STRATEGIC MANAGEMENT, Vol. XX (202x), No. x, pp. 0xx-0xx

DOI: 10.5937/StraMan2500008G

Received: April 11, 2025
Revised: June 04, 2025
July 01, 2025
Accepted: July 03, 2025
Published online: September 23, 2025

An NLP approach to skill analysis in ICT job advertisements from a gender perspective

Olivera Grljević

University of Novi Sad, The Faculty of Economics in Subotica, Subotica, Republic of Serbia
https://orcid.org/0000-0002-6028-1153

Tatjana Kecojević

The University of Manchester, School of Social Sciences, Manchester, United Kingdom
https://orcid.org/0009-0003-7074-6879

Abstract

Background: The gender gap in the Information and Communication Technology (ICT) sector is strongly pronounced across Europe, including Serbia. Women are underrepresented in ICT education, entry positions, and throughout career trajectory. Although surveys and official statistics are commonly used to study gender disparities and skills in demand, job advertisements remain an underexplored organic data source in these studies. They can provide valuable insights into required hard and soft skills and a basis for measuring the influence of used language in job descriptions on gender distribution of applicants.

Purpose: This paper investigates whether linguistic patterns and the framing of skill requirements in ICT job advertisements correlates with the gender distribution of applicants. The aim is to understand how specific wording or skill emphasis may differently attract male or female candidates.

Study design/methodology/approach: Authors analyse 3,643 ICT job advertisements from Serbia with associated data on self-reported gender of applicants. Using a five-step methodology, the authors apply exploratory text analysis and natural language processing techniques, including n-gram analysis, feature engineering, and co-occurrence networks, to identify hard and soft skills patterns across male- and female-majority job postings.

Findings/conclusions: The results offer insights into the language characteristics used in ICT job descriptions and gender-specific keywords reflecting hard and soft skills. Job advertisements that attract more male candidates use engineering-oriented terminology, such as programming and DevOps tools. In contrast, advertisements drawing more female candidates emphasise collaboration, teamwork, and emotional intelligence. Soft skills are more frequent in female-majority advertisements, while male-majority advertisements focus on narrower technical domains. Empirical findings suggest that how job advertisements frame requirements can reinforce gendered perceptions of role suitability. This has practical implications for human resource management and recruitment strategies in the ICT industry.

Limitations/future research: Limitations include the use of percentage-based gender data and the evolving nature of the ICT labour market. Future research will expand the dataset, improve gender classification, and explore longitudinal trends to track changes over time.

Keywords

ICT labour market, job advertisements, gender gap, NLP, exploratory text analysis, Serbia, soft skills, hard skills

Introduction

Digital transformation has reshaped labour market demands. According to the European Commission (2020) report, 90% of jobs require digital skills or expertise in science, technology, engineering, and mathematics (STEM). Professions reliant on these competencies, particularly in the Information and Communication Technology (ICT) sector, are driving economic development. Women's

participation in the workforce is recognized by the European Commission as a factor that strengthens national economies (European Commission 2020). However, the European ICT sector exhibits a pronounced gender gap. According to the DESI 2022 report, men constitute an average of 80% of ICT specialists in the European Union (European Commission, 2022). Some factors contributing to such disparity are educational barriers and the impact of artificial intelligence (AI) and automation on employment.

In the European Union, women comprise only 18% of ICT students, with 33% of STEM graduates (European Commission, 2022). In general, there are more women graduates in Serbia, but only 27% of them in ICT-related studies (Grljević et al., 2019), leading to missed talent and innovation opportunities (Kukić Đorđević & Čolić Mihajlović, 2023). Women's underrepresentation in ICT and STEM education spans to the workforce and poses a challenge for gender equality (World Economic Forum, 2020). In the EU, women constitute 18.9% of total employed ICT specialists and in Serbia this number is as high as 23.3% (Kukić Đorđević & Čolić Mihajlović, 2023). With career advancement gender gap is further pronounced. Women are underrepresented from entry-level positions to executive roles (Taplett et al., 2018). Research emphasizes low employment rates for women in ICT and higher attrition rates (Quirós, et al., 2018; Scott & Kapor Klein, 2017; Ashcraft et al., 2016). In Serbia, women are rarely present in high-ranking positions and most often apply for internships. Among already employed women, they constitute 55% of junior ICT specialists, 43% mediors, and only 2% seniors (Kukić Đorđević & Čolić Mihajlović, 2023). The same authors point to decrease of the share of women along corporate ladder – 17% of ICT managers are women, while the share of women in directors' positions in the ICT sector in Serbia is below 10%. While automation and digital transformation generate high-paying ICT and AI jobs, they simultaneously jeopardize clerical roles, disproportionally affecting women (Brussevich, et al., 2018), leading to pay gap and economic disparity. An increasing pay gap is present in Serbia, as well. According to Kukić Đorđević and Čolić Mihajlović (2023), general gender pay gap from 9.8% in 2018 reached 14.4% in 2022. Analysing earnings by education levels or occupations reveals an even more pronounced gap in Serbia, with ICT occupations leading.

Studying the gender gap is thus a relevant issue. However, it is insufficient to study how pronounced the gap is, but it is also necessary to deal with the causes, as well as the characteristics of the gender gap itself. Understanding its characteristics is vital for designing policies that address workforce challenges and improve access for women to well-paid careers and leadership opportunities.

The organic nature and easy accessibility contribute to job advertisements becoming the basis for analysis of skill requirements on labour market (Valavosiki et al., 2019). The literature indicates that content analysis of job advertisements provides good insights into required hard and soft skills (Lovaglio et al., 2018; Korbel, 2018; Valavosiki et al., 2019; Ilich & Akilina, 2017; Cosgrove et al., 2024). However, job advertisements are not utilized in the analysis of the relationship between required skills across candidate's genders, instead authors rely on surveys and official statistics (Hossain, et al., 2023; Bradić-Martinović & Banović, 2018; Bradić-Martinović, et al., 2024; Jevtić et al., 2023; Lazarević-Moravčević et al., 2023). Potential of more sophisticated data analysis techniques, such as text mining, natural language processing (NLP) techniques, and machine learning is reflected in their ability to more precisely identify and classify the required skills from advertisements (Jaiswal et al., 2025; Nasir et al., 2020; Pejic-Bach et al., 2020; Bäck et al., 2021), as well as to identify gender language patterns (Hu, et al., 2022). However, these techniques are rarely used to identify gender-specific ICT skills, such as (Simon et al., 2023), given that publicly available job advertisements are not disclosing information on candidate's gender. This leads to researchers rely on statistical data or surveys even though it is restricted to individuals willing to participate, which introduces potential bias. Consequently, there is a lack of studies that integrate these aspects using advanced analytical methods, as well as datasets that simultaneously capture labour market demand, supply, and candidates' socio-demographic characteristics. To the best of our knowledge, Serbia's ICT market has not been analysed and characterized in this manner. This study aims to contribute to bridging this gap.

Given the previously stated, the authors in this paper aim to characterize the gender gap in Serbia's ICT labour market through skill exploration and identification of possible gender-specific competencies. An empirical study addresses the following research questions:

RQ1: How do linguistic patterns and keywords in ICT job advertisements influence the gender distribution of applicants?
RQ2: What required skills in ICT job advertisements differentiate applicants by gender?

To answer these research questions authors utilize a unique dataset – 3,643 organic ICT job advertisements in Serbia with associated data on self-reported gender distribution of applicants – and apply a more comprehensive methodological approach. Self-reported gender data is limited to binary categories (female/male), based on the data provider's platform structure. Methodology comprises five steps, data extraction, data transformation and cleaning, n-gram analysis, feature engineering, and identification and interpretation of salient features. Through these steps, the authors combine content analysis (Krippendorff, 2013; Hsieh & Shannon, 2005), exploratory text analysis (Tukey, 1977; Allen et al., 2018), and natural language processing techniques (Jurafsky & Martin, 2025; Farzindar & Inkpen, 2015). Such combined approach to job advertisement analysis allows for quality text preparation, extraction of useful keywords and skills from texts, and their quantifications in order to identify patterns pertaining to gender distribution of applicants. The results provide insights into the linguistic characteristics of ICT job descriptions, highlighting how certain keywords reflecting hard and soft skills correlate with the gender composition of applicants. Job advertisements that include engineering-oriented terminology tend to receive higher engagement from male candidates, while those emphasising collaboration, teamwork, and emotional intelligence are more often associated with female applicant engagement. These empirical findings have practical implications for human resource management and recruitment strategies in the ICT industry. While this study focuses on gender-based patterns, future research could benefit from adopting an intersectional perspective examining how gender interacts with race, socio-economic status, and other identity dimensions in shaping digital labour market outcomes (Crenshaw, 1989; Noble, 2018).

The paper is structured as follows. Section 1 provides an overview of related literature. Section 2 presents the research methodology. Section 3 presents the results, while in the Section 4, the authors discuss results and reflect on proposed research questions. The last section refers to conclusions with references to practical implications and limitations of the research, as well as the future work.

1. Related literature

The organic nature and easy accessibility (Valavosiki et al., 2019) contribute to job advertisements becoming the basis for analysis of skill requirements and the gender gap in the ICT sector. Contrary to traditional methods to researching skill requirements, such as questionnaires or surveys, job advertisements provide insights into current labour market demands, specific requirements for technical (hard) and soft skills, and reflect dynamic changes in the industry. Author (Vieira da Cunha, 2009) was among the first to identify the analysis of collection of job advertisements can reveal long-term trends and shifts within a profession, which could provide measurable and comparable research insights. When correlated with gender data, analysis of job advertisements help identify gender-specific language patterns in the formulation of requirements for certain positions (Gaucher et al., 2011; Hu, et al., 2022), which influence the candidate's perception of which professions are "suitable" for a particular gender. Related literature is structured into two subsections to reflect on research studying ICT skills in demand from job advertisements and on research on gender disparities from the perspective of digital skills.

1.1. Analysis of ICT skills from job advertisements

Previous research focusing on ICT labour market demands, examined hard (Lovaglio et al., 2018) and soft skills (Korbel, 2018; Valavosiki et al., 2019; Ilich & Akilina, 2017; Pažur Aničić & Arbanas, 2015) separately, with more emphasis on the latter, given their relevance for ICT professionals is recognized early on (Purao & Suen, 2010; Zhang, 2012). Croatian companies sought soft skills almost twice as often as technical skills (Pažur Aničić & Arbanas, 2015). Hungarian and Serbian employees face higher expectations concerning soft skills compared to technical or hard skills (Strugar Jelača et al., 2025). Such findings point to the role educational institutions have in balancing study programmes and equipping students with both technical and non-technical skills, ensuring successful transition from university education to the labour market. A more advanced approach to studying skill gap present in university curricula with respect to AI industry

demands is present in the (Jaiswal et al., 2025) study. They used frequency analysis, machine learning, and NLP. Results indicate well-balanced AI curriculum in technical skills (e.g., programming, machine learning), while gap is present in data science, mathematics, and statistics.

