2022 14th International Conference on Computational Intelligence and Communication Networks (CICN) | 978-1-6654-8771-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CICN56167.2022.10008255

Comparative Semantic Resume Analysis for Improving Candidate-Career Matching

Asrar Hussain Alderham
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
aalderham@stu.kau.edu.sa

Emad Sami Jaha
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
ejaha@kau.edu.sa

***Abstract*—A resume, in general, is a commonly and widely used way for a person to present their competence and qualifications. It is usually written in different personalized methods in a variety of inconsistent styles in various file formats (pdf, txt, doc, etc.). The process of selecting an appropriate candidate based on whether their resume matches a list of job requirements is usually a tedious, difficult, time-consuming, and effort-consuming task. This task is deemed significant for extracting relevant information and useful attributes that are indicative of good candidates. This study aims to assist human resource departments to improve the candidate career matching process in an automated and more efficient manner based on inferring and analyzing comparative semantic resume attributes using machine learning (ML) and natural language processing (NLP) tools. The ranking support vector machine (SVM) algorithm is then used to rank these resumes by attribute using semantic data comparisons. This produces a more accurate ranking able to detect the tiny differences between candidates and give more unique scores to get an enhanced list of candidates ranked from the best to worst match for the vacancy. The experimental results and performance comparison show that the proposed comparative ranking based on semantic descriptions surpasses the standard ranking based on mere regular scores in terms of a distinction between candidates and distribution of resumes across the ranks with accuracy up to 92%.**

***Index Terms*—Resume, Semantic Attributes, Information Extraction, Comparative Description, Ranking**

I. Introduction

The internet has become crucial for helping job seekers find jobs and helping employers find distinctive candidates. Nowadays, applying for jobs is mostly done through online job portals or services. A human resources (HR) officer publishes job requirements, and applicants submit their resumes through a dedicated recruitment website. A resume exhibits job-related knowledge and information to the employers through a printed or electronic document. Usually, it contains personal information, education information, work experience, hard skills such as technical skills, and soft skills like leadership, time management, etc. In addition, a resume is a document that can take one of many structures and be presented in various file formats. There is a likely variability in the construction or context of a large number of applicants' resumes that need to be explored, reviewed, and filtered for better candidate nomination. As a result, extracting useful information and selecting potential candidates can be difficult and challenging for HR professionals. It is, therefore, imperative to keep pace with the astounding advances in machine learning (ML) technologies, as computing power becomes challenging when job requirements require further analysis and re-filtering of resumes by setting more precise criteria for extracting relevant resumes. A lot of research work has been conducted in the field of resume analysis, including classification, summarization, and extracting information using many different techniques. Such research efforts have faced a lot of problems regarding the resulting accuracy of selecting the most eligible candidates' resumes. These problems still need to be addressed with extended resume analyses and further research studies.

In this work, we aim to assist HR officers to improve the extraction, filtering, and selection of the most job-suited resumes in an automated, and more accurate, manner based on analyzing semantic resume attributes using natural language processing (NLP) tools. Thus, for each resume, two different groups of per-attribute semantic descriptions are used as relative and comparative labels. This form of comparative description is more precise and informative compared to categorical descriptions. However, comparative labels need to be derived from a large set of all possible comparisons per attribute by using a ranking method like the ranking SVM algorithm [1], resulting in discriminative ranking for all resumes per attribute or by all attributes.

The rest of this paper is organized as follows. The following Section II presents a brief background about candidate-career matching, the process of extracting information from resumes, and relative and comparative descriptions. Followed by related works of resume analysis with different techniques and approaches in section III. Then, Section IV is devoted to describing the data set, an overview of the proposed approach, and the methodology adopted in this work. Following this, Section V presents the details of the experiment and its results. Finally, the paper concludes in Section VI, and we offer directions for future work.

978-1-6654-8771-9/22/$31.00 ©2022 IEEE
DOI: 10.1109/cicn.2022.54

II. Background

A. Candidate-Career Matching

A candidate-career (job-candidate) matching task assigns the right job to the right candidate (applicant). Candidate-career matching is a tedious process to be carried out by HR officers. Therefore, it is mostly very difficult to thoroughly pass through all candidates' resumes and nominate some of them as good matches according to the job requirements, especially with a great number of resumes that need to be reviewed in a limited or short time [2]. Russell [3] stated that "the selection process is clearly the most critical and controllable variable in the development of a productive and successful work team."

