Annals of Data Science
https://doi.org/10.1007/s40745-024-00524-5

Applying BERT-Based NLP for Automated Resume Screening and Candidate Ranking

Asmita Deshmukh[1] · **Anjali Raut**[1]

Received: 17 September 2023 / Revised: 21 February 2024 / Accepted: 28 February 2024

Abstract

In this research, we introduce an innovative automated resume screening approach that leverages advanced Natural Language Processing (NLP) technology, specifically the Bidirectional Encoder Representations from Transformers (BERT) language model by Google. Our methodology involved collecting 200 resumes from participants with their consent and obtaining ten job descriptions from glassdoor.com for testing. We extracted keywords from the resumes, identified skill sets, and ranked them to focus on crucial attributes. After removing stop words and punctuation, we selected top keywords for analysis. To ensure data precision, we employed stemming and lemmatization to correct tense and meaning. Using the preinstalled BERT model and tokenizer, we generated feature vectors for job descriptions and resume keywords. Our key findings include the calculation of the highest similarity index for each resume, which enabled us to shortlist the most relevant candidates. Notably, the similarity index could reach up to 0.3, and the resume screening speed could reach 1 resume per second. The application of BERT-based NLP techniques significantly improved screening efficiency and accuracy, streamlining talent acquisition and providing valuable insights to HR personnel for informed decision-making. This study underscores the transformative potential of BERT in revolutionizing recruitment through scalable and powerful automated resume screening, demonstrating its efficacy in enhancing the precision and speed of candidate selection.

Keywords BERT · NLP · Resume · Ranking · Screening

✉ Asmita Deshmukh
asmitadeshmukh7@gmail.com

[1] Hanuman Vyayam Prasarak Mandal College of Engineering and Technology, Maharashtra 444605, India

Published online: 08 March 2024

1 Introduction

The process of talent acquisition has undergone a remarkable transformation in recent years, driven by the rapid advancement of Natural Language Processing (NLP) technologies Vajjala et al. [11]. In this era of dynamic technological evolution and intensifying competition for employment opportunities, the significance of efficient and accurate resume screening cannot be overstated Reynolds and Weiner [8]. Traditional manual screening methods often lead to inefficiencies, overlooking potential candidates, and introducing human biases into the selection process Carroll et al. [4]. In response to these challenges, this research embarks on a pioneering journey, proposing an innovative approach to automated resume screening through the utilization of state-of-the-art NLP technology, particularly the bidirectional encoder representation from transformers (BERT) language model developed by Google.

The fundamental premise of this study rests upon the power of BERT to decipher and comprehend intricate nuances within human language Wu et al. [12]. We present a novel strategy that harnesses the capabilities of BERT to automate and optimize the process of resume screening. This approach not only expedites the evaluation of candidate qualifications but also enhances the precision of candidate matching, facilitating a more streamlined and informed talent acquisition process.

To evaluate the efficacy of our automated screening approach, we undertook a com- prehensive data collection and preprocessing endeavor. A diverse pool of 200 resumes was meticulously gathered from willing participants, adhering to ethical guidelines and informed consent. Additionally, ten job descriptions were sourced from glassdoor.com to serve as benchmarking entities for testing our methodology. The extracted keywords and skill sets were systematically organized into structured tables, stored as Excel files. Through a meticulous process of ranking and prioritization, top keywords were isolated for subsequent analysis, with the elimination of stop words and punctuation ensuring data clarity and accuracy.

To enhance data fidelity, stemming and lemmatization techniques were judiciously applied, rectifying tense and semantic inconsistencies. The preinstalled BERT model, renowned for its contextual understanding of language, was then integrated into our methodology. Both the job descriptions and the top keywords extracted from resumes underwent processing through the BERT model, generating informative feature vectors.

A pivotal outcome of our study is the introduction of a novel similarity index, calculated based on the feature vectors. This index empowers the shortlisting of the most relevant candidates, thus expediting the decision-making process. Our experiments demonstrated a marked enhancement in screening efficiency and accuracy, showcasing the transformative potential of BERT-based NLP techniques in revolutionizing the recruitment landscape.

