Celal Bayar University Journal of Science

Determining a Proper Text Similarity Approach for Resume Parsing Process in a Digitized HR Software

Yusuf Özçevik[1*], Fatih Yücalar[1], Murat Demircioğlu[2]

[1] Manisa Celal Bayar University, Hasan Ferdi Turgutlu Faculty of Technology, Software Engineering, Manisa, Türkiye
[2] Bilge Adam Technologies, İstanbul, Türkiye
* yusuf.ozcevik@cbu.edu.tr
* Orcid: 0000-0002-0943-9226

Received: 28 December 2021
Accepted: 17 October 2022
DOI: 10.18466/cbayarfbe.1049845

Abstract

Resume parsing is one of the costly phases of a recruitment process. This phase has been alleviated in digitized human resources recently by using text processing approaches between a job advertisement content and resume of applicants. For this purpose, performing a text similarity calculation is one of the most commonly used approaches. However, there are lots of similarity calculation models and most of them are not targeted a recruitment process. Moreover, a subjective assessment of such approaches is required to provide a proper text processing in such a specific problem domain. Thus, in this paper, we offer to evaluate different similarity score calculation approaches through a recruitment case study with the help of a statistical assessment. For this purpose, a computer-aided resume evaluator on a set of resumes is proposed, a human evaluation on the same set of resumes is performed by the professions and the correlation between the outcomes is sought out. As a conclusion, a discussion among different similarity score calculation approaches available for resume processing is presented to find out a proper computer-aided resume evaluator for digitized human resources.

Keywords: Machine Learning based Recruitment, Digitized Human Resources, Text Comparison, Natural Language Processing

1. Introduction

Recruitment processes in many corporate companies consist of a series of major stages. These stages can be illustrated and enumerated as shown in Figure 1. Firstly, a job advertisement content is created according to the requirements and shortly after, the resume collection process from the applicants begins. Afterwards, resume evaluation stage is conducted to identify appropriate applicants and oral and/or written interviews are made according to the evaluation outputs. At the end of the interview evaluation process, the recruitment process is concluded by employing one of the applicants or discarding all of them. In particular, such a recruitment process is not feasible to be carried out with only human labor, especially the recruitment processes conducted in the human resources (HR) departments of large organizations. Therefore, it is inevitable that HR management should be digitized by computer-aided systems including Machine Learning (ML) techniques and Artificial Intelligence (AI) methods for the current industrial era.

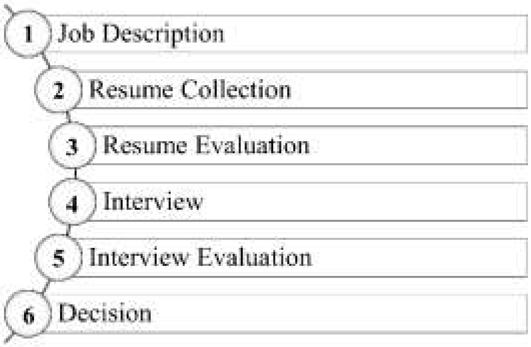

Figure 1. A sample recruitment process.

There are a wide variety of studies in the literature that offer to exploit ML and AI for digitized solutions in different practices of HR management. In [1], the authors reflect the potential of ML approaches to be exploited in HR management considering current technological trending. The authors discuss to place ML instead of human-beings into the center of recruitment processes. In [2], the authors offer to deploy different supervised ML algorithms such as decision tree, random forest, logistic regression, etc. to estimate employee turnover in a company. For this purpose, the authors analyze and evaluate a deck of approaches through the dataset they use. [3] asserts that HR management needs to be digitalized against the recent period of rapid changes in the industry. The authors support their assertion with the arisen of latest employment formats such as gig working, remote working, etc. In [4], the authors offer to use ML techniques in order to alleviate the burden of HR management departments as well as providing a concrete wage forecasting model. They practice a model with neural networks approach and present the evaluation results on the test scenario.