Research points that communication, problem-solving, and team work are most sought for competences in job advertisements among soft skills (Pažur Aničić & Arbanas, 2015; Korbel, 2018; Valavosiki et al., 2019; Ilich & Akilina, 2017). Authors (Valavosiki et al., 2019) associated required soft skills with six ICT job families recognized in the European e-Competence Framework – *Business Management*, *Technical Management*, *Design*, *Development*, *Service and Operations*, *Support*. Their findings indicate that communication, problem-solving, and teamwork are consistently sought in all six job families and that different job families require some specific soft skills, such as *Business or Technical Management* requires analytical and organisational skills, *Design and Support* job families presentation skills, *Development* requires self-motivation and desire to learn, while *Service and Operation* job family emphasize the need for customer orientation and independent work. These findings indicate the necessity for balanced skill development in the ICT sector, both during education and throughout professional development. Authors Cosgrove et al. (2024) mapped DigComp framework, (Vuorikari et al., 2022), to European Skills, Competencies, Qualifications and Occupations skill descriptors used in a database of European online job advertisement (OJA). In this way authors aligned digital skills employers demand from employees with education and training supply. DigComp covered 54.5% of skills mentioned in OJA (2018–2022), however, the coverage is uneven. The greater emphasis is on information and data literacy, content creation, and communication than on safety or problem-solving. The extent to which these skills are explicitly linked to digital environments varies, with information and data literacy more explicitly digital than communication and collaboration.

These studies utilize content analysis and frequencies, while more sophisticated analytical approaches could be used, such as text mining. Text mining is utilized to predict attractiveness of job advertisement based on job descriptions (Yunlu, 2023), to identify skills in demand for particular types of occupation, such as analysts positions (Nasir et al., 2020), or to identify industry specific knowledge, such as the case study of Industry 4.0 (Pejic-Bach et al., 2020), or to study trends, transitions in the job markets and skill demands, such as the case study of Finnish job market and investigation of the emergence of AI-related jobs (Bäck et al., 2021).

2.2. Gender digital divide

Authors rely on official statistics, such as Eurostat data (Martínez-Cantos, 2017) or surveys (Hossain, et al., 2023; Bradić-Martinović & Banović, 2018), if studying digital skills from the gender perspective. Such data provides information on both skill and gender. Global gender gap report by (World Economic Forum, 2020) differs in terms of using LinkedIn data. Based on five-year employment trends on the LinkedIn platform and similarities in required skills, eight job clusters have been defined with a growing employment trend: *People and Culture*, *Content Production*, *Marketing*, *Sales*, *Product Development*, *Data and AI*, *Engineering*, and *Cloud Computing*. This study indicates that women are underrepresented in technical-intense clusters, i.e., *Data and AI* (26%), *Engineering* (15%), and *Cloud computing* (12%). However, *Data and AI* can use the potential from the available talent pool, as women comprise 31% of other professions characterized by skills relevant to data and AI-related positions. Thus, it is possible to increase the proportion of female data scientists.

The issue of the gender gap in the Serbian labour market was studied in general by analysing the gender digital divide (Bradić-Martinović & Banović, 2018), through effects of digitalization and skills on inclusion of women in labour market (Jevtić et al., 2023), or by focusing on a specific sector – the authors Bradić-Martinović et al. (2024) explore gender disparities in Serbian tourism sector, while Lazarević-Moravčević et al. (2023) study gender inequality in science and education. These studies indicate presence of gender divide, such as 62.6% of women have low or no digital skills, contrary to 46.7% of men. Women's digital literacy is prerequisite for reduction of gender gap and potential increase of the number of female entrepreneurs. To the best of our knowledge, the most comprehensive study on the position of women in Serbia's ICT sector is the report published by The United Nations Development Programme. It offers detailed analysis of official statistical data, analysis of causes of gender imbalance pointing to education-related issues, such as stereotypes, lack of female role models,

lack of practical work, support, and motivation, and labour market-related issues, such as lack of female mentors, male-dominated environments and male-culture in companies, imbalance between work-home-family workload, perception gap (Kukić Đorđević & Čolić Mihajlović, 2023).

These studies are based on utilizing and analysing official statistical data. If more sophisticated techniques are used, such as machine learning, natural language processing, or text mining, they are utilized to decipher the language used in ICT job advertisement and link them to gender-based attributes. The pioneering research by Gaucher et al. (2011) reveals that masculine wording, used in male-dominated occupations, are less appealing to women indicating that gendered wording in job advertisements signals a candidate's fit for the position. Gender bias in job advertisements is studied in (Hu, et al., 2022) offering a mitigation approach in order to debias job posting text, while authors (Simon et al., 2023) used a LinkedIn profiles dataset on job candidates fit for IT-related positions to identify presence of gender bias caused by gender differences in textual self-presentation in LinkedIn.

2. Methodology

In an attempt to understand the gender gap present in the Serbia's ICT market, research aims to conduct NLP-based exploratory analysis of ICT job advertisements, with the specific focus on interpreting results in the context of gender distribution among applicants. The primary objective is to identify distinct skills within job advertisements and investigate whether these skills exhibit correlations with the gender distribution of applicants.

Authors use the dataset containing 3,643 ICT job advertisements published on Serbia's largest employment portal, *Poslovi.infostud* – a part of Inspira group and Alma Career. Job advertisement are linked to applicants' self-reported gender, derived from applicants' disclosures regarding their sex and is expressed as the percentage of male and female applicants per job posting, along with the percentage of applicants who did not disclose this information. The data on both job advertisements and applicants' gender data were provided by the company *Poslovi Infostud*, fully respecting data privacy regulations and without exposing any identifiable user information. The authors of this paper deployed named entity recognition as additional mechanism for ensuring anonymity of the data and removed all instances of named entities, such as cities, organisations.

The research presented in this paper is exploratory in nature and the research methodology is rooted in the methodology proposed in (Allen et al., 2018) and adapted to our case study. It comprises the following steps, as illustrated in Figure 1: (1) data extraction enables abbreviation of job advertisements retaining only relevant information, (2) data transformation and cleaning improves text quality enabling drawing more reliable conclusions from graphical representations, (3) n-gram analysis offers direction to frequency-based filtering, domain-dependent stopword detection, and identification of informative n-grams useful for further analysis, (4) feature engineering, and (5) identification and interpretation of salient features. Through these steps authors combine content analysis (Krippendorff, 2013; Hsieh & Shannon, 2005), exploratory text analysis (Tukey, 1977; Allen et al., 2018), and natural language processing techniques (Jurafsky & Martin, 2025; Farzindar & Inkpen, 2015). Such combined approach to job advertisement analysis allows for quality text preparation, extraction of useful keywords and skills from texts, and their quantifications in order to identify patterns pertaining to gender distribution of applicants. Each step is presented in more detail in the subsequent subsections.

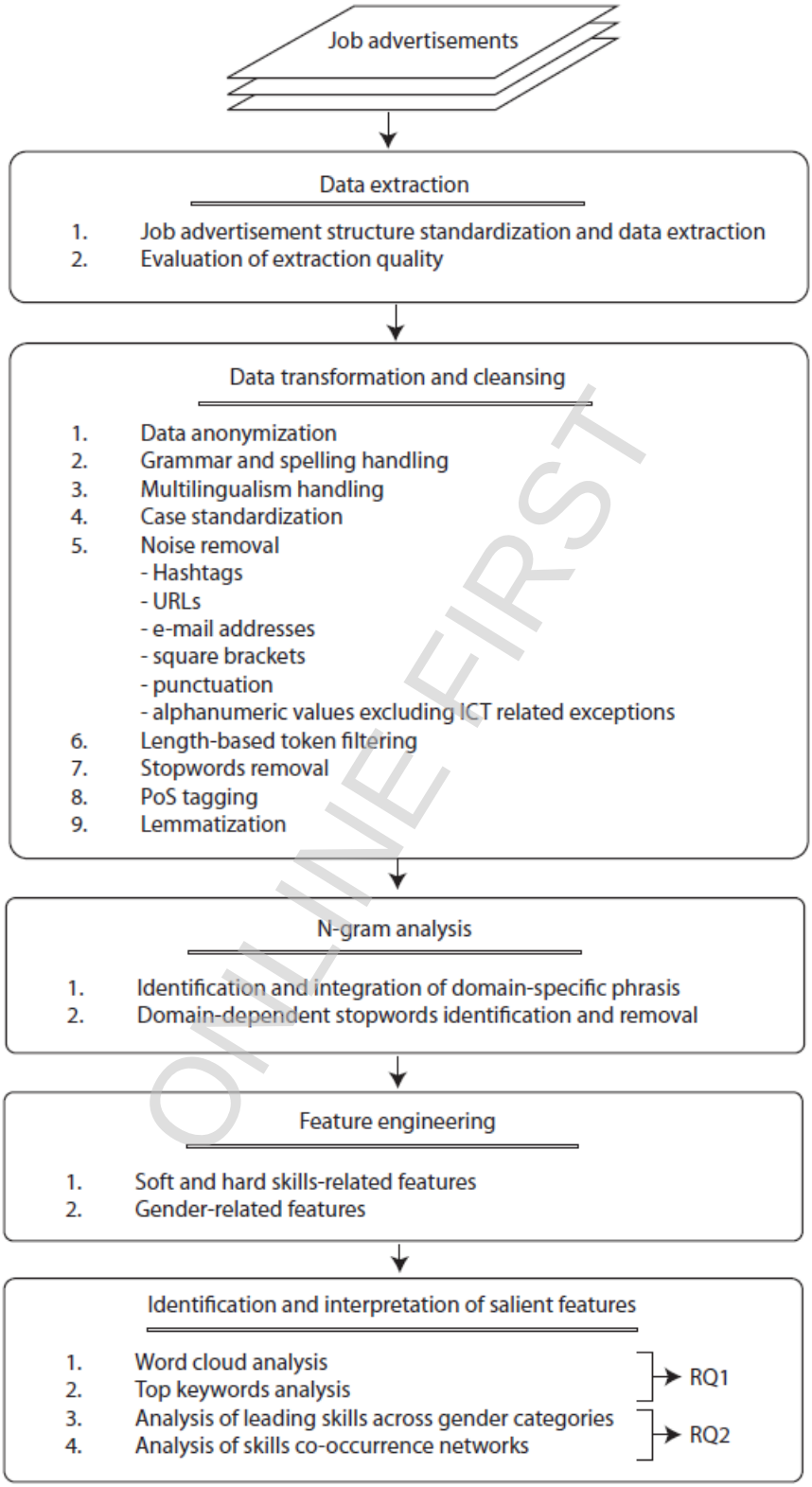


Figure 1 Methodology framework
Source: the authors

2.1. Data extraction

Job advertisements are commonly structured into five segments: a) description of a company, b) description of the position, c) required skills or knowledge, d) conditions a company offers to the future employees, and e) instructions on how to apply for the job. To address research questions, the focus of our analysis is on description of the position and requirements. Extraction of these particular text segments from full texts of job advertisements was challenging due to variations present in the way companies structure job postings. Not all of the job advertisements contained all five of the segments, nor the segments were always structured in the same order. For this reason we automated data extraction using Python, the ChatGPT-3.5 API, and structured prompting, similar to the approach in (Jiang, et al., 2024). This enabled extraction of position-related information into five segments: 1) position naming, 2) the narrative about the role or the job description, 3) responsibilities, 4) required technologies, 5) skills, both soft and hard. We have also anticipated the sixth, optional segment dedicated to non-specific details related to the job function. Figure 2 illustrates the effect of data extraction.

