The Corona Generation: (Not) Finding Employment during the Pandemic. *CESifo Forum* **2021**, *22*, 3–7.
6. Arntz, M.; Gregory, T.; Zierahn, U. *The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis*; OECD Publishing: Paris, France, 2016.
7. OECD. *OECD Skills Studies OECD Skills Strategy Lithuania Assessment and Recommendations*; OECD Publishing: Paris, France, 2021.
8. Hershbein, B.; Kahn, L.B. Do recessions accelerate routine-biased technological change? Evidence from vacancy postings. *Am. Econ. Rev.* **2018**, *108*, 1737–1772. [CrossRef]
9. Verma, A.; Lamsal, K.; Verma, P. An investigation of skill requirements in artificial intelligence and machine learning job advertisements. *Ind. High. Educ.* **2022**, *36*, 63–73. [CrossRef]
10. Deming, D.; Kahn, L.B. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *J. Labor Econ.* **2018**, *36*, S337–S369. [CrossRef]
11. Boselli, R.; Cesarini, M.; Mercorio, F.; Mezzanzanica, M. Using machine learning for labour market intelligence. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, 18–22 September 2017; pp. 330–342.
12. Brynjolfsson, E.; Horton, J.J.; Ozimek, A.; Rock, D.; Sharma, G.; Tuye, H.-Y. *COVID-19 and Remote Work: An Early Look at US Data*; National Bureau of Economic Research: Cambridge, MA, USA, 2020.
13. Autor, D.; Reynolds, E. *The Nature of Work after the COVID Crisis: Too Few Low-Wage Jobs*; Brookings Institution: Washington, DC, USA, 2020.
14. Kramer, A.; Kramer, K.Z. The potential impact of the COVID-19 pandemic on occupational status, work from home, and occupational mobility. *J. Vocat. Behav.* **2020**, *119*, 103442. [CrossRef]
15. Fabo, B. The Corona-Inducted Shift Towards Intermediate Digital Skills Across Occupations in Slovakia. In *Digital Labour Markets in Central and Eastern European Countries*; Routledge: Abingdon, UK, 2023; pp. 37–48.
16. Rebele, J.E.; Pierre, E.K.S. A commentary on learning objectives for accounting education programs: The importance of soft skills and technical knowledge. *J. Account. Educ.* **2019**, *48*, 71–79. [CrossRef]
17. Brunello, G.; Wruuck, P. Skill shortages and skill mismatch: A review of the literature. *J. Econ. Surv.* **2021**, *35*, 1145–1167. [CrossRef]
18. Wagner, J.A.; Hollenbeck, J.R. *Organizational Behavior: Securing Competitive Advantage*; Routledge: Abingdon, UK, 2020.
19. Ibrahim, R.; Boerhannoeddin, A.; Bakare, K.K. The effect of soft skills and training methodology on employee performance. *Eur. J. Train. Dev.* **2017**, *41*, 388–406. [CrossRef]
20. Heckman, J.J.; Kautz, T. Hard evidence on soft skills. *Labour Econ.* **2012**, *19*, 451–464. [CrossRef] [PubMed]
21. Asbari, M.; Purwanto, A.; Ong, F.; Mustikasiwi, A.; Maesaroh, S.; Mustofa, M.; Hutagalung, D.; Andriyani, Y. Impact of hard skills, soft skills and organizational culture: Lecturer innovation competencies as mediating. *EduPsyCouns J. Educ. Psychol. Couns.* **2020**, *2*, 101–121.
22. De Mauro, A.; Greco, M.; Grimaldi, M.; Ritala, P. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Inf. Process. Manag.* **2018**, *54*, 807–817. [CrossRef]
23. Autor, D.H. Work of the Past, Work of the Future. *AEA Pap. Proc.* **2019**, *109*, 1–32. [CrossRef]
24. Groysberg, B.; Lee, J.; Price, J.; Cheng, J. The leader's guide to corporate culture. *Harv. Bus. Rev.* **2018**, *96*, 44–52.
25. Isphording, I.E. Language and labor market success. In *International Encyclopedia of the Social & Behavioral Sciences*; Institute of Labor Economics: Bonn, Germany, 2014.
26. Berg, P.; Kossek, E.E.; Misra, K.; Belman, D. Work-life flexibility policies: Do unions affect employee access and use? *ILR Rev.* **2014**, *67*, 111–137. [CrossRef]
27. Bilal, M.; Malik, N.; Khalid, M.; Lali, M.I.U. Exploring industrial demand trend's in Pakistan software industry using online job portal data. *Univ. Sindh J. Inf. Commun. Technol.* **2017**, *1*, 17–24.
28. Clarke, M. Rethinking graduate employability: The role of capital, individual attributes and context. *Stud. High. Educ.* **2018**, *43*, 1923–1937. [CrossRef]
29. Mahany, A.; Khaled, H.; Elmitwally, N.S.; Aljohani, N.; Ghoniemy, S. Negation and Speculation in NLP: A Survey, Corpora, Methods, and Applications. *Appl. Sci.* **2022**, *12*, 5209. [CrossRef]
30. Kalyan, K.S.; Rajasekharan, A.; Sangeetha, S. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv* **2021**, arXiv:2108.05542.
31. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

32. Fellbaum, C. WordNet. In *Theory and Applications of Ontology: Computer Applications*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 231–243.
33. Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J. Gpt-neox-20b: An open-source autoregressive language model. *arXiv* **2022**, arXiv:2204.06745.
34. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 50–70. [CrossRef]
35. Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; Mikolov, T. Fasttext. zip: Compressing text classification models. *arXiv* **2016**, arXiv:1612.03651.