The recruitment, a specific practice of HR management, is also an appropriate candidate for a computer-aided system collaboration. In the literature, it is seen that there are discussions made and solutions offered to digitize recruitment processes at different levels considering the stages included in a recruitment process. In [5], the authors propose to use ML for pre-hire processes of an employee against a job advertisement. Moreover, they also utilize mathematical optimization to obtain a ML collaboration. Finally, they evaluate their model through the management of an entire recruitment process. In [6], the authors propose ML to classify job title for the resume of applicants as a decision support system in HR management. The authors focus on a single stage of recruitment process and from this regard, the study differs from the previous one. In [7], the authors study on AI and ML applications for HR management from a different perspective i.e., fairness. They highlight a fairness overview for recruitment processes and try to provide proper methods and tools in a computer-aided recruitment. In [8], the authors offer a fake job advertisement detection using ML techniques for recruitment. They deploy and analyze an ensemble classifier to find out spam contents in addition to a simple classifier analyze. As a consequence, they conclude that malicious recruitment detection can be obtained through the system model proposed. In [9], the authors propose to utilize ML techniques for functional recruitment of applicants. They mainly focus on the classification of applicants with respect to some features such as age, gender, total work experience, etc. They finally state that such ML collaboration can be successfully used in recruitment stages in HR management.

It can be claimed that one of the stages that effectively be supported by computer-aided systems is the resume evaluation stage. Many HR organizations prefer to collect resume documents and process them with the help of digital solutions. Hence, the oral/written interview processes can be carried out with fewer applicants by eliminating some of them. Most of the common methods used for the resume evaluation stage offer to use text comparison approaches. These approaches based on similarity score analysis introduced in the literature. [10] presents and discusses existing studies about the similarity score analysis between words, sentences, and documents. The authors introduce string-based, corpus-based and knowledge-based similarity approaches to categorize existing studies. Moreover, it also stated in the study that, a combination of such approaches is also applicable for case specific problem domains. [11] introduces a graph-based document similarity approach to assess the relatedness of documents under consideration. The author aims to take relationship between two documents into account when comparing them. [12] presents different problem domains where text similarity approaches can be applied. The authors provide a broad taxonomy among the approaches used and an extensive survey on the application fields. In [13], the author highlights different sentence-based text similarity approaches. Proper application fields of such methods are presented in detail in the study. In [14], the author focuses on information retrieving from text contents using text similarity especially for short texts. The authors offer a semantic-based approach in the study.

It can be inferred from the literature that string-based text similarity approaches fit to the resume parsing in the recruitment problem domain in which matching key terms between a concrete job advertisement and resume file contents are important. However, all the aforementioned studies require clear text content for the text comparison approach used. Hence, in HR management, this appears to be a restrictive element for resume documents collected as files. Moreover, the studies conclude the superiority of an approach depending on the problem domain generally within a subjective manner. In this context, this study proposes a resume evaluation method including a document pre-processing and text comparison approach that can be used in the early stages of the recruitment process for a digitalized HR infrastructure. Furthermore, a statistical analysis is conducted to evaluate the approaches within an objective manner and to determine the proper one(s) through a case study. The contributions of the proposed system model can be summarized as follows:

- Converting resume files from a variety of formats to clear text contents with a document pre-processing stage,
- Ranking resume contents by different string-based text comparison algorithms considering their compliance with the job advertisement content,
- Calculating Spearman Sorting Correlation Coefficient (SSCC) between the computer-aided system with different text comparison algorithms and an expert evaluation system,
- Deciding appropriate text comparison algorithms with the aid of statistical analysis for a digitalized HR infrastructure introduced in the case study.