The right matching between the candidate and the job is not only important for HR to hire the right applicants, but also useful for those applicants to avoid involving them in a job that is not right for them. Therefore, the capability of enhancing the matching process of candidates with careers, by correctly selecting and analyzing resumes, is critical for improving accurate candidate employability and establishing an effective employment process, which subsequently helps improve employee job performance.

B. Resume Parsing and Information Extraction

Resume parsing is the automatic process of extracting information via complex pattern matching or language analysis techniques from websites or inconsistently structured documents, such as resumes, in different formats and building a potential recruiting database [4]. Information extraction technologies, such as NLP, take natural language text as input and help to analyze the text efficiently and effectively in order to discover valuable and relevant knowledge that can produce useable structured information. Furthermore, NLP can be employed to understand human language in resumes and then extract their embedded useful information.

The primary task of NLP is to process unstructured text and generate a representation of its meaning through its pipeline processes, involving two levels: the syntactic level and semantic level. At the syntactic level, the raw text of a document is split into sentences using a sentence segmenter, and each sentence or statement is segmented into words and punctuation (i.e. tokens) using a tokenizer. Next, each token is tagged with part-of-speech tags (labelled in the form of noun, verb, adjective, adverb, and so on), which will prove very helpful in the named entity recognition (NER) step. This step identifies all of the occurrences belonging to a specific type of entity in the text. The final step involves using relation recognition to search for likely relations between different entities in the text at the semantic level, where each word is analyzed to determine the meaningful representation of the sentence [5].

As such, we have this background that we can use as a basis for our goal. Our proposed approach is intended to improve resume analysis by using comparative descriptions after extracting relative information from resumes using ML along with NLP techniques, where such comparative descriptions are expected to be more precise and informative in terms of distinguishing between resumes by their semantic attributes, as they have been found superior to their counterpart forms of description in other different domains by several research explorations [6], [1], [7].

C. Relative and Comparative Description

The parsing may fail and not bring in high-quality data if the search is more specific in describing job requirements, since the search in the majority of approaches is typically achieved based on an exact match of the search keyword or an explicit, predefined list of related words. However, other valuable implicit and semantic information is not analyzed or considered in the resulting retrieved data, which can usually and definitely be inferred and considered by HR officers while they are manually performing the search and selection of candidates. Therefore, to fill this semantic gap between how humans and machines analyze resumes, we propose and use a new method based on defining a number of semantic attributes that describe different aspects of the process, like personal information, education level, experiences, certifications, etc. These semantic attributes may be explicitly or implicitly included in a resume, as we need to parse them by searching keywords and giving each attribute a set of descriptive comparative labels that demonstrate what it implicates.

Such semantic attributes can either be binary attributes associated with categorical labels or relative attributes associated with comparative labels [1]. Categorical (absolute) labels can be defined as nameable descriptions used to describe semantic attributes, such as education level can be labeled with any of the categories (BSc, MSc, PhD, etc.). Relative attributes are used to describe the degree of presence using labels, such as "Very high," "High," "Average," "Low," "Very Low" and "None." Whereas comparative labels are named descriptions used to describe only relative attributes, for instance, such that they can be used describe the education level of one resume based on the comparison with the education level in another resume. In other words, these labels describe the degree of comparison of relative attributes, using suitable labels such as "Much higher," "Higher," "Same," "Lower" and "Much lower."

Thus, the relative attributes can represent the strength of the attributes being measured, but can also be comparable, making it easier to more accurately differentiate the small underlying differences in the descriptive attributes between one resume and another [8]. As indicated in [9] there may be several expected advantages to describing the resume features in a relative (comparative) format:

1) It makes resume descriptions clearer and more meaningful (for example, resume A is better than resume B).
2) It enables comparisons with a reference resume object (e.g. resume A is higher than Asrar's resume).

3) It makes attribute-based searching more efficient (e.g. search for a much higher resume with respect to skills).