In summary, this research presents an ingenious paradigm shift in resume screening practices, unveiling the transformative potential of BERT-based NLP. The subsequent sections delve into the intricate technical details, experimental setups, and nuanced implications of our automated screening approach, underpinning its role as a powerful and scalable solution for modern talent acquisition challenges.

Figure 1 illustrates the proposed block diagram of the automated resume screening procedure employing BERT and cosine transforms. Initially, a collection of 200 resumes is compiled from prominent resume websites like LinkedIn and gathered locally through direct candidate contact. These resumes are subsequently converted to PDF format and further transformed into Excel files. Upon the completion of Excel file preparation, word preprocessing tasks, such as lemmatization and stemming, are executed. Additionally, the process involves the removal of redundant and stop words from the resumes. All processed words are then input into the BERT model to generate encodings. These encodings are subsequently compared with job descriptions

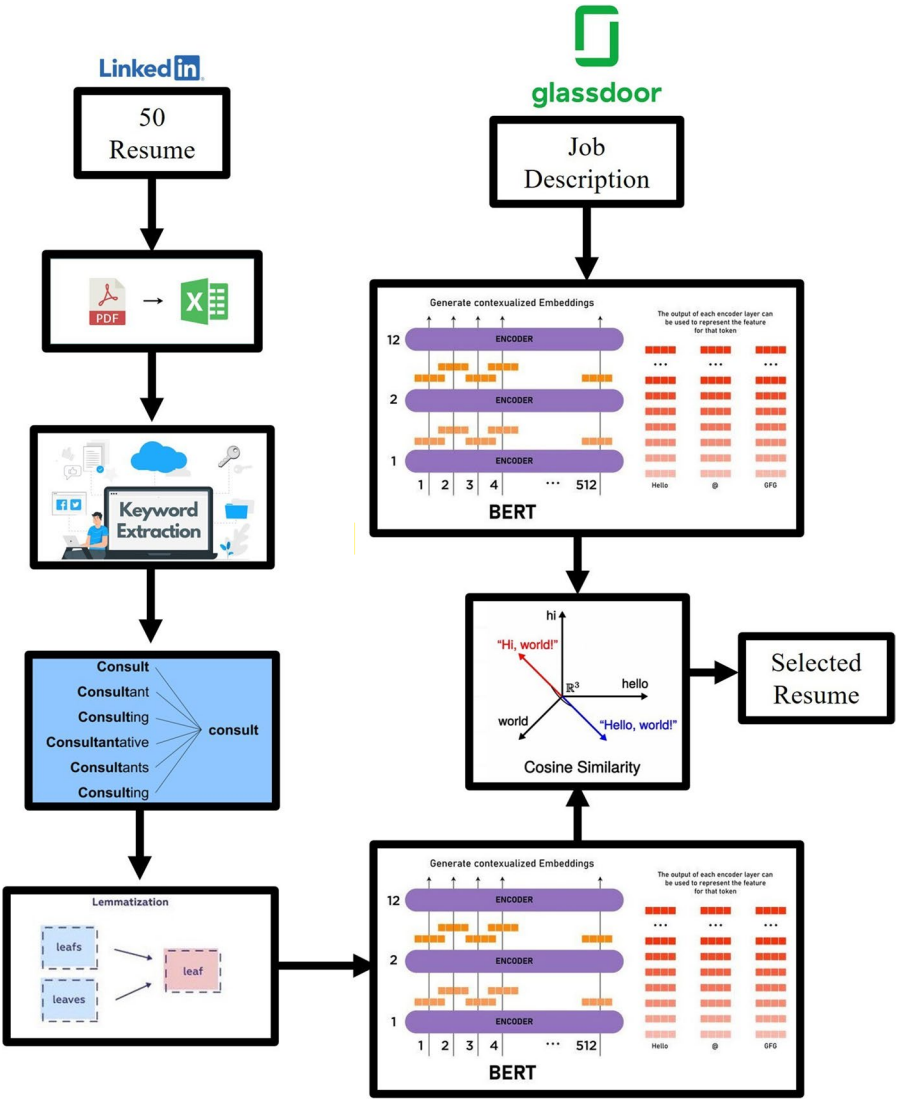

Fig. 1 Proposed mechanism for automated resume screening

sourced from reputable platforms like Glassdoor. Using cosine transforms, matching scores are computed, enabling the automated shortlisting of resumes, which are then presented to users for further evaluation and processing.