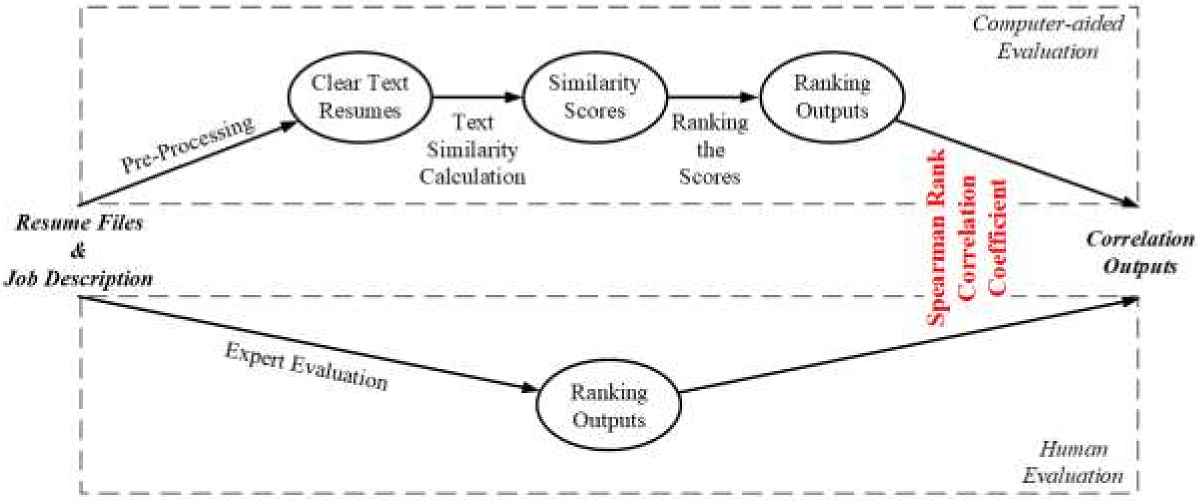

Figure 2. The proposed system model to determine a proper text similarity approach for resume parsing.

The rest of the paper is organized as follows: In section 2, the proposed system model is introduced in detail. Then, in section 3, the evaluation environment and methodology are stated, and the evaluation outcomes are presented with a discussion. Finally, in section 4, the study is concluded, and future directions are implied.

2. The Proposed System Model

The proposed system model to determine a proper text similarity approach for resume parsing is given with Figure 2. There are two different evaluation systems in the model named as computer-aided evaluation and human evaluation. The aim of the model is to identify proper computer-aided resume evaluation approaches on a verified dataset including resume files and apply it confidently through different recruitment processes for similar jobs. Thus, a resume pool including different applicant files and the job advertisement content are processed through a chain of stages by the proposed computer-aided system. Meantime, the data is also investigated by experts through a human evaluation process, simultaneously. Finally, different text similarity approaches used in the proposed computer-aided evaluation model are evaluated according to the correlation values obtained by a statistical method, SRCC, between the computer-aided evaluation and human evaluation. As a result, appropriate string-based text comparison approaches are identified according to the evaluation outcomes.

2.1. Computer-aided Evaluation

The computer-aided resume evaluator in the proposed system model is developed by a set of tools within the application stack illustrated with Figure 3. There is an operating system for both a development environment and a production environment at the bottom of the stack. Microsoft Windows 10 (MW 10) is used in the development environment (Dev) and Microsoft Windows 8.1 (MW 8.1) is used in the production environment (Prod). In the middle of the stack, Visual Studio (VS 2019) is chosen as the Integrated Development Environment (IDE) used through the development process and Internet Information Services (IIS 8.5) is used as the application server for the production environment. On the top of the stack, C# programming language and asp .net framework are determined as the programming technologies for the proposed system. As a summary, a computer-aided resume evaluator is obtained by the combination of these components of the application stack.

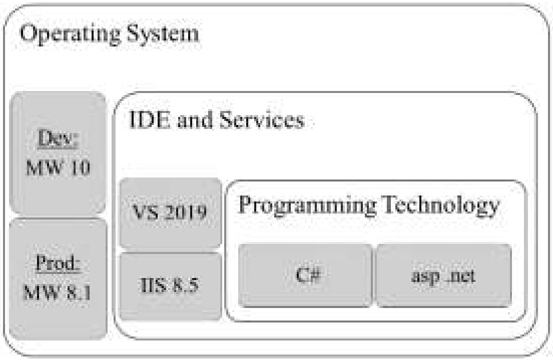

Figure 3. Application stack of the computer-aided evaluator used in the proposed system model.