To evaluate the quality of extraction, the authors randomly selected 124 job advertisements for manual evaluation. The first author of the paper along with the expert in the field, marked as annotator A1 and A2, respectfully, independently evaluated each text by comparing original job advertisement text with the extraction, taking into account that only information about job posting, required skills, and responsibilities should remain in the excerpt.

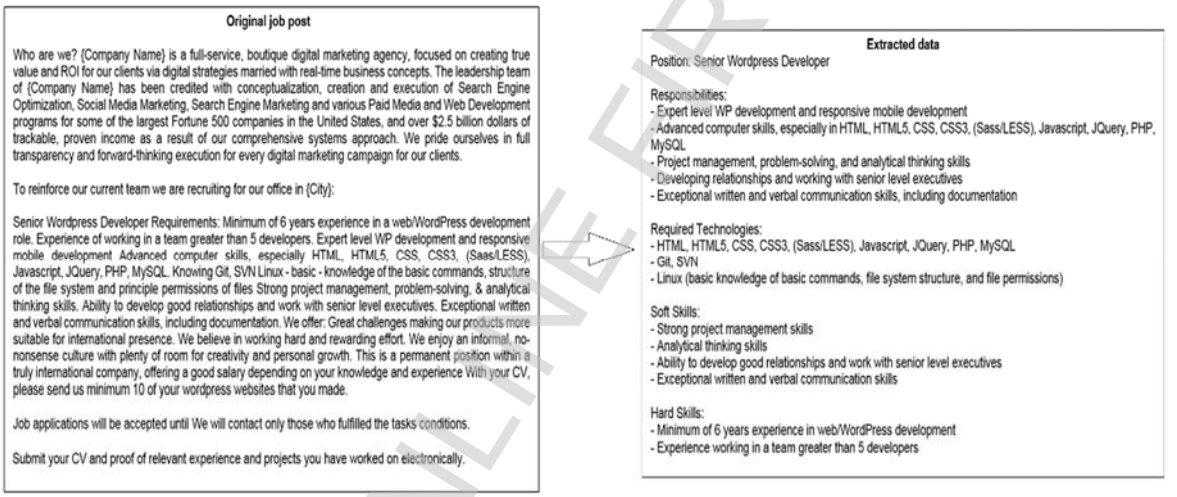


Figure 2 Illustration of data extraction from job advertisement

Source: the authors

Both evaluators rated each extraction with values 1, indicating the extraction was complete and successful, or 0, indicating the extraction was incomplete and unsuccessful. Cohen kappa κ was applied to evaluate the inter-rater agreement (IRA), or the extent to which evaluators make the same judgment on the successfulness of extraction (Rau & Shih, 2021; McHugh, 2012; Tan et al., 2024). "Kappa is a statistical measure of the agreement in assignment of units to categories in excess of what would be expected by chance, based on each rater's tendency to assign units to each category" (Rau & Shih, 2021). It is calculated as:

$$\kappa = \frac{Pr_{(a)} - Pr_{(e)}}{1 - Pr_{(e)}} \quad (1)$$

where $Pr_{(a)}$ and $Pr_{(e)}$ are the actual observed agreement and expected frequencies of agreement, respectively.

Table 1 indicates the number of excerpts assigned to each category by each evaluator: 119 excerpts were judged by both evaluators as successfully extracted text, and 3 excerpts as unsuccessfully extracted, indicating an observed total percent agreement $Pr_{(a)}$ of 98.39% ($Pr_{(a)} = (119+3) / 124$).

As shown in Table 2, based on the evaluators' overall preference for each category, it is expected that 116 excerpts are categorized by both evaluators in category 1 ($(120/124)*(120/124)*124 = 116$), and 0 in category 2 (4/124*4/124*124). An expected

agreement $Pr_{(e)}$ is 93.55% by chance alone ($Pr_{(e)} = (116+0) / 124$).

Table 1 Observed agreement and (disagreement), based on the ratings assigned to each excerpt by evaluators A1 and A2

		Evaluator A2		
	Category	**1**	**0**	**Total**
Evaluator A1	**1**	119	(1)	120
	0	(1)	3	4
	Total	120	4	124

Source: Authors

Table 2 Expected agreement and (disagreement) by chance, based on the proportion of excerpt assigned to each category by each evaluator

		Evaluator A2		
	Category	**1**	**0**	**Total**
Evaluator A1	**1**	116	(4)	120
	0	(4)	0	4
	Total	120	4	124

Source: Authors

Kappa is calculated according to the expression (1) as $\kappa = \frac{0.9839 - 0.9355}{1 - 0.9355} = \mathbf{0.75}$. The value of the IRA indicated by kappa was interpreted using the guidelines given in (Landis & Koch, 1977), according to which there are six categories of agreement: poor (−1–0), slight (0.01–0.2), fair (0.21–0.4), moderate (0.41–0.6), substantial (0.61–0.8) and excellent (0.81–1). Two evaluators achieved substantial agreement ($\boldsymbol{\kappa = 0.75}$) and we conclude that evaluators substantially agree with the automated extraction of key information from the full job advertisement texts and that we can proceed with such automation for the full data set.

2.2. Data transformation and cleansing

Data pre-processing is the most important task in any analytical project, consuming most of the time and effort of the analyst. The goal is to improve data quality and transform data into a suitable format for machine learning (Grljević, 2023). Pre-processing of unstructured data, such as ICT job descriptions, is even more demanding than working with structure data. This is due to the extensive vocabulary used in texts where each unique word, phrase, or a symbol represents a feature. This leads to high dimensionality of data in use. Without adequate pre-processing all of these words would be used to represent text, in this case study job advertisements, regardless of the fact they do not contribute equally to the semantics or meaning of the text. An effective data preparation is key to reducing dimensionality and identifying a simplified set of feature to represent texts (Feldman & Sanger, 2013). In the empirical study, authors applied several procedures pertaining to data anonymization, data transformation, or cleaning the noise from data.

The goal of *data anonymization* is to protect the privacy of companies posting job advertisements. To achieve removal of all mentions of companies we opted for identification and removal of named entities using spaCy open-source Python library for natural language processing and its implementation of entity recognizer for organisations (SpaCy, 2024). While entity recogniser identified certain technologies as organisation, before removing identified named entities the authors implemented manually crafted exceptions referring to the predominant technologies, such as *Microsoft Dynamics 2016*, *Microsoft Azure*, *SAP*, etc. As the main analytical goal is set towards modelling ICT skills on the Serbian job market, we also removed all mentions of geopolitical entities using spaCy library. This aided to pertaining only content referring to the job position, skills, or technologies the company requires for the position.

Job postings officially represent the company and one would expect high standards regarding grammar and spelling. However, the authors of this paper found wide disrespect towards grammar and spelling, particularly in job postings written in the Serbian language. This mostly refers to words containing diacritic marks, which are often written in colloquial style, such as instead of using letter *š* employers used *sh*, or instead of using *č* employers used *ch*. This was corrected manually as it posed limitation for translation of texts to English language in the next step.

Job advertisements are posted in various languages. English written job posts comprise 95.39% of dataset (3.475), Serbian 4.53% (165), and German 0.08% (3). Although non-English advertisements represent a relatively small portion of the dataset (approximately 5%), to ensure full representation of market requirements authors opted for their retainment. The authors decided on language unification as the multilingual strategy. Using Google Translate API the language of the job post is detected and translated to English. In this way, the language is unified. Prevalence of English in the dataset and the availability of the NLP resources for English language motivated this decision. Selected approach is in line with multilingual strategies identified by Laureate et al. (2023) and studies exploring user behaviour (Tang et al., 2022; Marcolin et al., 2021; Kirilenko et al.,

2021; Grljević et al., 2025). The authors manually reviewed and compared original and translated texts to ensure that vocabulary related to skills is preserved, mitigating in this way known limitations of machine translation that might impact technical and context-specific terminology.

The content is thereafter pre-processed in Python using regular expressions or NLTK (natural language toolkit) libraries for the following data cleansing and dimensionality reduction procedures:

1. Case standardization. As different forms of the same word, such as capitalized, uppercased, or lowercased, are considered by algorithms a different feature that increase the dimensionality of the data, the authors standardized the case by lowercasing the text of all job postings.
2. Noise removal. In textual data noise refers to all instances that are not contributing to the meaning or adding new knowledge or information about the text. With respect to noise, we have removed: hashtags, URLs, email addresses, square brackets, punctuation, alphanumeric values excepting technology names (e.g., css3, html5, neo4j), and numbers.
3. Excessive whitespaces and newlines are removed from the job descriptions for this purpose.
4. Length-based token filtering. Words with less than 3 characters are removed from the texts, as they can be considered as stopwords.
5. Stopwords removal. Stopwords refer to frequent words in spoken and written language that do not add new knowledge or attribute to the semantics of the text, such as conjunctions (Grljević et al., 2022), and as such are removed from the texts.
6. Tokenization implies text segmentation on its constituent words (tokens), allowing the further pre-processing on a token level.
7. Authors utilized part-of-word (PoS) tagging as a part of preparation for text normalization. All instances of adjectives, adverbs, verbs, and nouns are tagged, and subsequently lemmatized. Lemmatization is the process of word reduction to their base form, lemma. It is selected as an approach for text normalization since it preserves semantic integrity and retains meaning (Manning et al., 2008).

Figure 3 illustrates job posting prior and after textual data pre-processing.