2 Literature Review

In recent literature, several studies have leveraged advanced Natural Language Processing (NLP) techniques, particularly the bidirectional encoder representation from transformers (BERT) language model, to enhance various aspects of the recruitment process.

2.1 Resume Parsing and Matching Solutions

Bhatia et al. [2] present an integrated solution encompassing resume parsing and candidate selection using BERT. Their approach employs a heuristic and BERT-based model to standardize resumes and rank them based on job suitability.
Lavi et al [6] introduce ”consultantBERT,” a fine-tuned Siamese Sentence-BERT model that excels in matching job seekers with job postings. It surpasses both unsupervised and supervised baselines, exhibiting significant improvement in ROC-AUC.

2.2 BERT-Based Solutions

Li et al. [7] delve into the BERT-BiLSTM-CRF model for extracting information from English resumes. The integration of BERT enhances semantic understanding, while the BiLSTM-CRF framework refines information extraction. Experimental outcomes underline the model's efficacy in improving accuracy.

Bhoir et al. [3] propose an innovative hybrid Spacy Transformer BERT and Spacy NLP approach for parsing resumes, demonstrating its superiority over existing parsers in terms of accuracy and efficiency.

Tallapragada et al. [10] advocate for BERT vectorization in resume classification. Their methodology employs BERT tokenization and embeddings to decipher the contextual meaning of resumes, enhancing classification accuracy. Meanwhile, Jagdish et al. [5] employ BERT to construct a relationship graph of Chinese characters within the financial domain. Their lightweight blockchain-based BERT model (B- BERT) extracts information from unstructured financial resumes and populates user information templates to form accurate user relationship graphs.

2.3 Hybrid Approach

Athukorala et al. [1] targeted limited human resources in IT companies, presenting a Business Intelligence Assistant for Human Resource Management. Their solution, comprising a Structured Resume Analyzer, Smart Candidate Ranker, Employee Engagement Survey Generator, and Business Intelligence Processor, streamlines processes and decision-making in HR management.

Collectively, these studies underscore the transformative potential of BERT-based NLP techniques in reshaping recruitment strategies, optimizing candidate matching, and streamlining HR processes [9]. The subsequent sections delve into the intricate details of these methodologies, their experimental evaluations, and the broader implications for modern talent acquisition practices.

3 Methods

3.1 Data Collection and Preprocessing

The research commenced by gathering a dataset comprising 200 resumes from diverse participants, ensuring ethical compliance through their informed consent. Additionally, 10 job descriptions were sourced from glassdoor.com to serve as benchmarks for testing the automated screening techniques. The collected resumes underwent thorough pre- processing steps to ensure data quality. All extracted words were processed to identify skill sets and keywords relevant to the applicants' qualifications and experience.

3.2 Keyword Extraction and Table Creation

The identified skill sets and keywords were organized systematically by compiling them into structured tables using Excel files. In alignment with best practices, the tables underwent further refinement through the elimination of stop words and punctuation marks, mitigating noise and enhancing the accuracy of the subsequent analysis.

3.3 Keyword Ranking and Feature Extraction

To assign priority to the extracted keywords, a ranking mechanism was employed based on their significance to the desired attributes of prospective candidates. The top n keywords were selected for further analysis. Subsequently, stemming and lemmatization techniques were applied to address variations in tense and meaning, ensuring data coherence and precision.

3.4 BERT Model Integration

The research capitalized on the capabilities of the preinstalled BERT model by Google, a state-of-the-art bidirectional encoder representation. Both the job descrip-

tions and the top keywords extracted from the resumes were individually passed through the BERT model and its corresponding tokenizer. This process generated feature vectors, encapsulating contextual information and semantic meaning.

3.5 Similarity Calculation

The BERT-generated feature vectors were instrumental in calculating the similarity index between each job description and the corresponding resumes. By evaluating the extent of match between the feature vectors, the research identified the highest similarity index.

3.6 Resume Shortlisting

Resumes with the highest similarity indices were selected, effectively shortlisting candidates with the most relevant qualifications and experience. The shortlisted resumes were subsequently forwarded to the HR department for manual verification and processing. HR professionals undertook a thorough review of the shortlisted resumes, making final decisions regarding candidate selection based on the BERT-enhanced screening results.