III. Related Work

A considerable amount of work has been published on the automatic extraction and analysis of information from resumes for improving candidate-career matching and selection. Although there is a wide range of techniques and approaches used in unstructured resume analysis, most researchers focus on ML and NLP approaches for extracting entities from resumes in order to ensure the best match between applicants and job requirements.

A. Resume Analysis and Candidate Selection using NLP

Several researchers are interested in this scope [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. In [10], the author describes the proposed system, which is made up of several modules: section-based segmentation, the filtration, and category-based matching. A ML-based algorithm called "learning-to-rank" is used to build the applicant learning model for ranking, and NLP is used to pull out information.

The authors in [11] proposed a new model that uses NLP methods to summarize resumes in various formats. It does so by extracting information that is relevant to the job requirements and listing each summarized resume in a text file through four integrated modules. They found that the auto-summarisation model of resumes is not an efficient solution because it might not work well in all industries as different skills are needed for different job roles.

In [12], the model works in two steps. First, NER and NLP are used to extract the relevant features from the resumes. LinearSVM is then used to sort the resumes into the correct categories. Second, they use cosine similarity to find similarities between the resume and the job description. Their model was tested in resume classification and achieved an accuracy of 78.53%. With similar technologies used in [13] they implemented a system that extracts job-related knowledge and skills, compares applicants' resumes to the job descriptions, and displays the similarity percentage using cosine similarity for a cost-effective candidate ranking.

In [14], they proposed a resume parser system using the NER of Stanford CoreNLP, and pattern matching. Based on the extracted skills, they used the Term Frequency and Inverse Document Frequency (tf-idf) technique and the logistic regression classification technique to predict the candidate's field of expertise. The overall resume prediction accuracy of their system was 91.47%.

In [15], they proposed semantical and contextual rich information extraction (IE) using the advanced NLP library, SpaCy, for phrase matching. They identify a table/dictionary containing various skill sets, then parse the resumes to search for matching skill words, and count the frequency of these words in various categories. To select the appropriate candidates, they used the Matplotlib tool to visually represent the information after parsing.

In [16] they used optical character recognition (OCR) to extract data from a resume based on NLP and a ranking algorithm. Then, they planned to develop a job portal where employees and applicants could post their resumes, making the recruiting process easy and efficient. When comparing extracted entities and needed keywords, resumes were ranked based on their technical skills. The results were supplied in visualised forms, such as pie charts and bar graphs.

The research work of [17] involved developing a system that automates the eligibility examination and evaluation of master's program candidates. Machine learning, NLP, and three classification algorithms Naïve Bayes, SVM and Random Forest—are used. Hence, candidates were accepted based on scores given after extracting the relevant attributes. Those scores were used to classify and decide whether a resume fell under the accepted or rejected category. The classification results were cross-validated by online video interviews.

In [18], the authors addressed a problem that the resume parser had when supporting languages other than English. They created a system of interconnected NLP models. Their models were evaluated on a data set of 1,686 resumes written in Norwegian, Swedish, Finnish, Polish and English, where the system achieves F1 scores of 0.86, 0.88, 0.86, 0.87 and 0.82 at the resume section level, respectively.

The proposed prototype in [19] designed a hybrid resume parser service using text mining algorithms to extract information from Polish resume documents for use in IT recruitment. The system combines three NER tools (Liner2, NERF, Babelfy) with the anchor NER service and dictionary methods, such as Fuzzy Dict and Competence Lexem.

The education section of a resume required scalable annotated data for NER. A semi-supervised model consisting of CNN, Word Embeddings and Bi-LSTM are used in [20] to identify degrees and institute names on a resume based on NER. This model predicts unclassified education section entities and is corrected using a correction unit. Then, the corrected predictions are added to the training set, resulting in better accuracy of 92.06% than other previously trained model with accuracy.

In [21] the authors designed a resume classifier application that uses NLP to obtain only relevant information, and an ensemble learning-based voting classifier consisting of six individual classifiers to categorize a candidate's profile into a suitable domain based on their interest, work experience and expertise. The resume classifier produces a bar graph of the candidate's domain suitability.

This study in [22] aims to develop an automated resume classification system (RCS) by job category. The study's main contribution is pre-processing resumes for corpus and extracting vectorized representations via NLP techniques. Among nine evaluated ML models on the parsed resumes dataset, the SVM classifier outperformed the others by 96%.