There are 8 different string-based text similarity approaches considered in the proposed system model to establish a proper computer-aided evaluation. A common feature of these approaches is that they produce normalized values between [0,1] as the similarity scores. Through the implementation of these approaches in the proposed computer-aided evaluator, the distance values are calculated so that the lower value corresponds to the more similarity between two text sequences. Thus, in particular, a higher distance value obtained from these approaches refers a greater dissimilarity i.e., the distance between text contents.

The definitions and details of these approaches used in the study are explained as follows.

2.1.1. Normalized Levenshtein

The Levenshtein distance given in [10] computes a similarity score between two words considering the minimum number of character edits. Here, a character edit refers to an operation needed to make one of the words to the same to the other. The operation can be either an insertion, a deletion, or a substitution for a single character. The Normalized Levenshtein is computed by dividing the Levenshtein value to the length of the longest string under consideration as given in [15]. Hence, a normalized similarity score in the interval of [0,1] is obtained.

2.1.2. Jaro-Winkler

Jaro-Winkler algorithm [16] is an extended version of the Jaro algorithm [17] and computes a distance between two text sequences considering both the number of edits and the number of common characters in the strings. Moreover, a prefix matching is weighted in a comparison. In other words, the similarity between the beginnings of two strings is scored with a higher similarity value. Hence, it is known to be perform better for comparing short strings such as names, adjectives, terms, etc.

2.1.3. Metric Longest Common Subsequence

The Longest Common Subsequence (LCS) given in [10] is used to measure the length of longest common parts of two different strings. Hence, a similarity value without an upper bound can be achieved. On the other hand, the Metric LCS [18] is used to determine a similarity score between two words considering a normalization on the value obtained from LCS. This value is calculated as the division of LCS value by the length of the longest one of the two strings under consideration. Thereby, a value ranging between [0,1] is obtained for the Metric LCS score.

2.1.4. Normalized N-Gram

The N-Gram algorithm [10] scores consecutive occurrence of n characters between two different strings. On the other hand, the Normalized N-Gram [19] achieves a normalization on the similarity score. The normalization is obtained by dividing the similarity score to the original length of the longest word being compared.

2.1.5. Cosine similarity

The Cosine similarity approach [20] computes a cosine value between the angle of vector representation of two text sequences, *V1* and *V2*. The original cosine angle is achieved by the inner (dot) product of these non-zero vectors as V1 · V2. Then, it is divided by the value of length products of these vector, $|V1|\times|V2|$, to obtain a normalized value. The concrete formula to compute Cosine similarity score (*CSS*) ranging [0,1] interval can be given as in Equation 2.1:

$$Css = V1{\cdot}V2 \,/\, |V1| \times |V2| \qquad \textbf{(2.1)}$$

where *V1* and *V2* are the vector representation of two text sequences.

2.1.6. Jaccard index

Jaccard index computes on the common set of words generated by the samples included in two text sequences. Jaccard similarity coefficient [21] is obtained by dividing the intersection of the samples to the union of them. In other words, it gives the computation for the number of shared terms over all unique terms existing in the text sequences. The formula to compute the Jaccard index value (*Ji*) can be denoted as in Equation 2.2:

$$\begin{aligned} Ji &= \backslash A\cap B\backslash \,/\, \backslash A\cup B\backslash \\ &= \backslash A\cap B\backslash \,/\, (\backslash A\backslash + \backslash B\backslash - \backslash A\cap B\backslash) \qquad \textbf{(2.2)} \end{aligned}$$

where *A and B* represent the set of samples in different text sequences, $|A\cap B|$ represents the number of samples in common, $|A\cup B|$ represents the number of unique samples in the union of A and B, $|A|$ and $|B|$ represent the number of samples in A and B, respectively.