The methodology adopted in this research draws upon the integration of BERT-based NLP techniques with traditional resume screening processes. The application of advanced NLP methods for automated resume screening promises to significantly expedite the talent acquisition process, optimizing the identification of suitable candidates while minimizing human bias and resource expenditures. The following sections of the paper will delve into the experimental setup, the outcomes of the automated screening process, and the implications of the study for modern recruitment practices.

Figure 2 illustrates the flowchart depicting the automated resume screening and candidate ranking mechanism proposed in this study.

Figure 2 depicts the flowchart of the proposed system. Initially, the resume is read and converted into a readable file format. Next, keywords are extracted and passed on to the function for lemmatization and stemming. The remaining keywords are then ranked and the BERT function is applied to extract the feature vector. Similarly, the job description (JD) is also converted to feature extraction. Both features are then subjected to a cosine similarity check. Once all resumes are checked, they are rearranged according to their matching with the JD, and a rank is generated to inform the employer to enhance data quality. The ranking of these keywords allowed us to focus on the most significant attributes sought in potential candidates.

4 Results

Figure 3 presents a visualization demonstrating the impact of job description length on both the sum of feature vectors and the total number of feature vectors used in the analysis.

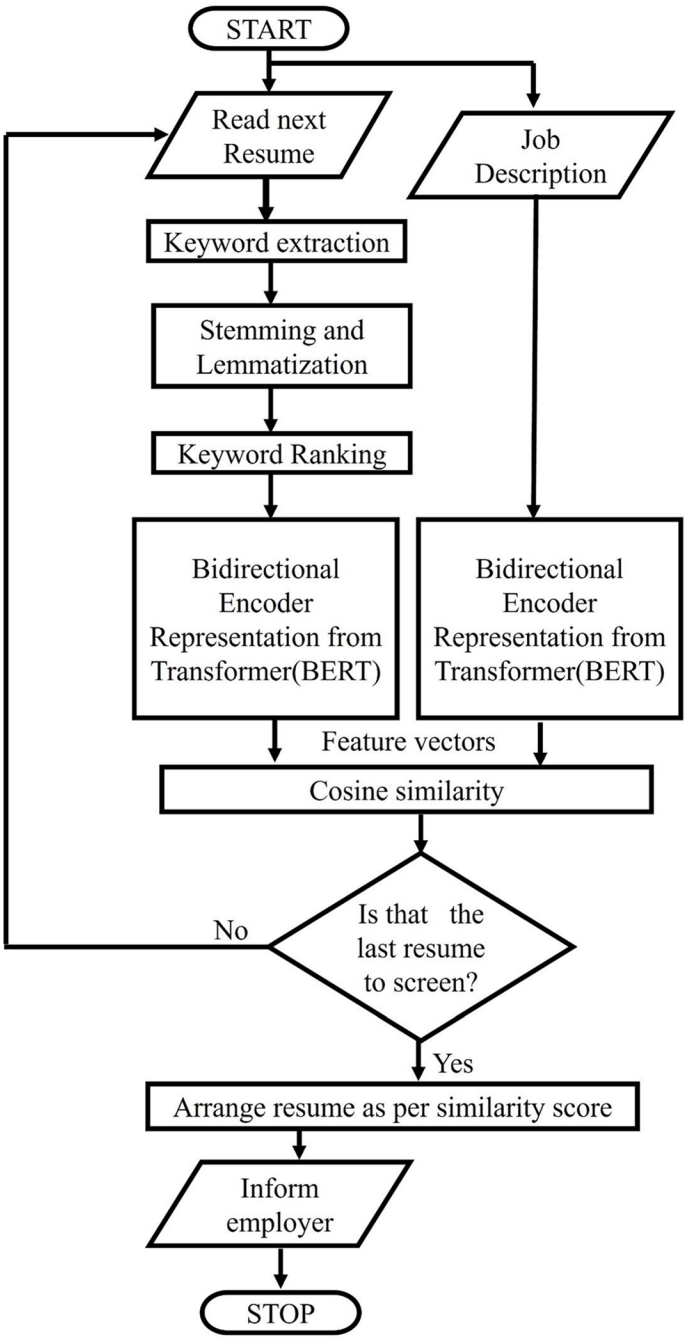

Fig. 2 Flowchart of proposed automated resume screening and candidate ranking mechanism