TABLE I
COMPARING OUR PROPOSED APPROACH WITH PREVIOUS STUDIES

Study	ML	NLP	Scoring	Ranking	Comparative	Language
[10]	√	√	√	√	-	English
[11]	√	√	√	-	-	English
[12]	√	√	√	√	-	English
[13]	√	√	√	√	-	English
[14]	√	√	-	-	-	English
[15]	√	√	√	-	-	English
[16]	√	√	√	√	-	English
[17]	√	√	√	-	-	English
[18]	√	√	-	-	-	English, Norwegian, Swedish, Finnish, Polish
[19]	√	√	-	-	-	Polish
[20]	√	√	√	-	-	English
[21]	√	√	-	-	-	English
[22]	√	√	-	-	-	English
Our work	√	√	√	√	√	English

Table I presents our proposed approach compared with a number of previous related studies.

B. *Comparative Description Approach*

The capability of the comparative descriptions is yet to be extensively investigated as pertains to resume analysis. Thus, we widen the exploration to embrace other prominent applications utilizing comparative descriptions, such as human identification by semantic/soft biometrics. This research was motivated by the successful use of comparative descriptions in other applications; therefore, we aim to apply and extend such effective capability as a novel contribution in the domain of resume ranking and selection processes for effective candidate-career matching.

In [7], the authors looked at subject comparisons in the Soton gait database by comparing one target to multiple subjects. When they compared one target to more than one subject, they found that the comparative descriptions performed 17% better than the absolute/categorical descriptions.

Also, in [1], the authors used semantic clothing traits as soft biometrics for human identification. They further explored their validity and efficiency through corresponding comparative descriptions, allowing for more accurate differentiation.

The study in [6] analyzed a data set using gender as a comparative attribute, and found that comparative annotations are more discriminatory than categorical labels. The study's approach on 100 annotated subject images showed correct-match reliability in the top 7% with 10 comparisons, or the top 13% with only five comparisons.

Additionally, in [9], the authors studied human identity through comparative facial soft biometrics using Labelled Faces in Wild (LFW) dataset. Comparative soft biometrics allow each person to be uniquely identified in the database by creating a biosignature with their exact physical traits compared to others. Such comparative descriptions improved searches based on a given comparative trait (e.g., searching for someone younger).

IV. METHODOLOGY

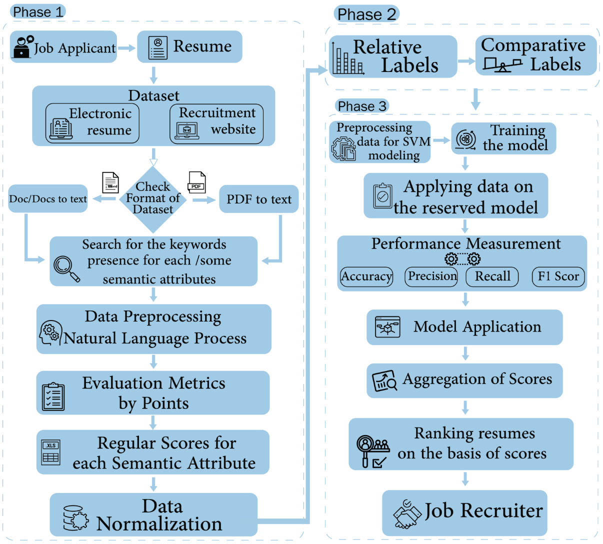

Fig. 1. Overview of the proposed semantic resume analysis approach

As in the literature, there is a need to improve the process of automatically extracting data from resumes and mining their information [23].This can assist in the process of selecting candidates and ranking them according to the job's requirements. In point of fact, recruiters may have their own subjective ideas about which aspects they want to emphasise, depending on the circumstances and nature of the requirements that their company is looking to fulfil at that particular time. Therefore, in this research, we propose a novel approach to improve the candidate-career matching process, based on comparative semantic resume analysis. Our approach is composed of three phases: the semantic attribute extraction phase using NLP techniques, the relative and comparative labelling phase, and the ranking phase using the ranking SVM algorithm. The objective of our research is to provide a ML-based model that does not solely rely on training data to match candidate