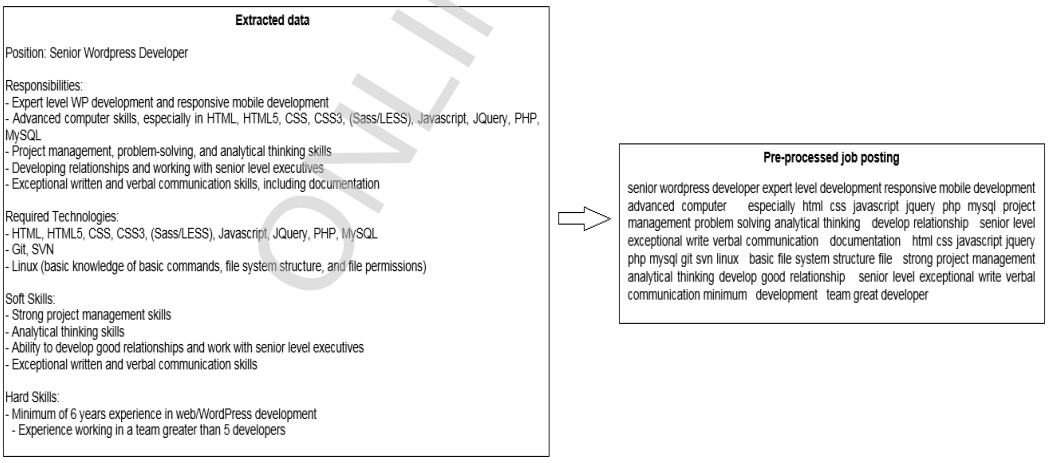


Figure 3 Illustration of pre-processed job advertisements
Source: the authors

2.3. N-gram analysis

An n-gram is a sequence of n tokens or words naturally occurring in text. Single token is referred to as unigram, a sequence of two tokens is called a bigram (e.g., "strong programming", "programming skills"), while a sequence of three tokens is a trigram (e.g., "strong programming skills"), and so forth. To ensure semantical richness of data, we utilize n-grams to identify domain specific phrases that should be observed unified in the further analysis, as well as those that

can be considered as domain-dependent stopwords and should be excluded from the further analysis.

By manually inspecting unigrams, authors were able to identify domain-dependent stopwords, such as *experience, skill, work, knowledge,* or *position*, which were removed from the corpus. Also words occurring 5 or less than five times in the corpus can be considered as uninformative, such as *retirement, water*, *food*. For this reason and with the goal of retaining only words that contribute to modelling soft and hard skills differentiating ICT job advertisements by gender, the authors conducted frequency-based filtering to remove all words occurring 5 or less than 5 times in the corpus. Figure 4 illustrates 20 most frequently used words in ICT job advertisements in the resulting corpus. Prevailing keywords in the corpus are team, development, test, technology, design, software, English, etc. Unigrams offer an initial intuition on common trends in ICT market.

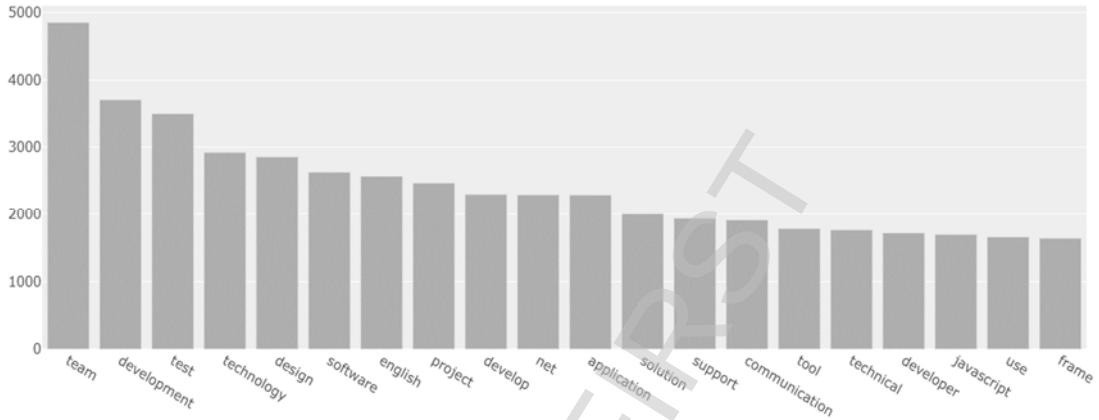


Figure 4 The most frequent unigrams in IT job advertisements
Source: the authors

Bigram analysis allowed for identification of domain-dependent phrases that are not contributing to the context, such as *require technology*, *soft skills*, *hard skills*, which were removed from the corpus, as well as informative phrases that should be retained as features, such as *artificial intelligence, machine learning, SQL server*. Figure 5 illustrates top 20 bigrams according to frequencies, after filtering.

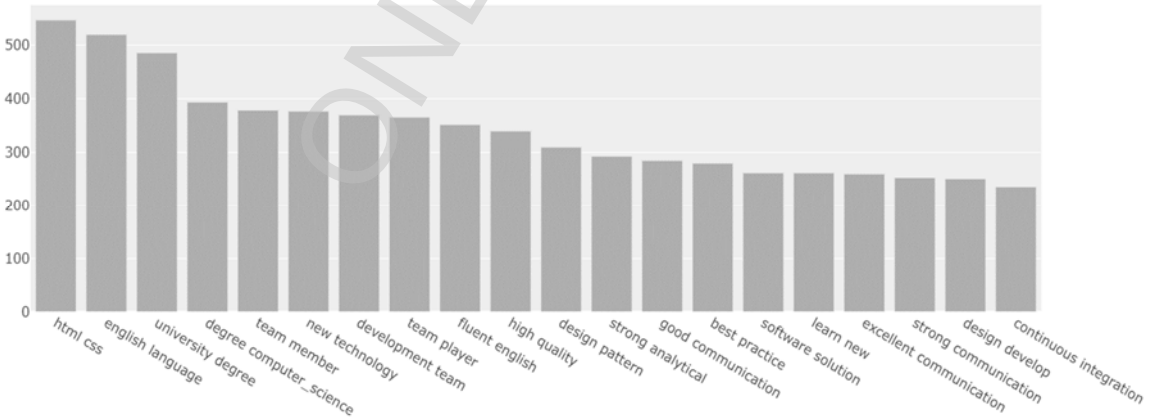


Figure 5 The most frequent bigrams in IT job advertisements
Source: the authors

Bigrams and trigrams, illustrated with Figure 6, also offer the first glimpses into the context of the ICT advertisements. From the most frequent bigrams we can observe that key technologies refer to frontend technologies, HTML and CSS, companies are seeking people with English language skills and university degree, preferably computer science, while most sought soft skills are

ability to work in team, analytical, and communication capabilities. Trigrams are indicating that among other important soft skills employers are valuing learning new technologies, problem solving, and hard skills refer to writing clean, high quality code.

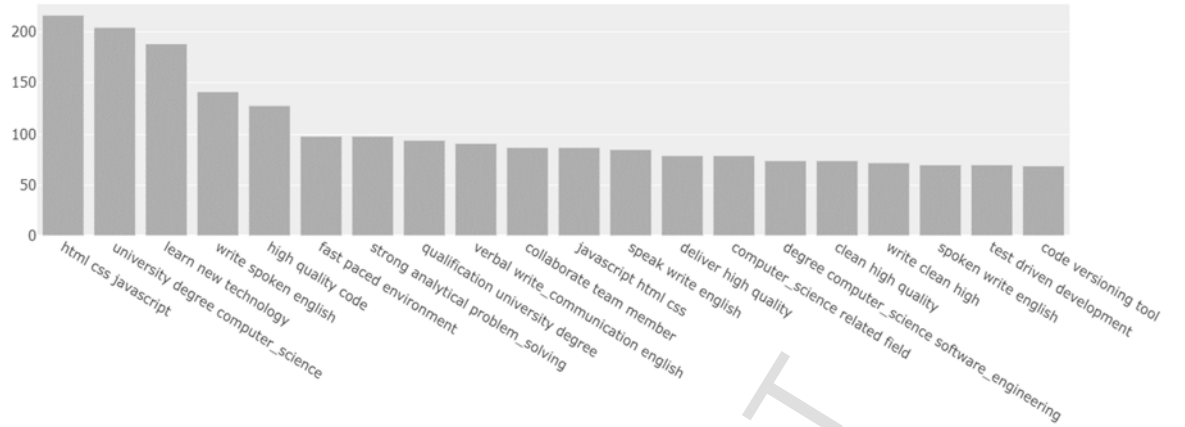


Figure 6 The most frequent trigrams in IT job advertisements
Source: the authors

2.4. Feature engineering

Feature engineering is a crucial step in machine learning, converting pre-processed data into a structured format suitable for automated analysis (Popov, 2022). Three types of features are generated: hard and soft skills sets of features and gender-related feature.

2.4.1. Hard and soft skills features

Unigrams and informative phrases from bigram analysis are used to generate the list of unique words, which were subsequently classified into hard and soft skills through Python code, invoking ChatGPT-3.5 Turbo for AI-driven categorization. Each skill is represented as a separate feature in a column, with values of 0 if the skill is absent and 1 if it is present in the job advertisement. The authors manually inspected resulting skills to filter out instances not strictly related to job skills, such as *three*, *popular*, *accept*, etc., resulting in a 645 keywords reflecting hard skills and 448 keywords reflecting soft skills.

2.4.2. Gender-related features

Data on gender distribution are expressed as percentages of male, female applicants, and applicants whose sex was undisclosed, not as raw counts. Advertisements for which the gender of the candidates is indicated as 100% unknown are removed from the dataset, leaving 3,638 posts for further analysis. Although the percentage data on the gender of candidates provides an overview of gender representation, it limits the granularity of statistical analysis and subgroup comparisons. A discussion of this limitation and its implications is included in the concluding section.

Advertisements are categorized and labelled according to gender ratios. Based on the percentage data, authors derived new attribute *Gender_bins* indicating majority gender of applicants using the following manually crafted rules:

- Female-majority ad – at least 51% of applicants are female.
- Male-majority ad – at least 51% of applicants are male.
- Neutral ad – advertisements where the gender distribution is approximately balanced, with no more than 51% of applicants from either gender.

Table 3 illustrates the gender distribution of ICT job advertisements, indicating severe skewness towards male-majority advertisements. For clarity and brevity, the terms "job advertisements" and "ads" are used interchangeably in the remaining of the paper.