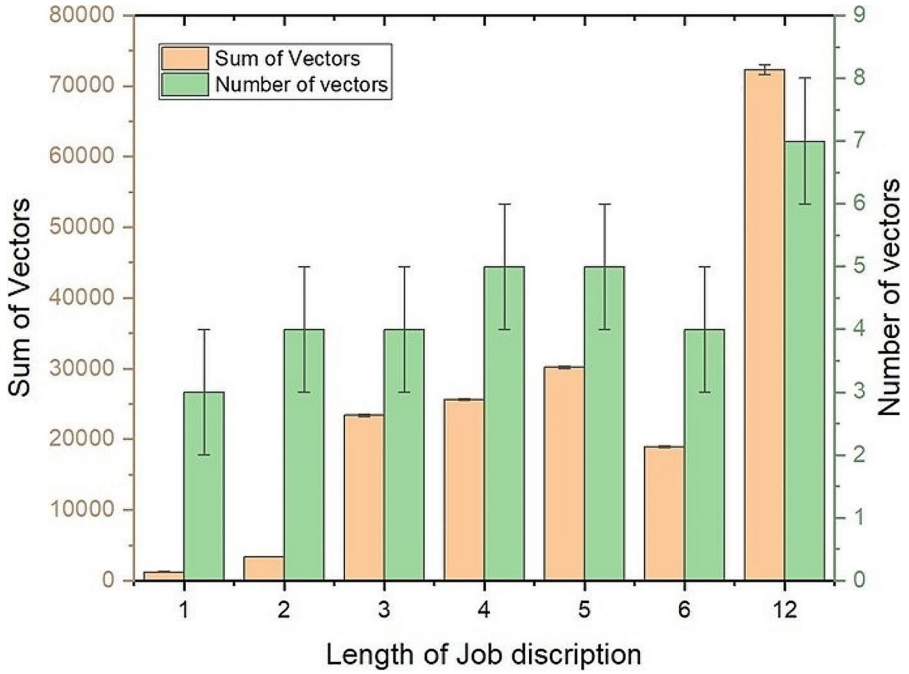

Fig. 3 Effect of length of job description versus sum of feature vectors and number of feature vectors

The application of BERT-based Natural Language Processing (NLP) for automated resume screening and candidate ranking yielded insightful outcomes, highlighting the efficacy and potential of this innovative approach. In this section, we present a comprehensive overview of the results obtained from our experiments, showcasing the performance and impact of our proposed methodology.

We initiated our study by collecting 200 resumes from diverse participants, each obtained with their informed consent. Additionally, we sourced ten job descriptions from glassdoor.com to serve as the basis for evaluating the effectiveness of our automated screening techniques.

Figure 4 illustrates the relationship between the quantity of resumes and the execution time of the proposed algorithm. It is evident that as the number of resumes rises, the processing time also experiences a linear increase. Rate can reach more than 1 resume per second.

Figure 5 depicts the graphical representation illustrating the connection between the number of words under consideration and the similarity index. The graph's behavior demonstrates that the similarity index experiences rapid initial growth, but it saturates after reaching a certain threshold of word count. Based on this observation, we can infer that for the proposed method, an optimal range of 6 to 10 words is adequate for achieving the most effective resume filtering results.

The automated resume screening process involved multiple stages. First, we extracted all words from the collected resumes and identified skill sets and relevant keywords. These keywords were then organized into structured tables stored