36. Salton, G. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, NY, USA, 1983.
37. Reimers, N.; Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv* **2019**, arXiv:1908.10084.
38. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013.
39. Arora, S.; Liang, Y.; Ma, T. A simple but tough-to-beat baseline for sentence embeddings. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
40. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
41. Harris, Z.S. Distributional structure. *Word* **1954**, *10*, 146–162. [CrossRef]
42. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]
43. Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1188–1196.
44. Bellman, R.; Kalaba, R. On adaptive control processes. *IRE Trans. Autom. Control* **1959**, *4*, 1–9. [CrossRef]
45. Wang, Y.; Yao, H.; Zhao, S. Auto-encoder based dimensionality reduction. *Neurocomputing* **2016**, *184*, 232–242. [CrossRef]
46. Dong, Y.; Du, B.; Zhang, L.; Zhang, L. Dimensionality reduction and classification of hyperspectral images using ensemble discriminative local metric learning. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2509–2524. [CrossRef]
47. Thomas, R.; Judith, J. Hybrid dimensionality reduction for outlier detection in high dimensional data. *Int. J.* **2020**, *8*, 5883–5888.
48. Li, M.; Wang, H.; Yang, L.; Liang, Y.; Shang, Z.; Wan, H. Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction. *Expert Syst. Appl.* **2020**, *150*, 113277. [CrossRef]
49. McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv* **2018**, arXiv:1802.03426.
50. Sumithra, V.; Surendran, S. A review of various linear and non linear dimensionality reduction techniques. *Int. J. Comput. Sci. Inf. Technol.* **2015**, *6*, 2354–2360.
51. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
52. Du, X.; Zhu, F. A novel principal components analysis (PCA) method for energy absorbing structural design enhanced by data mining. *Adv. Eng. Softw.* **2019**, *127*, 17–27. [CrossRef]
53. Iannucci, L. Chemometrics for data interpretation: Application of principal components analysis (PCA) to multivariate spectroscopic measurements. *IEEE Instrum. Meas. Mag.* **2021**, *24*, 42–48. [CrossRef]
54. Fan, C.; Sun, Y.; Zhao, Y.; Song, M.; Wang, J. Deep learning-based feature engineering methods for improved building energy prediction. *Appl. Energy* **2019**, *240*, 35–45. [CrossRef]
55. Van Der Maaten, L. t-SNE. 2019. Available online: https://lvdmaaten.github.io/tsne (accessed on 25 March 2023).
56. Linderman, G.C.; Steinerberger, S. Clustering with t-SNE, provably. *SIAM J. Math. Data Sci.* **2019**, *1*, 313–332. [CrossRef]
57. Kobak, D.; Linderman, G.C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* **2021**, *39*, 156–157. [CrossRef]
58. Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I.W.; Ng, L.G.; Ginhoux, F.; Newell, E.W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **2019**, *37*, 38–44. [CrossRef] [PubMed]
59. Böhm, J.N.; Berens, P.; Kobak, D. A unifying perspective on neighbor embeddings along the attraction-repulsion spectrum. *arXiv* **2020**, arXiv:2007.08902.
60. Arunkumar, N.; Mohammed, M.A.; Abd Ghani, M.K.; Ibrahim, D.A.; Abdulhay, E.; Ramirez-Gonzalez, G.; de Albuquerque, V.H.C. K-means clustering and neural network for object detecting and identifying abnormality of brain tumor. *Soft Comput.* **2019**, *23*, 9083–9096. [CrossRef]
61. Jain, A.K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **2010**, *31*, 651–666. [CrossRef]
62. Singh, A.; Yadav, A.; Rana, A. K-means with Three different Distance Metrics. *Int. J. Comput. Appl.* **2013**, *67*, 13–17. [CrossRef]
63. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **2017**, *42*, 1–21. [CrossRef]
64. Campello, R.J.; Moulavi, D.; Zimek, A.; Sander, J. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data* **2015**, *10*, 1–51. [CrossRef]
65. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Rec.* **1996**, *25*, 103–114. [CrossRef]

66. Dueck, D.; Frey, B.J. Non-metric affinity propagation for unsupervised image categorization. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
67. Guan, R.; Shi, X.; Marchese, M.; Yang, C.; Liang, Y. Text clustering with seeds affinity propagation. *IEEE Trans. Knowl. Data Eng.* **2010**, *23*, 627–637. [CrossRef]
68. Fang, Q.; Sang, J.; Xu, C.; Rui, Y. Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning. *IEEE Trans. Multimed.* **2014**, *16*, 796–812. [CrossRef]
69. Ng, A.; Jordan, M.; Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2001**, *14*, 849–856.
70. Janani, R.; Vijayarani, S. Text document clustering using spectral clustering algorithm with particle swarm optimization. *Expert Syst. Appl.* **2019**, *134*, 192–200. [CrossRef]
71. Lukauskas, M.; Ruzgas, T. A New Clustering Method Based on the Inversion Formula. *Mathematics* **2022**, *10*, 2559. [CrossRef]
72. Lukauskas, M.; Ruzgas, T. Reduced Clustering Method Based on the Inversion Formula Density Estimation. *Mathematics* **2023**, *11*, 661. [CrossRef]
73. Venna, J.; Kaski, S. Neighborhood preservation in nonlinear projection methods: An experimental study. In Proceedings of the Artificial Neural Networks—ICANN 2001: International Conference, Vienna, Austria, 21–25 August 2001; pp. 485–491.
74. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
75. Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [CrossRef]
76. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.-Theory Methods* **1974**, *3*, 1–27. [CrossRef]