2.1.7. Sorensen-Dice coefficient

Sorensen-Dice text similarity approach [22] resembles to the one given by Jaccard considering the utilization of sample sets. However, it can be argued that Sorensen-Dice approach is more intuitive because it can be dedicated as the percentage of overlap between the two sets of samples and twice the number of common ones. Sorensen-Dice coefficient (*SDc*) can be calculated according to Equation 2.3:

$$SDc = (2 \times \backslash A\cap B\backslash) \,/\, (\backslash A\backslash + \backslash B\backslash) \qquad \textbf{(2.3)}$$

where $|A|$ and $|B|$ represent the number of samples in different text sequences and $|A\cap B|$ denotes the number of samples in common between the sets.

2.1.8. Ratcliff-Obershelp

Ratcliff-Obershelp approach stated in [23] simply calculates the number of matching characters divided by the total number of characters in both strings under consideration. The subset of matching characters refers to the ones obtained by LCS approach, in addition to the matching ones in the unmatched region on either side of the LCS parts.

2.2. Human Evaluation

In traditional recruitment processes in an HR organization, resumes are evaluated through a group of expert investigation from different perspectives. Hence, the human evaluation module of the proposed framework includes a traditional human labor processing for resume evaluation. For this purpose, evaluators from both industry and academia participate to the human evaluation process and a consensus is aimed to be achieved by a decision-making process. In this way, all applicants are evaluated through the experience and intuition of the evaluators. As a consequence, the applicants are ranked jointly by the evaluators and a ranking between the applicants is achieved by human evaluation.

2.3. Spearman Rank Correlation Coefficient (SRCC)

SRCC [24] is a statistical approach to test the significance of the correlation between the rankings made by different evaluators. To define SRCC, let *E1* and *E2* are independent ranking evaluations for *n* elements and these evaluations assign numbers as *1*, *2*, *3*, ... , n to the ranking of the elements. If the difference between the order value that evaluations *E1* and *E2* assign to the same element is *d*, then SRCC can be defined as:

$$SRCC=1-[(6\times\sum d^2)/(n^3-n)] \qquad \textbf{(2.4)}$$

where *n* is the number of evaluations made by a single evaluator in the system.

3. Results and Discussion

In this section, the preliminary work and the assumptions about the problem domain are mentioned. The evaluation environment based on the case study to evaluate the proposed system is introduced. The evaluation outcomes are also presented and discussed as follows.

3.1. Evaluation Environment

The dataset used for the evaluation environment of the proposed system model is described in Table 1. According to the table, the data is collected from three different job advertisement in which the recruitment process has completed. The job advertisements are all chosen from information technologies (IT) domain because there has been an unexpected churn rate for current employees in IT firms, recently. The frequent job change of IT employees leads a continues recruitment workload for HR departments. Hence, we believe that such computer-aided solutions can alleviate the burden of HR departments in which employee circulation is high. A brief description about the positions is given in the table, whereas detailed advertisement contents are used through the evaluation. There are total of 18 applicants in the date set in which each six apply for one of the job advertisements.

Table 1. Evaluation environment configuration.

Job #	Position Description	# of Applicants
Job-1	Software developer with C# language and .net framework knowledge	6
Job-2	Network and system administrator for Windows systems	
Job-3	Business analysts for software development processes	

The computer-aided evaluation on the dataset is conducted on a machine with 64-bit Windows 10 OS, 16 GB RAM and 3.7 GHz processing speed. The experiments completed within a reasonable amount of time. Hence, the running time of the system according to the different text similarity approaches is not stated for the sake of simplicity. As a result, a computer-aided ranking is obtained for the applicants.

3.2. Evaluation Methodology

The main workflow of the study has three major components independently proceed with each other. The first one can be stated as a preliminary work that describes a job advertisement and collects resume files of different applicants. In the study, there are 3 different job advertisements prepared to collect the appeals. The content of each advertisement is described including the keywords about the job and prerequisite knowledge level of applicants. Since the content of the job description may affect the accuracy of the evaluation results, the contents are prepared by the professionals of an HR department those have experiences about common practices in recruitment processes.