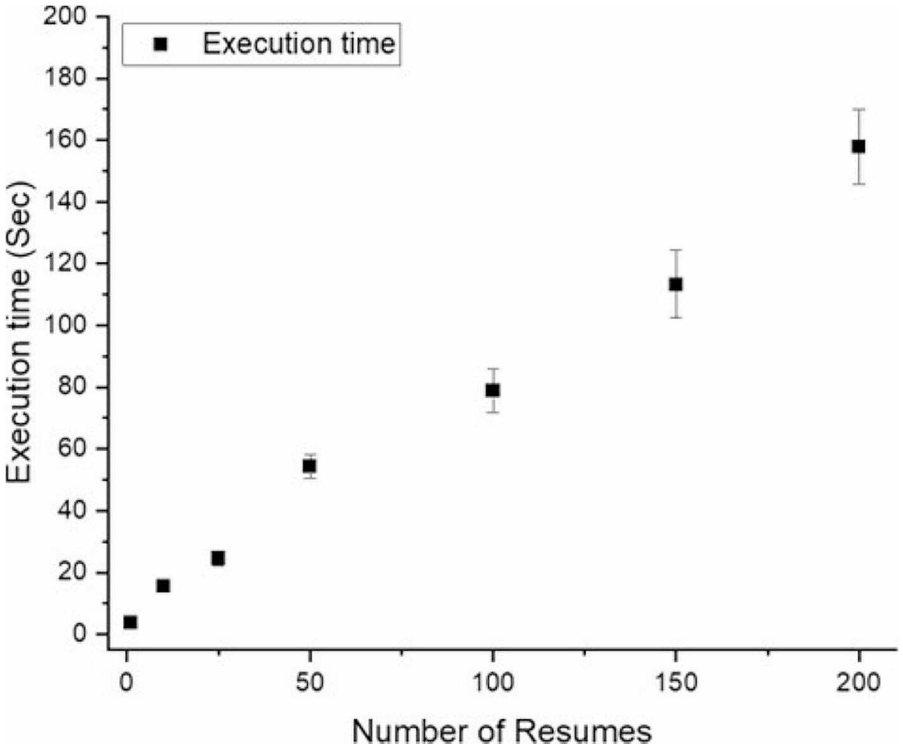

Fig. 4 Graph showing the effect of the number of resumes on the execution time of the proposed algorithm. As the number of resumes increases the processing time also increases linearly

using Excel files, and unnecessary stop words and punctuation were meticulously eliminated.

We analyzed a total of 200 resumes from various disciplines to assess the impact on the accuracy of the proposed algorithm (Fig. 6). The results indicated superior performance with engineering-related terms (0.96), while the accuracy was comparatively lower for screening arts backgrounds (0.75). Resumes from commerce backgrounds demonstrated an acceptable accuracy level, just over (0.85).

To enhance data precision, stemming and lemmatization techniques were employed, correcting tense and meaning variations present in the resume text. Subsequently, the preinstalled BERT model and tokenizer played a pivotal role in generating feature vectors for both job descriptions and extracted resume keywords.

The core outcome of our approach was the calculation of a similarity index for each resume, facilitating the identification of the most relevant candidates. By utilizing the highest similarity index, we shortlisted resumes that exhibited a substantial alignment with the respective job descriptions. These shortlisted resumes were then made available for manual processing by the HR department.

The results of our experiments indicate a notable improvement in the efficiency and accuracy of resume screening through the integration of BERT-based NLP techniques. The calculated similarity scores and computation times offered valuable