Table 3 ICT job advertisements distribution across gender categories

Male majority ads	Female majority ads	Neutral ads
3465	101	72

Source: Authors

2.5. Identification and interpretation of salient features

This step identifies deviations or variations, i.e., searches for deviations between groups of texts. Using domain knowledge, the authors in the interpretation phase interpret the identified salient

features in the context of the domain from which the texts originate. As suggested in (Allen et al., 2018), for this purpose, we used word clouds, analysis of leading words within the framework of observed phenomenon, as well as co-occurrence networks indicating pairs of skills that often occur together in job advertisements. The results of exploratory analysis are in detail presented in the subsequent sections. Discovered knowledge can help in setting guidelines for potential improvements in hiring strategies.

3. Results

3.1. Word clouds of gender labelled job advertisements

A word cloud is a graphical presentation of the most frequently occurring terms in a corpus. Words closely associated with a specific corpus are emphasized by larger font size. Contrary, less frequent terms appear in smaller font sizes. In this way they extract key information from a corpus and present it visually, enabling prompt comprehension of topics and essential content of the text (Ren et al., 2024). Figures 7-9 illustrate generated word clouds for distinct genre related subsets of IT job advertisements using the WordCloud library in Python. The colours are randomly allocated to distinguish individual words, without conveying any semantic significance.

Prominent words in the word cloud of male majority ads, illustrated in Figure 7, such as *application*, *system*, *development, solution, process, technology,* etc.*,* indicate these ads are more directed towards technical aspects and greater emphasis is on software development, platforms, and engineering-oriented terminology. Female majority ads, illustrated in Figure 8, are characterized by words such as *project*, *team*, *product*, *data*, *design,* etc., indicating the focus is more on teamwork, collaboration, coordination, product-related and analytics-oriented roles. In neutral ads, Figure 9, software testing and user-oriented roles are more in focus compared to other two groups of advertisements.

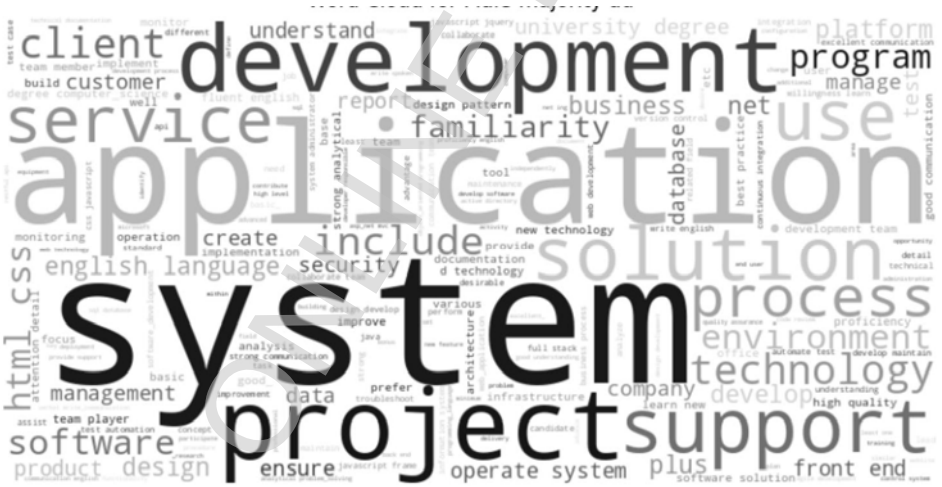


Figure 7 Word cloud of male-majority job advertisements
Source: the authors

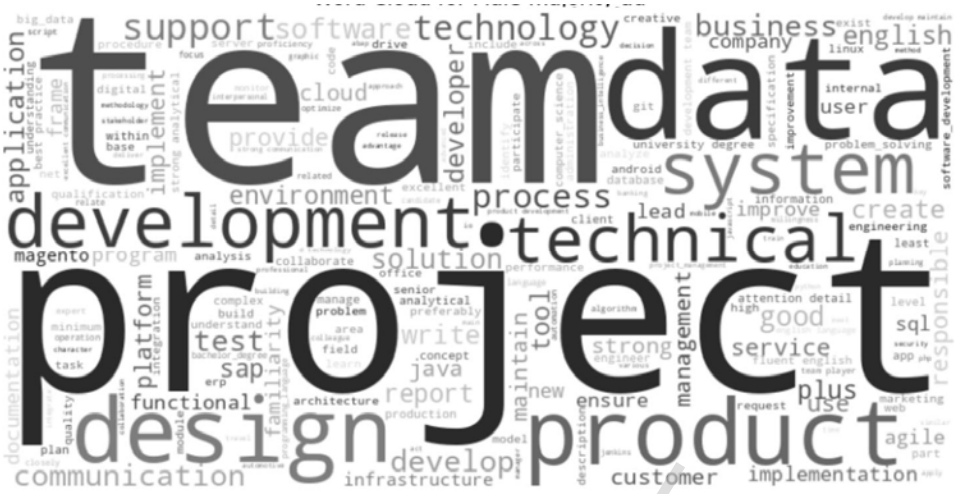


Figure 8 Word cloud of female-majority job advertisements
Source: the authors

Figure 9 Word cloud of neutral job advertisements
Source: the authors

3.2. Distinguished words in gender labelled job advertisements

Keywords that contribute more to advertisement content are derived using the Term Frequency-Inverse Document Frequency (TF-IDF) measure. This is a weighted value that indicates relative importance of a term to a document compared to the relevance it has to all documents in the corpus (Grljević, 2023; Zhou et al., 2024). TF-IDF improves semantic clarity by downweighing commonly occurring words with low distinguishing power while highlighting less frequent terms that carry greater discriminatory significance (Luo et al., 2021). This measure is constituted of two measures Term Frequency (TF) and Inverse Document Frequency (IDF), (Markov & Larose, 2007; Grljević, 2023). TF measures term frequency within a document and is defined as:

$$TF_{t,d} = \begin{cases} 1 + \log\big(count(t,d)\big) & , \ if\ count(t,d) > 0 \\ 0 & , \quad else \end{cases} \quad (2)$$

Here, t denotes the term, and d the document. In the research presented in this paper d refers to a single job ad. The logarithmic adjustment mitigates the overemphasis on frequently appearing terms. IDF, emphasizing terms that occur in fewer documents, is computed as:

$$IDF_t = \log(\frac{N}{df_t}) \quad (3)$$

where N represents the total number of documents in the corpus, and df_t the count of documents containing term t. Terms appearing in fewer documents receive higher IDF values, while universally present terms are assigned the lowest weight (1). Logarithmic scaling smoothens the measure for large corpora.

The significance of a term t in a document d is determined by combining TF and IDF:

$$TF\text{-}IDF_{(t,d)} = TF_{t,d} \; x \; IDF_t \quad (4)$$

Table 4 illustrates 10 top distinguishing words in job advertisements according to gender categories, based on their TF-IDF scores. Emphasized words indicate unique words that are particularly indicative of respective gender categories. Male majority advertisements are particularly characterized by .NET application and software development. Female majority advertisements are focused on data, product, SAP, and technical background, while neutral advertisements are user-centred.

Table 4 Top distinguishing words in job advertisements according to gender categories

Male-majority	Female-majority	Neutral
test	**data**	team
team	team	test
development	**product**	software
net	development	development
design	project	technology
software	**sap**	develop
project	design	design
technology	test	**developer**
application	**technical**	**customer**
develop	**strong**	**solution**

Source: the authors

3.3. Skills across gender labelled job advertisements

The total number of job postings is imbalanced and skewed toward the male-majority category and using raw skill frequencies as a sole measure could mislead conclusions. The authors applied a normalization procedure to raw skill frequencies to facilitate a comparison of skills prevailing in male-majority and female-majority job advertisements and indicate disparities in skills across gender categories.

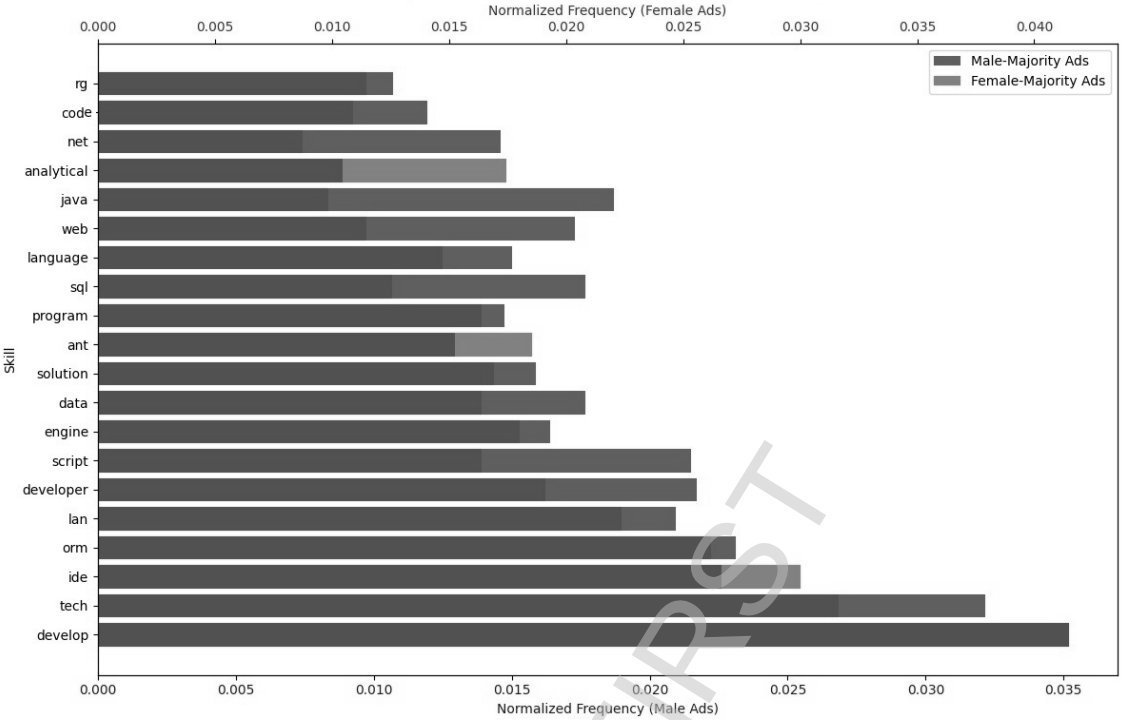
The frequency of each skill within male-majority and female-majority job advertisements was normalized by dividing the raw frequency of each skill by the total frequency of all skills in the respective category. Such a normalization approach transforms absolute skill frequencies into relative proportions, making the relative importance of specific skills within each gender category comparative and not dependent on differences in the total volume of job advertisements. Both hard and soft skills are normalized. The mathematical representation of the normalization process is as follows:

$$Male\_Normalized_i = \frac{Male\_Frequency_i}{\sum_{j=1}^{n} Male\_Frequencies_j} \quad (5)$$