The second workflow includes computer-aided evaluation of the applicants by parsing the resume files collected. For this purpose, some pre-processing steps are required to prepare the evaluation data and perform evaluations between different text similarity approaches. First of all, the resume data of different applicants are collected as pdf files having a predefined template. Then, the C# PDF library is utilized to obtain clear text information extracted from these files. Accordingly, text similarity analysis can be performed by a software program that utilizes some existing libraries or develops the aforementioned text similarity approaches on its own. There are some known libraries available in github repositories such as *String Similarity*, *Novigo*, *Fuzzy Wuzzy*, etc. those are developed with different programming languages. However, some of the presented text similarity approaches are not implemented in each of the libraries. Thus, in this study, *String Similarity* library is utilized as it is developed with C# programming language that is compatible with the application stack of the proposed framework and missing implementations are also completed. Finally, the computer-aided evaluator ranks the applications by calculating a text similarity score between the job description and the resume contents provided. Hence, an order of preference is obtained among the applicants.

The third workflow implicates the human evaluation of traditional recruitment processes. The human evaluation process is conducted with the participation of three experts whose have experience on IT technologies and software development from 8 to 23 years. The applicants for each of the positions are evaluated through intuition and experience with the participation of all human evaluators. Consequently, an expert ranking is obtained between the applicants and a number is assigned to each appeal that indicates the order of preference. The expert evaluation results are used to calculate a correlation with each of the text similarity approaches by applying SRCC. Thus, the appropriateness of each approach is examined.

3.3. Evaluation Results

The evaluation results for the case study are demonstrated in Table 2. The rows in the table represent different job advertisements, whereas the columns show different text similarity approaches. There are SRCC values in each table cell that correspond to the rank correlation score between computer-aided evaluation and human evaluation. An SRCC value varies between [-1,1] range. The value equals one 1 if there is a perfect matching between the rankings of the computer-aided evaluation and human evaluation. On the other hand, the value equals to -1 if the rankings in the reverse order.

Table 2. Evaluation results showing SRCC values between computer-aided evaluation and human evaluation under different job descriptions.

Job #	Text Similarity Approaches							
	Normalized Levenshtein	Jaro Winkler	Metric LCS	Normalized N-Gram	Cosine similarity	Jaccard index	Sorensen Dice	Ratcliff Obershelp
Job 1	0,54	-0,20	0,54	0,54	0,03	-0,08	-0,08	0,08
Job 2	0,26	0,14	0,26	0,77	0,14	-0,03	-0,03	0,09
Job 3	0,54	0,03	0,54	0,65	-0,14	0,09	0,09	0,03

A visual representation of the evaluation results is given with a polar graph in Figure 4. The center of the octagon corresponds to the value of -1, while the corners of it refers to the value of 1. Hence, SRCC values between human evaluation and different text similarity approaches is clearly seen.

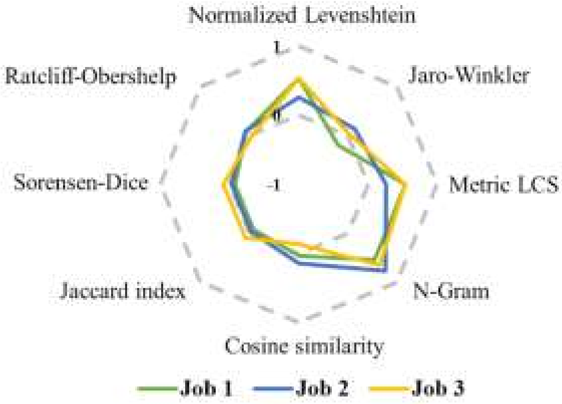

Figure 4. Polar graph of SRCC values between human evaluation and different text similarity approaches.