TABLE II
PROPOSED RESUME SEMANTIC ATTRIBUTES TO BE EXTRACTED AND ANALYZED

Semantic attributes	Keywords
1- Personal information completeness	Name, Nationality, Address, Mobile No., Phone, Contact No, Email, Country, ...etc.
2- Education level	Bachelor, Master, PhD, GPA.
3- Technical skills level	Programming Languages, Databases, Tools, Framework, Operating systems, ...etc.
4- Professional experience	Experience, Worked, Participated, ...etc.
5- Personal skills (Soft skills)	Communication, Teamwork, Problem-Solving, Creativity, Time Management, Leadership, ...etc.
6- Awards	Winner, Awarded, ...etc.
7-Hobbies and interest	Surf, Play, Games, News, Read, Design, Video, Sports, ...etc.
8- Additional Qualification	Course, Level, Program, Trained, ...etc.
9- Professional Certificate	Certified, Scrum, Professional, Master, Administrator, Management, Cisco, ...etc.
10- Suitable age requirement	Years old, Age, Birth, Birth year.
11- Career objective	Objective, Gain, Opportunity, Looking for, Looking forward, ...etc.
12- Project Experience	Client, Project, Role, Responsibilities, Environment.
13- Languages	English, Arabic, ILETS, STEP, ...etc.

resumes to career requirements, but also enables dynamic learning and ranking per any combination of the proposed resume semantic attributes. Figure 1 summarises all of the methodology phases and presents the flowchart of the proposed approach of semantic resume analysis:

A. Semantic Information Extraction phase

To make the recruitment process simple, efficient and automated, some advanced studies, like [15] and [13], solved the parsing problem with specific searches and unstructured data by searching the whole document for words in predefined tables or dictionaries, or using a ”Bag of Words” (BoW) to list important words and count the number of detected words-of-interest. As a result, we proposed a candidate selection system based on keywords, such that if the keywords exist, it will count their presence and then allocate the given points as evaluation metrics or as HR-specified job requirements. Table II demonstrates the proposed semantic attributes that will be extracted from resumes, as well as all relevant keywords that will be defined and used in our approach. The tools and techniques used at this phase are:

1) Keywords Detection or Information Extraction:

- Extracting files from the archive directory of resumes dataset and generating a list.
- Checking the format of the input file from the dataset (pdf, doc/docs), and converting them to text.
- Extracting semantic attribute information. First, the personal information, name, email, and mobile number are parsed separately using the natural language toolkit (NLTK) and SpaCy models of NLP. Since resumes have rather varied structures and formats, for the 10 remaining semantic attributes, the keywords are similarly searched.

2) Normalization: For many procedures to work properly, the data must be normalised or standardised, with either way, such as using a mean of zero and a variance of one to to normalize the data samples. As a result, since resumes have unbalanced scores after extracting attributes, we used normalisation (MinMaxScaler) to transform features (the regular scores of the attributes) by scaling each attribute to a corresponding range (between zero and one). Then we will simply use the normalization score to get the relative scores, which range from 0 to 5 as explained in the next part.

The transformation is given by:

$$X\_std = \frac{(X - X.min(axis=0))}{(X.max(axis=0) - X.min(axis=0))} \qquad (1)$$
$$X\_scaled = X\_std * (max - min) + min$$

where $min, max = feature\_range$.

Table III shows an example of evaluation metrics by scores given for the personal information attribute, where the total is eight points for the whole personal information attribute and shows the related normalization score which is calculated by “(1)” where $feature\_range$=(0,8), X=8, min=2, and max=8.

TABLE III
EXAMPLE OF EVALUATION METRICS BY SCORES, REGULAR SCORE AND NORMALIZATION SCORE FOR THE PERSONAL INFORMATION ATTRIBUTE.

Personal information	Score
Name	2
Nationality	1
Address	1
Mobile No.	2
Email	2
Total (Regular score)	8
Normalized score	1

B. Relative and Comparative Labeling Phase

After we extract the important semantic attributes from the resume, we come to the comparative description part, to which we contribute to improving the resume analysis domain. In our work, apart from simple traditional evaluation metrics (a regular scoring system), we use the proposed comparative and relative labels to more precisely score or rate a resume.