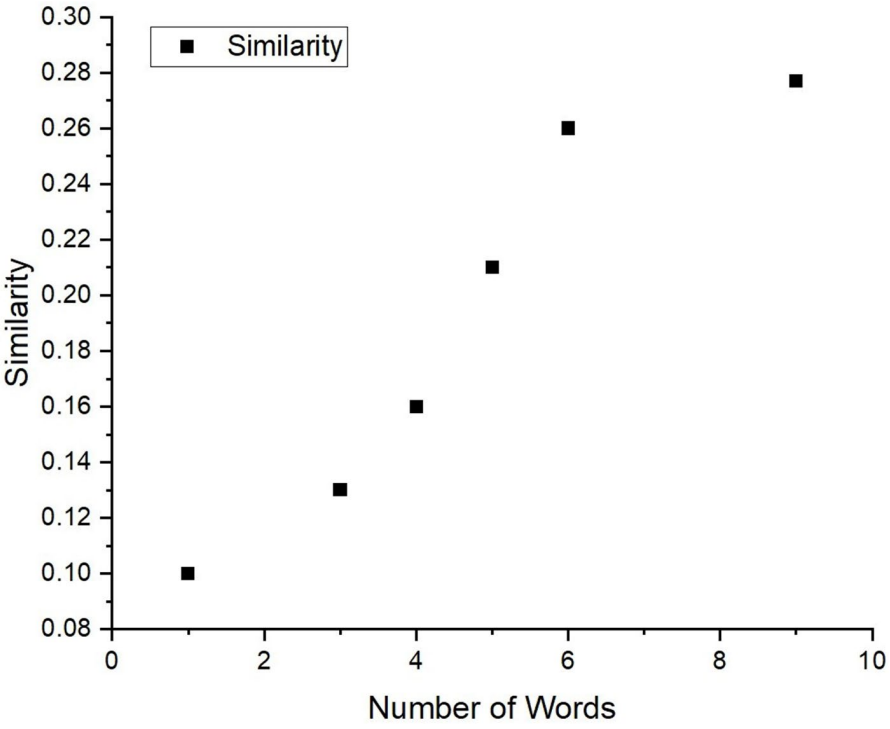

Fig. 5 Graph showing the relation between the number of words under consideration and similarity index. It can be seen from the nature of the graph that similarity saturates after a certain number of words and initially increases drastically. Hence we can conclude for the proposed method around 6 to 10 words are sufficient for achieving best resume filtering

insights into the performance of our approach, underlining its potential to streamline the talent acquisition process. Notably, our methodology demonstrated a significant reduction in manual effort and time required for screening, while also providing a structured and data-driven approach to candidate ranking.

Furthermore, our approach provided valuable insights into the potential of BERT-based NLP in revolutionizing the realm of recruitment. The successful integration of this technology showcases its ability to automate and optimize the initial stages of candidate evaluation, thereby empowering HR professionals to make informed decisions and saving valuable time and resources.

The results of our research underscore the transformative impact of BERT-based NLP on automated resume screening and candidate ranking. The significant improvements in efficiency, accuracy, and data-driven decision-making validate the potential of this approach to revolutionize traditional recruitment processes, ultimately enhancing the overall quality of talent acquisition for organizations. The subsequent sections delve into a deeper analysis of these results and their broader implications for the field of human resources and recruitment.

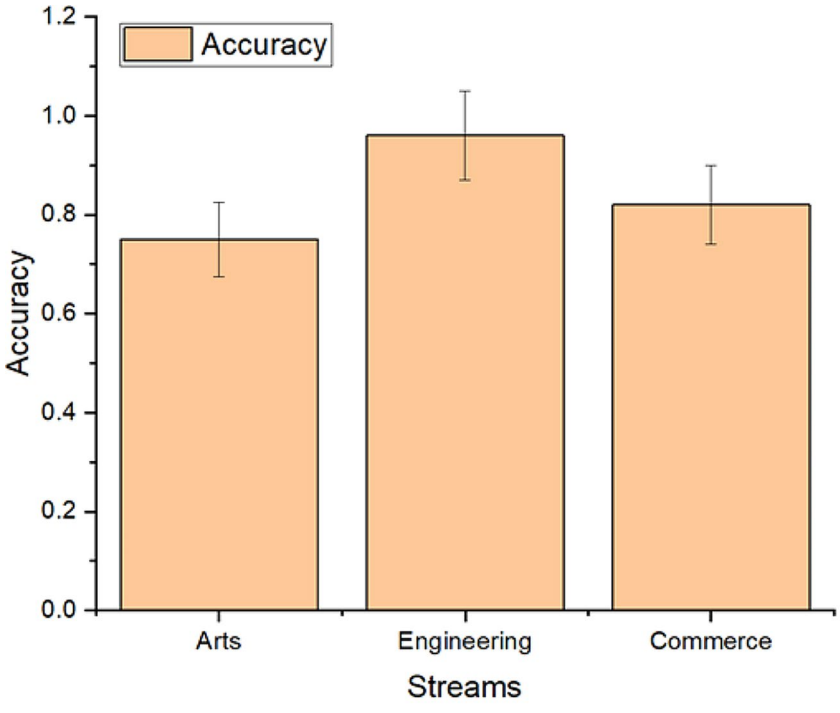

Fig. 6 Bar graph illustrating the impact of academic background on the performance of the proposed algorithm in terms of accuracy

5 Conclusions

This study harnesses BERT-based Natural Language Processing (NLP) to transform automated resume screening and candidate ranking. We showcased BERT's impact on efficiency and accuracy, significantly reducing manual effort and time. Integrating advanced NLP streamlined talent acquisition, empowering HR for informed, data-driven candidate ranking.