3.4. Discussion

The evaluation results put a clear comparison between different text similarity approaches. It is clearly seen from Table 2 and Figure 4 that N-gram text similarity approach best fits to be used in a computer-aided resume evaluation. An adequate correlation between human evaluation and N-gram approach can be stated according to SRCC values obtained from different experiments. Moreover, a moderate correlation for human evaluation vs. Normalized Levenshtein and human evaluation vs. Metric LCS pairs can also be stated. When we investigate these approaches, it is seen that both of them utilize edit distance while computing a text similarity. Thus, the correlation between them is predictable.

The remaining five of the eight text similarity approaches can be stated as not giving a meaningful correlation to the human evaluation. The SRCC values obtained from Jaro Winkler, Cosine similarity, Jaccard index, Sorensen-Dice coefficient and Ratcliff Obershelp approaches are around zero that indicates an independent ranking from human evaluation. Moreover, some of them produce the same SRCC values. For example, Jaccard index and Sorensen-Dice coefficient are totally correlated with each other. This situation can be explained in a way that these approaches have similar strategies to compute a text similarity score. They both work on character sets considering their intersection and union. It seems that having twice weight on the intersection set in Sorensen-Dice approach does not change the ranking.

4. Conclusion

Digitized HR management is one of the interesting topics studied recently. There is a variety of applications used in today's HR management in which computer-aided systems are consulted for decision-making processes. One of the application fields targets recruitment processes where human labor is heavily used in traditional HR. There is a set of main stages in a recruitment process and traditional approaches using human labor are costly especially in medium and large corporate companies where frequent staff circulation occurs. An initial stage of a recruitment processing is resume parsing against a job advertisement that requires time and effort of HR departments. However, the cost can be alleviated with a proper computer-aided system that is able to objectively evaluate an application based on the information provided. For this purpose, lots of approaches have been studied in the literature to compare text similarities through a similarity score calculation. Hence, it is essential to determine a proper approach for the success of a computer-aided decision-making process. In this study, a system model to determine proper text similarity approaches is presented and discussed. An evaluation environment is prepared for three different recruitment processes of IT firms in which staff turnover has been very frequent, recently. Accordingly, a statistical correlation between the proposed computer-aided evaluation and human evaluation is investigated to determine the appropriateness of different text similarity calculation approaches. Thus, SRCC method is applied to calculate correlation scores between the human evaluator and different text similarity methods deployed in the computer-aided evaluation system. Finally, a discussion is put forth to present the performance of different approaches. As a consequence, it is believed that digital solutions can contribute to HR departments for achieving cost-effective and objective recruitment processes compared to the human labor-centric traditional approaches.

The dataset provided in the case study is planned to be extended as a future work to investigate further issues. Thus, the suitability of the text similarity calculation approaches will also be sought out for specifically IT domain as well as different functionalities of the proposed digitalized solution will be developed. Moreover, other stages in a recruitment process are also planning to be digitized with recent computer driven approaches especially for IT firms where there has been a frequent turnover between the employees. For example, determining the unexpected churn rate (i.e., attrition rate) for the employees might be very beneficial to have more precise planning about resources and other processes. Hereby, a fully digitized HR management can be achieved with the help of computer-aided components.

Acknowledgement

The study is technically and financially supported by Bilge Adam Technologies. The authors would like to thank the company.

Author's Contributions

Yusuf Özçevik: Drafted and wrote the manuscript, performed the experiments for the case study and helped for the result interpretation.
Fatih Yücalar: Assisted in the evaluation analysis on the case study, result interpretation and helped in manuscript preparation.
Murat Demircioğlu: Assisted in proper resume dataset preparation and human evaluation process for the case study.

Ethics

There are no ethical issues after the publication of this manuscript.

References

[1]. Rąb-Kettler, K, Lehnervp, B. Recruitment in the Times of Machine Learning. In: Management Systems in Production Engineering, Sciendo, 2019, pp 105-109.

[2]. Zhao, Y, Hryniewicki, M, K, Cheng F., Fu B., Zhu X. Employee Turnover Prediction with Machine Learning: A Reliable Approach. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Systems and Applications, Springer International Publishing, 2018, pp 737–758.