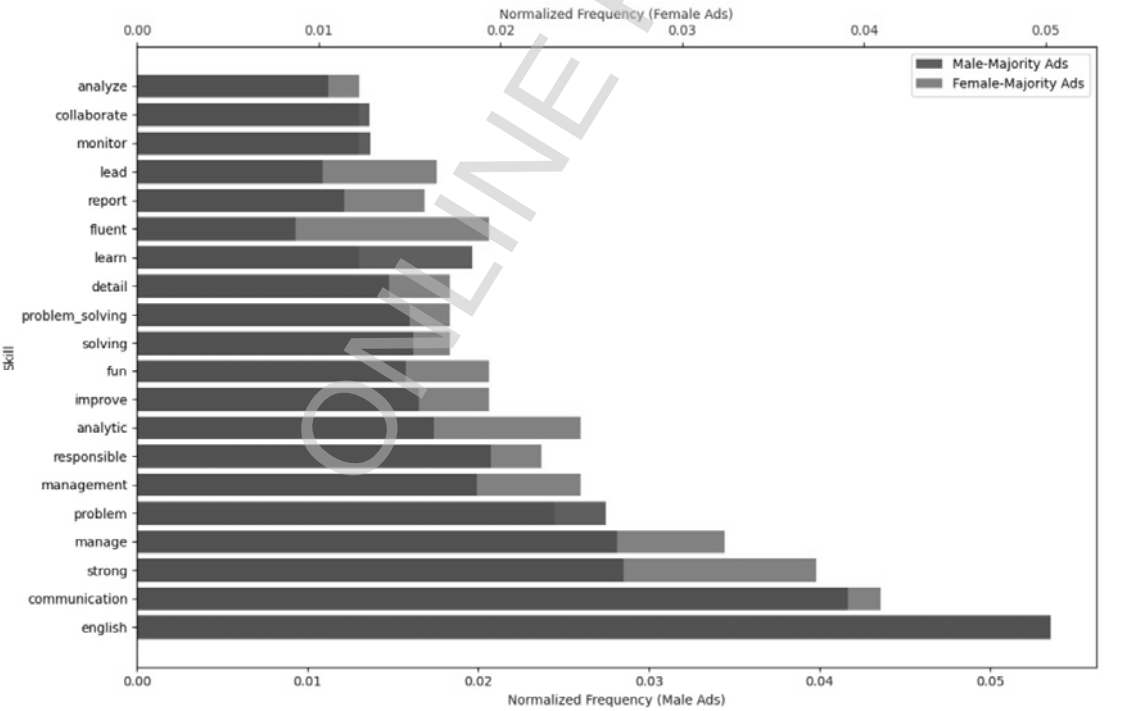
$$Female\_Normalized_i = \frac{Female\_Frequency_i}{\sum_{j=1}^{n} Female\_Frequencies_j} \quad (6)$$

where, $Male\_Normalized_i$ and $Female\_Normalized_i$ represent the proportion of job advertisements within male-majority and female-majority categories, respectively, that mention skill i. $Male\_Frequency_i$ and $Female\_Frequency_i$ denote raw frequencies of skill i appearing in each gender category. $\sum_{j=1}^{n} Male\_Frequencies_j$ and $\sum_{j=1}^{n} Female\_Frequencies_j$ represent the sum of all skill mentions in each of the gender categories, serving as denominators ensuring proportional scaling within this category. The total number of distinct skills in the dataset is denoted as n.

Figure 10 illustrates resulting top 20 hard (10a) and soft skills (10b) across gender categories. Figure 10a indicates higher rates in male-majority postings for hard skills, such as *develop*, *developer*, *tech*, *script*, *java, .net, sql*, that indicate male category is dominated by technological skills. On the contrary, female-majority advertisements put emphasis on *analytical*, *ant*, and IDE, indicating potential differences in how job advertisements for different gender-majority roles are structured. Figure 10b indicates that soft skills are present with higher frequencies in female majority job advertisements, except for skills associated to identification and resolutions of problems, willingness to learn, monitoring, and collaboration, while all other most frequently occurring skills are linked to female-majority advertisements.



a) Hard skills



b) Soft skills

Figure 10 Prevailing skills in male-majority vs. female-majority job advertisements
Source: the authors

3.4. Skill co-occurrence networks

Co-occurrence networks represent text mining technique that analyses joint occurrence of pair of keywords in the documents (van Eck & Waltman, 2014). In the case study keywords are hard or soft skills extracted from job advertisements. The assumption is that keywords which frequently appear together in the same documents have a relationship to one another (Narong & Hallinger, 2023). Co-occurrence network is derived based on the skills frequencies using Pearson correlation coefficients. In this way it is quantified how often skills are mentioned together in job advertisements and facilitated identification of dependencies between skill sets. To extract only strong correlations among skill sets, a threshold is set to 0.75 and identified skill pairs are visualized using Python's implementation of network graph. Nodes represent individual skills, while edges represent co-occurrence relationships indicating which skills are often required together. The following of the section is structured into subsections to separately present hard and soft skills co-occurrence networks. The corresponding network visualizations are provided as supplementary material to the paper (Grljević & Kecojević, 2025).

3.4.1. Hard skills co-occurrence network

Three types of co-occurrence networks are created. The first refers to hard skill co-occurrence network regardless of genders providing an insights about general patterns present in job descriptions (Supplementary material – (Grljević & Kecojević, 2025) – Figure 1). The results suggest presence of five distinct clusters of skills. Authors have associated names according to technical domains clusters represent:

- *Programming and development* is represented with skills, such as Java, JavaScript, jQuery, Python, and MongoDB, which suggest strong co-occurrence in software development and data-related roles.
- *Networking* cluster groups skills, such as ASP.NET, WebSphere, GlassFish. These skills are associated with infrastructure and networking roles.
- *Database and cloud* cluster comprises skills such as PostgreSQL, Unix, and diagnostics that might indicate relevance in cloud computing, system administration, and database management.
- *Software testing and automation* cluster. The co-occurring skills, such as TestLink, Testrail, and Robot Framework, indicate group of ICT roles related to software testing and automated tests.
- *Embedded systems and IoT* cluster links skills such as Autosar, WinCC, and CPU.

We also observe separately hard skill co-occurrence patterns in job advertisements that are male majority (Supplementary material – (Grljević & Kecojević, 2025) – Figure 2) and female majority (Supplementary material – (Grljević & Kecojević, 2025) – Figure 3). Male majority hard skill co-occurrence network is sparser compared to female majority network. This is reflected in fewer connections with many isolated clusters compared to denser co-occurrence network in female majority advertisements indicating multiple hard skills appear in more co-occurring pairs. Advertisements male candidates are more interested in put emphasis on skills with technical and infrastructure focus or software engineering and testing focus. Technical and infrastructure focused skills are reflected in the presence of skills related to programming languages and database (e.g., Java, Python, PostgreSQL, MongoDB), networking and security related skills (e.g., IPsec, HTTPS, TCP, proxy), DevOps and automation skills (e.g., CI/CD, WebLogic, Robot Framework, Docker), or embedded and hardware oriented skills (e.g., microcontrollers, controller, GPS). Software engineering and testing focus is reflected in mentions of related tools, such as TestRail, TestLink, PHPUnit, Cordova. Such co-occurrence network indicates that ICT job postings drawing more attention among male candidates are more specialised in various domains, such as security, embedded systems, DevOps, testing, which formed separate clusters in the co-occurrence network.

High connectivity, observed in hard skill co-occurrence network of female majority job advertisements indicates that certain skills frequently co-occur in job advertisements associated with higher female applicant engagement. This pattern may reflect a broader or versatile skill set expected in these roles. Job advertisements linked to female-majority application profiles appear to span a wide range of subdisciplines, including software development (e.g., JavaScript, AngularJS,

Node.js, GraphQL, frontend development with Bootstrap or React), data and cloud infrastructure (e.g., Azure, Kubernetes, RabbitMQ, Ansible, virtual machines, PostgreSQL, Hive), project management (e.g., product owner, scrum, Bitbucket, Subversion), business analytics (e.g., business analysts, visualisations, regression), as well as security and compliance (e.g., authentication, authorisation, secure configurations).

3.4.2. Soft skills co-occurrence network

The resulting co-occurrence network was extremely sparse when threshold is set at 0.75. This indicates that small number of skills in ICT job advertisements frequently appear together with correlation above 0.75, meaning they tend to be mentioned independently instead in strong pairs. Such sparsity limits the interpretability and usefulness of the network. To obtain a more informative graph with meaningful soft skill clusters, i.e., groups of skills that might be relevant for both job seekers and HR managers, and to achieve improved interpretability of the graph we varied the threshold. A threshold of 0.5 was selected as it provided a balance between including meaningful co-occurrence patterns and maintaining a readable network structure. Using this threshold we observed soft skills co-occurrence network for the overall dataset of job advertisements to gain insights into the general patterns present in job descriptions (Supplementary material – (Grljević & Kecojević, 2025) – Figure 4). The results suggest presence of four distinct groups of soft skills. Authors have attributed names according to prevailing keywords used in job descriptions:

- *Communication and collaboration* are represented with keywords, such as *communicativeness, communicate, discuss, open-minded, trustworthy, teammate,* etc. It represents an important personal threads for ICT roles where effective communication, teamwork, and interpersonal skills often have a crucial role for successful completion of projects, or collaborations across teams or departments.
- *Leadership and organisational skills* are represented with keywords, such as *management*, *training*, *organise, motivate, guide*, etc. They emphasise importance of such skill set in managing projects, teams, and ensuring undisturbed workflow.
- *Problem-solving and analytical thinking* are referred to with keywords such as *problem solving, suggest, anticipate, analytical thinking, estimating*, etc. This set of skills is closely related to critical thinking that is necessary for resolving complex problems and working with data.
- *Adaptability and performance under pressure* are represented with keywords like *stressful, stress, rapid, immediate,* etc. The co-occurring skills indicate that ICT needs professionals who can handle high pressure and adapt to changes.

We also observe separately soft skill co-occurrence patterns in job advertisements that are male majority (Supplementary material – (Grljević & Kecojević, 2025) – Figure 5) and female majority (Supplementary material – (Grljević & Kecojević, 2025) – Figure 6). Similar to hard skills co-occurrence networks, male majority soft skill co-occurrence network is sparser than female, which is reflected in fewer connections. Also we can observe less overlaps in co-occurrence network suggesting well-defined sets of expected competences. The results indicate that job advertisements male candidates are more interested in put emphasis on soft skills linked with autonomy and independence (e.g., *independent*, *autonomous*, *problem solving*, *trustworthiness*), analytical thinking (e.g., *statistics*, *methodical*, *intelligent*, *analytical thinking*), leadership (e.g., *manage*, *teammate*, *guide*, *conscientious*), efficiency and performance (e.g., *consistency*, *monitoring*, *improvement*, *rapid*). Such co-occurrence network indicates that ICT job postings drawing more attention among male candidates are emphasizing problem-solving, leadership, independent decision making which align with technical ICT roles and managing technical teams.