To infer the relative scores, the points are totaled for each attribute in the regular scoring system after normalization. We multiply each point by 5 and then round it down. The labels of relative descriptions are defined as “Very high,” “High”, “Average,” “Low,” “Very low” and “None”. Table IV shows an example of relative labels and their corresponding scores. After assigning the most appropriate scores based on relative labels for each semantic attribute in the resume, we will accordingly use comparative labels to describe a pairwise comparison per attribute for all of the attributes between each resume and all other resumes in the dataset (one subject with all others). The comparative labels include “Much higher,” “Higher,” “Same,” “Lower” and “Much lower.” For example, when the relative score of personal information is 5 for resume A and is 4 for resume B, then the comparative label for resume A will be labelled as “Higher” than resume B, with the new corresponding score equal to 1, as listed in Table V. This shows an example of comparative labels and their corresponding scores.

We put the comparative scores as the following: the score will be “1” or “-1” if the difference between the relative scores of two resumes is 1 or 2, the score will be “2” or “-2” if the difference between the relative scores of two resumes is more than 2, and the score will be “0” if the relative scores of two resumes are the same. As a result, resume A will be compared to each of the other resumes, one by one, per attribute. Note that when we compare resumes A and B, we do not need the opposite comparative label comparing B and A, because we only compare the possible pairwise combinations, not the permutations. This is due to the fact that the combination relations refer to the combination of n elements taken k at a time without repetition and the order of elements selection does not matter here. Therefore, a k-combination of a set S is a subset of k distinct elements of S. If the set has $n$elements, the number of k-combinations is equal:

$$\frac{n!}{k!\,(n-k)!} \tag{2}$$

whenever $k \leq n$, and which is zero when $k > n$

TABLE IV
RELATIVE LABELS AND CORRESPONDING SCORES.

Relative score system	Relative labels
5	Very high
4	High
3	Average
2	Low
1	Very low
0	None

C. Ranking SVM Phase

The final ranking of a resume per semantic attribute will be based on comparative labels with respect to the other N-1 resumes, using the ranking SVM algorithm. SVM-Rank is a technique for sorting lists of objects using pairwise difference vectors to adaptively arrange given comparable peers based on a certain criterion. The objective of using the ranking SVM is to sort the list of resumes with respect to each attribute, and we can infer the ordering list of resumes for all attributes.

TABLE V
COMPARATIVE LABELS AND CORRESPONDING SCORES.

Comparative score system	Comparative labels
2	Much higher
1	Higher
0	Same
-1	Lower
-2	Much lower

This will enable further search capabilities based on how relevant the retrieved objects are to a particular search query. In our work, all resumes based on comparative scores are placed as pairwise inputs into the SVM ranking algorithm for the prediction of the comparative labels, to obtain finally the ordering list based on comparisons of all resumes. This method helps HR query the list of candidates ranked based on single or multiple semantic attributes in their resume. As such, we can apply a search to retrieve the resume that is “Much higher” than the others in a certain attribute. To rearrange the resumes based on the resulting rankings, we can count (sum) their nascent relative measurement scores based on multiple or all attributes for each resume. The ranking is given in descending order to present the resumes from the best to the worst. A soft-margin Ranking SVM method is used, for a given set of attributes A, to learn a ranking linear function r_a for each attribute, similar to the way used in [1]:

$$r_a(x_i) = w_i^T x_i \tag{3}$$

where w_a is the coefficient of the ranking function r_a and x_i is a feature vector of attributes of a resume being ranked. Rearranging a set of comparisons into two groups can be thought of as a representation of the pairwise relative constraints needed to learn a ranking function. The first group is a set of dissimilarity comparisons D_a of ordered pairs so that $(i,j)\epsilon D_a \Rightarrow i > j$. The second group is a set of similarity comparisons S_a of non-ordered pairs so that $(i,j)\epsilon S_a \Rightarrow i = j$. Then, the following formula is used to get the w_a coefficients of r_a from the D_a and S_a sets:

$$\begin{array}{ll} \text{minimize} & \left(\frac{1}{2}\left\|w_a^t\right\|^2 + C\sum \xi_{ij}^2\right) \\ \text{subject to} & w_a^T(x_i - x_j) \geq 1 - \xi_{ij}; \quad \forall (i,j)\epsilon D_a \\ & \left| \, w_a^T(x_i - x_j) \, \right| \leq \xi_{ij}; \quad \forall (i,j)\epsilon S_a \\ & \xi_{ij} \geq 0 \end{array} \tag{4}$$