[3]. Connelly, C, E, Fieseler, C, Černe, M, Giessner, S, R, Wong, S, I. 2021. Working in the digitized economy: HRM theory & practice. *Human Resource Management Review*, 31(1), 100762.

[4]. Zhu, H, 2021. H. Research on Human Resource Recommendation Algorithm Based on Machine Learning. *Scientific Programming*, pp 2021.

[5]. Pessach, D, Singer, G, Avrahami, D, Ben-Gal, H, C, Shmueli, E, Ben-Gal, I. 2020. Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134, 113290.

[6]. Javed, F, Luo, Q, McNair, M, Jacob, F, Zhao, M, Kang, T. Carotene: A Job Title Classification System for the Online Recruitment Domain, IEEE First International Conference on Big Data Computing Service and Applications, 2015, pp. 286-293.

[7]. Mujtaba, D, Mahapatra, N. Ethical Considerations in AI-Based Recruitment, IEEE International Symposium on Technology and Society (ISTAS), 2019, pp. 1-7.

[8]. Bandyopadhyay, S, Dutta, S. 2020. Fake Job Recruitment Detection Using Machine Learning Approach. *International Journal of Engineering Trends and Technology*, vol 68.

[9]. Reddy, D, J, M, Regella, S, Seelam, S. Recruitment Prediction using Machine Learning, International Conference on Computing, Communication and Security (ICCCS), 2020, pp 1-4.

[10]. Gomaa, W, Fahmy, A. 2013. A Survey of Text Similarity Approaches. *International journal of Computer Applications*, vol 68

[11]. Paul, P. Efficient Graph-Based Document Similarity. In: The Semantic Web. Latest Advances and New Domains, Springer International Publishing, 2016, pp 334–349.

[12]. Wang, J, Dong, Y. 2020. Measurement of Text Similarity: A Survey. *Information*, vol 11(9).

[13]. Farouk, M. 2019. Measuring Sentences Similarity: A Survey. *Indian Journal of Science and Technology*, 12(25), 1–11.

[14]. Rao, J. Bridging the Gap between Relevance Matching and Semantic Matching for Short Text Similarity Modeling, In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp 5370–5381.

[15]. Yujian, L, Bo, L. 2007. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 29 (6), pp 1091-1095.

[16]. Dreßler, K, Ngomo, A, N. 2017. On the Efficient Execution of Bounded Jaro-Winkler Distances. *Semantic Web*, vol 8, pp 185-196..

[17]. Del, M, Angeles, M, García-Ugalde, F, Valencia, R, Nava, A. Analysis of String Comparison Methods During De-Duplication Process. International Conference on Advances in Databases, Knowledge, and Data Applications, Rome, Italy, 2015, pp 57-62.

[18]. Bakkelund, D. 2009. An LCS-based string metric. *Olso*, Norway: University of Oslo.

[19]. Kondrak, G. N-Gram Similarity and Distance. In String Processing and Information Retrieval, Springer Berlin Heidelberg, 2005, pp 115–126.

[20]. Gunawan, D, Sembiring, C, Budiman, M. 2018. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *Journal of Physics: Conference Series*, vol 978.

[21]. Bag, S, Kumar, S, K, Tiwari, M, K. 2019. An efficient recommendation generation using relevant Jaccard similarity. *Information Sciences*, vol 483, pp 53-64.

[22]. Salim, D, Perdana, N, J, Mulyawan, B. 2020. Application of the case based reasoning & sorensen-dice coefficient method for fitness exercise program. *IOP Conference Series: Materials Science and Engineering*, vol 1007(1), pp 012188.

[23]. Kalbaliyev, S. Text Similarity Detection Using Machine Learning Algorithms with Character-Based Similarity Measures. In Digital Interaction and Machine Intelligence, Springer International Publishing, 2021, pp 11–19.

[24]. Zar, J.H. Spearman Rank Correlation: Overview. In: Wiley StatsRef: Statistics Reference Online, 2014.