Results demonstrated enhanced precision in candidate selection, with similarity scores providing a quantitative measure of alignment between resumes and job descriptions. Our scalable approach, using existing BERT models and tokenizers, highlights adapt- ability within current recruitment infrastructures. Success suggests future refinements, paving the way for more sophisticated resume screening systems.

In a broader context, this research contributes to transforming recruitment practices. Automated screening emerges as a potent tool in navigating competitive job markets. BERT-based NLP's potential to reshape candidate evaluation is underscored, promising elevated standards in talent acquisition. The successful application of BERT-based NLP in automated screening and ranking signals a paradigm shift in recruitment. The fusion of linguistic analysis and technology foretells a future marked by heightened efficiency, accuracy, and objectivity. As organizations

embrace the digital age, such innovative approaches promise to redefine talent acquisition standards.

The proposed methodology places a strong emphasis on specific skill sets. Consequently, individuals who may have omitted or indirectly referenced their skill set might face challenges in comprehending it. The system exhibits a high degree of stringency concerning resumes, overlooking the potential that a selected candidate could acquire the necessary skills. Therefore, there is an opportunity for improvement by incorporating the ability to grasp skill sets based on prior education and experience into the proposed methods. While the system is currently fair, achieving higher accuracy could potentially render the need for human involvement in HR departments obsolete, raising concerns about AI encroaching on human employment opportunities.

Author Contributions The conceptualization was jointly undertaken by Asmita Deshmukh (AD) and Anjali Raut (AR). AD was responsible for data collection, coding, and experimentation. Additionally, AD took the lead in preparing the initial draft of the manuscript, while AR handled corrections. Furthermore, data analysis and graphic design were conducted by AD.

Funding This study is not linked to any funding sources.

Data Availability Upon a reasonable request, both the data and code utilized in this research will be provided.

Declarations

Conflict of Interest The authors affirm that they have no conflicts of interest or competing interests to disclose.

References

1 Athukorala C, Kumarasinghe H, Dabare K et al (2020) Business intelligence assistant for human resource management for IT companies. In: 2020 20th International Conference on Advances in {ICT} for Emerging Regions ({ICTer}). IEEE
2 Bhatia V, Rawat P, Kumar A et al (2019) End-to-end resume parsing and finding candidates for a job description using Bert. arXiv preprint arXiv:191003089
3 Bhoir N, Jakate M, Lavangare S et al (2023) Resume Parser using hybrid approach to enhance the efficiency of Automated Recruitment Processes
4 Carroll M, Marchington M, Earnshaw J et al (1999) Recruitment in small firms: processes, methods, and problems. Empl Relations 21(3):236–250
5 Jagdish M, Shah DU, Agarwal V et al (2022) Identification of end-user economical relationship graph using lightweight blockchain-based BERT model. Comput Intell Neurosci 2022:6546913
6 Lavi D, Medentsiy V, Graus D (2021) Consultantbert: fine-tuned siamese sentence- Bert for matching jobs and job seekers. arXiv preprint arXiv:210906501
7 Li X, Shu H, Zhai Y et al (2021) A method for resume information extraction using BERT-BiLSTM-CRF. In: 2021 {IEEE} 21st International Conference on Communication Technology ({ICCT}). IEEE
8 Reynolds DH, Weiner JA (2009) Online recruiting and selection: innovations in talent acquisition. Wiley
9 Sk¨enderi E (2023) Text Representation Methods for Big Social Data 12

10 Tallapragada VVS, Raj VS, Deepak U et al (2023) Improved Resume Parsing based on Contextual Meaning Extraction using BERT. In: 2023 7th International Conference on Intelligent Computing and Control Systems ({ICICCS}). IEEE
11 Vajjala S, Majumder B, Gupta A et al (2020) Practical natural language processing: a comprehensive guide to building real-world NLP systems. O'Reilly Media
12 Wu L, Qiu Z, Zheng Z et al (2023) Exploring large language model for graph data understanding in online job recommendations. arXiv Preprint arXiv :23070572213