ξ_{ij} is the misclassification bias, and C is the trade-off between maximization of margin and minimization error. The resulting optimal w_a function can then be used to (explicitly) rank all training samples according to a. “Equation(3) is used to map a feature vector x_i to a feature vector consisting of a number of real-value relative measurements.”

V. Experiments and Results

The experiment of our proposed approach was conducted on a dataset with 228 resumes in different textual file formats, including doc, docx, and pdf. We have collected various technical specialists' resumes from different sources like LinkedIn, Kaggle, Github, etc. After we implemented the three phases of our approach, we generated 25,878 comparisons between resumes for each attribute. We aggregated the scores and ranked the resumes per each attribute, and for all attributes, to select the candidate who matched the career requirements, as explained in Section IV. The resumes were ranked in descending order, starting with the one with the highest total score and ending with the one with the lowest. Then we conducted our analysis of results based on four comparison aspects:

1) Ranking on the basis of the regular scores. Fig. 2 showed the histogram distribution of resume rankings for all attributes based on the regular scores.
2) Ranking on the basis of the comparative scores. Fig. 3 showed the histogram distribution of resume rankings for all attributes based on the comparative scores.
3) Ranking on the basis of the comparative scores and regular scores. Fig. 4 showed the histogram distribution of resume rankings for all attributes based on the comparative scores and regular scores.
4) Ranking on the basis of the comparative scores and relative scores. Fig. 5 showed the histogram distribution of resume rankings for all attributes based on the comparative scores and relative scores.

TABLE VI
Results of Rankings with Different Four Basis

Basis of rankings	**Max Rank**	**No. of unique scores**	**No. of redundant scores**
Regular scores	75	30	198
Comparative scores	207	187	41
Comparative and Regular	198	171	57
Comparative and Relative	**209**	**194**	**34**

TABLE VII
Accuarcy of Rankings with Different Four Basis

Basis of rankings	**Accuracy**
Regular scores	33%
Comparative scores	90%
Comparative and Regular scores	87%
Comparative and Relative scores	**92%**

We summarize the four ranking results in the following Table VI, which compares the efficiency in terms of distribution and discrimination in the rankings of resumes with high unique values, where all three rankings based on comparative scores outperform the ranking using mere regular scores, which offers very weak resumes' differentiation with high (undesired) redundant values and highly (confusable) similarities in resumes' scores. Note that, with respect to resume distributions (as shown in Table VI, Fig. 3, Fig. 5, and Fig. 4), the more unique scores, the better the ranking result.