The soft skill co-occurrence network in female majority advertisements reveals a pattern similar to that of hard skills, with a high degree of interconnectivity suggesting a diverse and multi-dimensional skill set. Job advertisements associated with higher female applicant engagement more frequently emphasise competencies such as communication and teamwork (e.g., *communicativeness*, *coach*, *growth*, *community*), management (e.g., *manage*, *milestone*, *willingness*), emotional intelligence (e.g., *zeal*, *care*, *friendly*), critical thinking (e.g., *problem-solving*, *investigate*, *insight*), and adaptability (e.g., *adjustable*, *compatible*).

3. Discussion

Word clouds and distinguishing words across gender categories indicate potential gendered language patterns in job descriptions and skills and responsibilities that may attract different applicant pools, such as the use of engineering-oriented terminology in male dominated job advertisements contrary to collaboration-oriented and product-related terminology in female dominated advertisements. Natural language pre-processing techniques enabled extraction of hard and soft skills from job descriptions. On the basis of these skills, labour market demands were analysed and skills that differentiate applicants' interests by gender are identified. Frequency analysis revealed that soft skills are present with higher frequencies in female majority and hard skills in male majority job advertisements, while co-occurrence networks offered insights into common skill groupings indicating how different hard and soft skills are used in job descriptions. Table 5 comparatively presents key hard skills related findings. Job advertisements male candidates prefer favour specialised skills and are more focused on security, networking, and DevOps. The female majority advertisements are more focused on business tools and agile methodologies. High connectivity in the co-occurrence network suggests that job advertisements associated with higher female applicant engagement tend to describe multidisciplinary roles requiring a broader set of skills.

Table 5 Comparative insights from hard skills analysis

Aspect	Male-majority ads	Female-majority ads
Skill specialisation	Specialised clusters	General clusters
Key fields	Security Embedded systems Databases DevOps	Frontend development Agile Business analysis
Technical focus	Infrastructure-heavy skills System-level skills	Application oriented Business focus

Source: the authors

Table 6 comparatively presents key findings related to soft skills. Job advertisements male candidates prefer are more focused on structured leadership, independent decision making, autonomy in work, and efficiency. The female majority advertisements are more focused on collaborative problem solving, teamwork, high adaptability, mentorship, and emotional intelligence. These differences in prevalent soft skills suggest that job advertisements associated with male-majority applications tend to describe technical or managerial roles, while those linked to female-majority applications more often emphasise customer-centric or HR-related responsibilities.

Table 6 Comparative insights from soft skills analysis

Aspect	Male-majority ads	Female-majority ads
Key skills	Autonomy Independence Structure	Communication Teamwork Emotional intelligence Adaptability
Focus	Efficiency driven Goal-oriented Hierarchical	People-oriented Mentorship driven Collaborative

Source: the authors

Conclusion

Proposed methodology, based on NLP techniques and exploratory text analysis, offers a promising direction in job advertisement analysis. NLP enables a scalable and unbiased way to uncover patterns in job advertisements' language that might otherwise go unnoticed. By analysing linguistic patterns and skill requirements in 3,643 job advertisements in Serbia, the authors found that role descriptions and the gender composition of applicants correlate. Job advertisements more appealing to male candidates are more specialised, putting an emphasis on technical and infrastructure-heavy skills, such as microcontrollers, GPS, robotics. This reflects a higher demand for engineering-oriented roles. According to the results, female candidates are more prone to business-oriented positions, which highlight team collaboration, teamwork, emotional intelligence, and data-related tasks. These roles suggest integration of coordination skills with technical knowledge and more prominent demand for agile methodologies. Differences suggest that even a subtle shifts in framing can influence the future candidate's perception of their fitness for a role.

Resulting insights contribute to broader efforts aimed at creating more equitable and diverse digital labour markets and have a practical implications for human resource managements and recruitment strategies in the ICT sector. They can support diversity-aware recruitment practices, as results align with a growing body of research

recommending inclusive recruitment practices, such as the use of gender-neutral language in job advertisements or software able to detect bias. A more inclusive recruitment strategies could be supported by understanding the influence language used in job descriptions has on application behaviour. Such knowledge enables employees to refine the job descriptions and ensure a more inclusive appeal, particularly for roles where female representation is historically low, thus, contributing to reducing the gender imbalance in ICT. Beyond recruitment, discovered patterns in formulation of job descriptions may guide career planning and professional development efforts, by informing the design of training programmes, mentorship pathways, and curriculum development in ICT education. This could encourage a more balanced development of technical and interpersonal competences across all genders and contribute to a more equitable and resilient ICT workforce.

Although proposed methodology offers valuable insights into gender-related patterns in job advertisements, limitations should be acknowledged. Given the ICT labour market is highly dynamic and technologies are evolving changing the skills in demand, dataset should be regularly updated to ensure the relevance of the findings over time and enable tracking of evolvement of gendered patterns over time. Another limitation stems from the way gender data is obtained, as percentages instead of raw frequencies. This restricts the granularity of the analysis and restricts opportunities for more extensive analyses. Future data collection should aim to include actual counts to enhance precision. Lastly, diversity observed in female-majority category in both soft and hard skills may be due to the limited data availability. This reinforces the importance of strengthening the dataset, ensuring the data representation is of more quality and balancing the representation of each category. This will improve analytical robustness.

As part of our future research, we aim to explore more advanced NLP techniques, including embedding-based and transformer models, alongside methods, such as market-basket analysis and community detection for identification of clusters of skills and enhance the depth and breadth of our analysis. Future research could also include more dimensions, such as educational background of candidates, socio-economic status, location. This will enable improved understanding of how multiple factors shape access to ICT employment opportunities.

Declarations

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Funding

Not applicable.

Acknowledgements

We are thankful to the Inspira group for providing the data and Professor Zita Bošnjak for participating as an expert in evaluation of data extraction procedure.

References

Allen, T. T., Sui, Z., & Akbari, K. (2018). Exploratory text data analysis for quality hypothesis generation. *Quality Engineering, 30*(4), 701–712. https://doi.org/10.1080/08982112.2018.1481216

Ashcraft, C., McLain, B., & Eger, E. (2016). *Women in tech: the facts.* National center for women and information technology. Available at: https://www.ceoplaybook.co/wp-content/uploads/2019/11/womenintech_facts_fullreport_05132016-1.pdf

Back, A., Hajikhani, A., & Suominen, A. (2021). Text Mining on Job Advertisement Data: Systematic Process for Detecting Artificial Intelligence Related Jobs. *CEUR Workshop Proceedings*, (pp. 111-124). Available at: http://ceur-ws.org/Vol-2871/paper9.pdf

Bradić-Martinović, A., & Banović, J. (2018). Assessment of Digital Skills in Serbia with Focus on Gender Gap. *Journal of Women's Entrepreneurship and Education, 2018*(1-2), 54-67. https://doi.org/10.28934/jwee18.12.pp54-67

Bradić-Martinović, A., Lazić, M., & Banović, J. (2024). Exploring Gender Disparities in Digital Skills: Evidence from the Serbian Tourism Sector. *Journal of Women's Entrepreneurship and Education, 3-4*, 180-205. https://doi.org/10.28934/jwee24.34.pp180-205

Brussevich, M., Dabla-Norris, E., Kamunge, C., Pooja, K., Khalid, S., & Kochhar, K. (2018). *Gender, Technology, and the Future of Work.* IMF staff discussion notes. Available at: https://www.imf.org/-/media/Files/Publications/SDN/2018/SDN1807.ashx

Cosgrove, J., Sostero, M., & Bertoni, E. (2024). *Mapping DigComp digital competences to the ESCO skills framework for analysis of digital skills in EU online job advertisements.* European Commission, Joint Research Centre. Luxembourg: Publications Office of the European Union. Available at: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC137060/JRC137060_01.pdf

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989(1), Article 8. Available at: http://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

European Commission. (2020, March 5). A Union of Equality: Gender Equality Strategy 2020-2025. Brussels. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0152

European Commission. (2022). *The Digital Economy and Society Index (DESI).* European Union. Available at: https://digital-strategy.ec.europa.eu/en/policies/desi

Farzindar, A., & Inkpen, D. (2015). *Natural Language Processing for Social Media.* Morgan & Claypool.

Feldman, R., & Sanger, J. (2013). *The text mining handbook: Advanced approaches in analyzing unstructured data.* Cambridge University Press.

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. *Journal of Personality and Social Psychology, 101*(1), 109 –128. https://doi.org/10.1037/a0022530

Grljević, O. (2023). *Analiza sadržaja društvenih medija: Napredni pristupi analizi nestrukturisanih podataka.* Subotica: Ekonomski fakultet u Subotici.

Grljević, O., Bošnjak, Z., & Bošnjak, S. (2019). Sustainable Development through Gender Equality – A Case of Higher Education of Data Scientists. *Journal of Women's Entrepreneurship and Education, 3-4*, 73-85. https://doi.org/10.28934/jwee19.34.pp73-85

Grljević, O., Bošnjak, Z., & Kovačević, A. (2022). Opinion mining in higher education: a corpus-based approach. *Enterprise Information Systems, 16*(5), 1773542. https://doi.org/10.1080/17517575.2020.1773542

Grljevic, O., & Kecojević, T. (2025). An NLP approach to skill analysis in ICT job advertisements from a gender perspective: Supplementary material. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.29468132.v2

Grljević, O., Marić, M., & Božić, R. (2025). Exploring Mobile Application User Experience Through Topic Modeling. *Sustainability, 17*, 1109. https://doi.org/10.3390/su17031109

Hossain, M., Akter, S., Nishu, A. N., Khan, L., Shuha, T. T., Jahan, N., Rahman, M. M., Khatun, M.T. (2023). The gender divide in digital competence: a cross-sectional study on university students in southwestern Bangladesh. *Frontiers in Education, 8*, 1258447. https://doi.org/10.3389/feduc.2023.1258447

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qual Health Res., 15*(9), 1277-88. https://doi.org/10.1177/1049732305276687

Hu, S., Al-Ani, J. A., Hughes, K. D., Denier, N., Konnikov, A., Ding, L., Xie, J., Hu, Y., Tarafdar, M., Jiang, B., Kong, L., Dai, H. (2022). Balancing Gender Bias in Job Advertisements With Text-Level Bias Mitigation. *Frontiers in Big Data, 5*, 805713. https://doi.org/10.3389/fdata.2022.805713

Ilich, L., & Akilina, O. (2017). Impact of ICT on Labor Market Development: Main Trends and Prospectives. *Електронне наукове фахове видання "Відкрите Освітнє Е-Середовище Сучасного Університету", 3*, 55–68. https://doi.org/10.28925/2414-0325.2017.3.5568

Jaiswal, K., Kuzminykh, I., & Modgil, S. (2025). Understanding the skills gap between higher education and industry in the UK in artificial intelligence sector. *Industry and Higher Education, 39*(2), 234–246. https://doi.org/10.1177/09504222241280441

Jevtić, B., Vucekovic, M., & Tasić, S. (2023). The Effects of Digitalization and Skills on Women's Labor Market Inclusion- Serbian Gap Study. *Journal of Women's Entrepreneurship and Education, Special Issue: "Strengthening opportunities and solutions for women entrepreneurs in Asia and Europe"*, 58-75. https://doi.org/10.28934/jwee23.pp58-75

Jiang, H., Xia, S., Yang, Y., Xu, J., Hua, Q., Mei, Z., Hou, Y., Wei, M., Lai, L., Li, N., Dang, Y., Zhou, J. (2024). Transforming free-text radiology reports into structured reports using ChatGPT: A study on thyroid ultrasonography. *European Journal of Radiology, 175*, 111458. https://doi.org/10.1016/j.ejrad.2024.111458

Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language models.* Stanford online. Available at: https://web.stanford.edu/~jurafsky/slp3

Kirilenko, A. P., Stepchenkova, S. O., & Dai, X. (2021). Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply? *Tourism Management, 83*, 104241. https://doi.org/10.1016/j.tourman.2020.104241

Narong, D.K., & Hallinger, P. (2023). A Keyword Co-Occurrence Analysis of Research on Service Learning: Conceptual Foci and Emerging Research Trends. *Education sciences, 13*(4), 339. https://doi.org/10.3390/educsci13040339

Korbel, P. (2018). *Internet job postings: preliminary.* Adelaide SA 5000, Australia: National Centre for Vocational Education Research. Available at: https://www.ncver.edu.au/__data/assets/pdf_file/0023/2931440/Internet-job-postings-preliminary-skills-analysis-technical-paper.pdf

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology.* Beverly Hills, USA: Sage Publications.

Kukić Đorđević, Z., & Čolić Mihajlović, V. (2023). *Women in Serbia's ICT sector.* Belgrade, Serbia: UNDP and UNDP Accelerator Lab in Serbia. Available at: https://lab.undp.org.rs/wp-content/uploads/2024/03/Research-eng-web_26_02_2024.pdf

Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Laureate, C. D.P., Buntine, W., & Linger, H. (2023). A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review, 56*, 14223–14255. https://doi.org/10.1007/s10462-023-10471-x

Lazarević-Moravčević, M., Mosurović Ružičić, M., & Minović, J. (2023). Gender Inequality in Education and Science: The Case of Serbia. *Journal of Women's Entrepreneurship and Education, 3-4*, 143-166. https://doi.org/10.28934/jwee23.34.pp143-166

Lovaglio, P. G., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 11*(2), 78-91. https://doi.org/10.1002/sam.11372

Luo, Y., He, J., Mou, Y., Wang, J., & Liu, T. (2021). Exploring China's 5A global geoparks through online tourism reviews: A mining model based on machine learning approach. *Tourism Management Perspectives, 37*, 100769. https://doi.org/10.1016/j.tmp.2020.100769

Manning, S. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Marcolin, C. B., Becker, J. L., Wild, F., Behr, A., & Schiavi, G. (2021). Listening to the voice of the guest: A framework to improve decision-making processes with text data. *International Journal of Hospitality Management, 94*, 102853. https://doi.org/10.1016/j.ijhm.2020.102853

Markov, Z., & Larose, T. (2007). *Data Mining the Web - Uncovering Patterns in Web Content, Structure, and Usage*. Wiley Series on Methods and Applications in Data Mining.

Martínez-Cantos, J. L. (2017). Digital skills gaps: A pending subject for gender digital inclusion in the European Union. *European Journal of Communication, 32*(5), 419–438. https://doi.org/10.1177/0267323117718464

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med (Zagreb), 22*(3), 276-82. Available at: https://pubmed.ncbi.nlm.nih.gov/23092060/

Nasir, S. A. Md, Yaacob, W. F. W., & Aziz, W. A. H. W. (2020). Analysing Online Vacancy and Skills Demand using Text Mining. *Journal of Physics: Conference Series, 1496*(2020), 012011. https://doi.org/10.1088/1742-6596/1496/1/012011

Noble, S.U. (2018). *Algorithms of oppression: How search engines reinforce racism.* New York: NYU Press.

Pažur Aničić, K., & Arbanas, K. (2015). Right Competencies for the right ICT Jobs – case study of the Croatian. *TEM Journal, 4*(3), 236-243. https://doi.org/10.18421/TEM43-03

Pejic-Bach, M., Bertoncel, T., Meško, M., & Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. *International Journal of Information Management, 50*, 416-431. https://doi.org/10.1016/j.ijinfomgt.2019.07.014

Popov, A. (2022). Feature engineering methods. *Advanced Methods in Biomedical Signal Processing and Analysis*, 1-29. https://doi.org/10.1016/B978-0-323-85955-4.00004-1

Purao, S., & Suen, H. (2010). Designing a multi-faceted metric to evaluate soft skills. *SIGMIS-CPR '10: Proceedings of the 2010 Special Interest Group on Management Information System's 48th annual conference on Computer personnel research on Computer personnel research* (str. 88-91). New York, United States: Association for Computing Machinery. https://doi.org/10.1145/1796900.17969

Quirós, C., Guerra Morales, E., Rivera Pastor, R., Fraile Carmona, A., Sáinz Ibáñez, M., & Madinaveitia Herrera, U. (2018). *Women in the Digital Age.* Available at: https://openaccess.uoc.edu/bitstream/10609/149134/1/Women_Tari%CC%81n_EC.pdf

Rau, G., & Shih, Y.S. (2021). Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of English for Academic Purposes, 53*, 101026. https://doi.org/10.1016/j.jeap.2021.101026

Ren, H., Liu, Y., Naren, G., & Lu, J. (2024). The impact of multidirectional text typography on text readability in word clouds. *Displays*, *83*, 102724. https://doi.org/10.1016/j.displa.2024.102724

Scott, A., & Kapor Klein, F. (2017). *Tech leavers study: A first-of-its-kind analysis of why people voluntarily left jobs in tech.* Kapor centre for social impact. Available at: https://www.issuelab.org/resource/tech-leavers-study-a-first-of-its-kind-analysis-of-why-people-volunarily-left-jobs-in-tech.html

Simon, V., Rabin, N., & Gal, H.C.-B. (2023). Utilizing data driven methods to identify gender bias in LinkedIn profiles. *Information Processing & Management, 60*(5), 103423. https://doi.org/10.1016/j.ipm.2023.103423

SpaCy. (2024, May 20). *Entity recognizer*. Available at: https://spacy.io/api/entityrecognizer

Strugar Jelača, M., Slavić, A., Juhász, T., & Gáspár, T. (2025). Students' perception of the importance of soft skills in the business context of Hungary and Serbia. *STRATEGIC MANAGEMENT*, online first. https://doi.org/10.5937/StraMan2400021S

Tan, K. S., Yeh, Y., Adusumilli, P. S., & Travis, W. D. (2024). Quantifying Interrater Agreement and Reliability Between Thoracic Pathologists: Paradoxical Behavior of Cohen's Kappa in the Presence of a High Prevalence of the Histopathologic Feature in Lung Cancer. *JTO Clinical and Research Reports*, *5*(1), 100618. https://doi.org/10.1016/j.jtocrr.2023.100618

Tang, F., Yang, J., Wang, Y., & Ge, Q. (2022). Analysis of the Image of Global Glacier Tourism Destinations from the Perspective of Tourists. *Land, 11*(10), 1853. https://doi.org/10.3390/land11101853

Taplett, F., Krentz, M., Tsusaka, M., & Ziegler, B. (2018). *Winning the Race for Women in Digital.* Available at: https://www.bcg.com/publications/2018/winning-race-women-digital.aspx

Tukey, J. W. (1977). *Exploratory data analysis.* London, UK: Pearson.

Valavosiki, V.-A., Stiakakis, E., & Chatzigeorgiou, A. (2019). Development of a Framework for the Assessment of Soft Skills in the ICT Sector. U A. Sifaleras, & K. Petridis, *Operational Research in the Digital Era – ICT Challenges* (str. -). Cham: Springer. https://doi.org/10.1007/978-3-319-95666-4_8

van Eck, N.J., & Waltman, L. (2014). Visualizing Bibliometric Networks. In: Ding, Y., Rousseau, R., Wolfram, D. (eds) *Measuring Scholarly Impact: Methods and Practice* (pp. 285-320). Cham: Springer. https://doi.org/10.1007/978-3-319-10377-8_13

Vieira da Cunha, M. (2009). The information professional's profile: An analysis of Brazilian job vacancies on the internet. *Information Research, 14*(3), 5–16. Available at: https://files.eric.ed.gov/fulltext/EJ869361.pdf

Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *DigComp 2.2: The Digital Competence Framework for Citizens - With new examples of knowledge, skills and attitudes.* Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/490274

World Economic Forum. (2020). *GLOBAL GENDER GAP REPORT.* Cologny/Geneva: World Economic Forum. Available at: https://www3.weforum.org/docs/WEF_GGGR_2020.pdf

Yunlu, L. (2023, August 7). A Study on Predicting the Attractiveness of Job Advertisements Based on the Text Mining of Job Descriptions. https://doi.org/10.17605/OSF.IO/6HV3K

Zhang, A. (2012). Peer Assessment of Soft Skills and Hard Skills. *Journal of Information Technology Education: Research, 11*, 155-168. Available at: http://www.jite.org/documents/Vol11/JITEv11p155-168Zhang1119.pdf

Zhou, J., Ye, Z., Zhang, S., Geng, Z., Han, N., & Yang, T. (2024). Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data. *Heliyon*, *10*(16), e35945. https://doi.org/10.1016/j.heliyon.2024.e35945

✉ **Correspondence**

Olivera Grljević

University of Novi Sad, Faculty of Economics in Subotica
Segedinski put 9-11, 24000, Subotica, Serbia

E-mail: olivera.grljevic@ef.uns.ac.rs