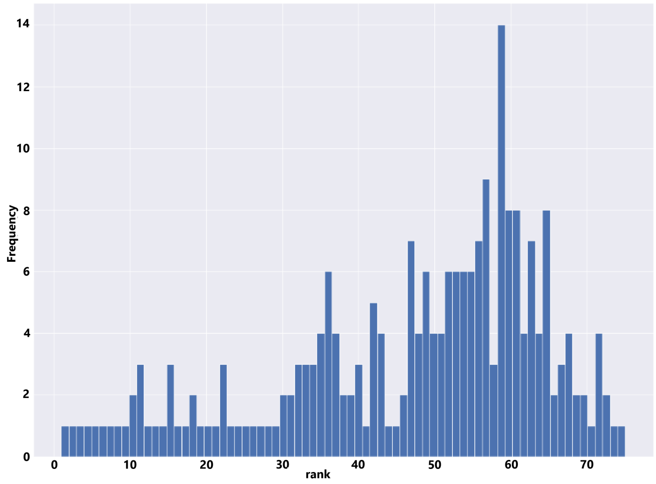

Fig. 2. Resume rankings for all attributes based on the regular scores

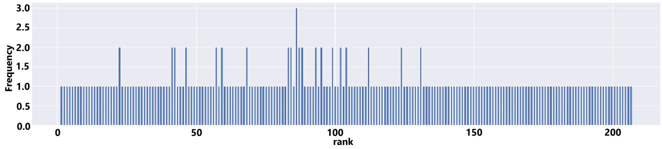

Fig. 3. Resume rankings for all attributes based on the comparative scores

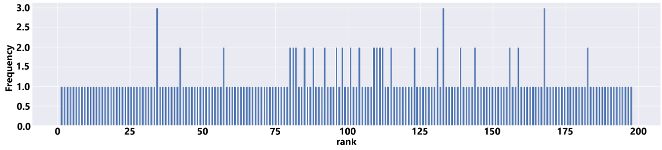

Fig. 4. Resume rankings for all attributes based on the comparative scores and regular scores

We show that this method is very effective by figuring out comparing its accuracy as shown in Table VII. The quality of ranking can be measured as the accuracy of differentiating (as much as possible) between N compared resumes, resulting in as many as possible different unique scores S, which is, in the best-case scenario, equal to the number of compared resumes (i.e., $S = N$), This implies that each resume has a unique

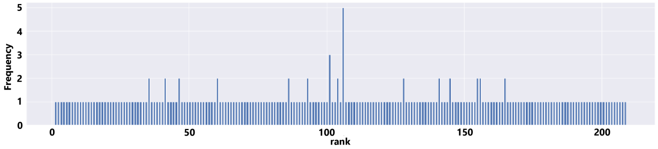

Fig. 5. Resume rankings for all attributes based on the comparative scores and relative scores

(distinct) score. This is done by computing the percentage of the total sum of all unique and redundant scores with respect to their frequencies using the following formula:

$$Accuracy = \frac{1}{N}\left(\sum_{i}^{S}\frac{1}{f_i}\right) * 100 \tag{5}$$

where N is the number of resumes to be ranked. S is the total number of unique scores assigned to N number of resumes to be ranked, and f_i is the corresponding ith frequency of the ith unique score. As such, we note that all methods that rank the resumes based on comparative scores outperform the ranking based on regular scores. Also, the ranks taken from the comparative scores with the relative scores as the basis achieve the highest accuracy in ranking the resumes. This is also separately applicable to each attribute

The following Fig. 6 shows the rankings' histogram of the resumes for the personal information attribute, taking the comparative and relative scores as the basis.

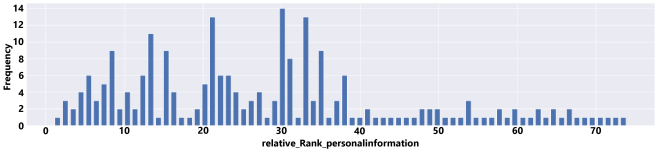

Fig. 6. Resume rankings for personal information attribute on the basis of the comparative scores and relative scores

Finally, Fig. 7 depicts the top ten resumes' (i.e. applicants) rankings with totalling scores on a comparative and relative basis.

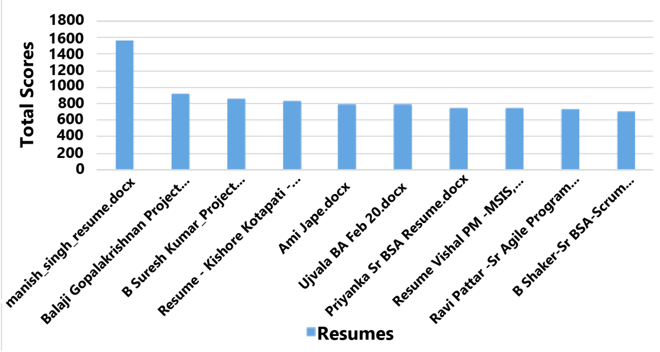

Fig. 7. Top ten resume' rankings

VI. Conclusions and Future work

In the model proposed in this study, improving candidate ranking and selection is performed via three phases: semantic information extraction for 13 attributes using NLP tools; relative and comparative descriptions, which are inferred based on a comparison between every two resumes per attribute; and the ranking SVM algorithm for ordering all resumes per attribute. The comparative description in the candidates' selection process has been proposed to be more precise and informative, and it helps to distinguish the tiny difference between resumes. The histogram result proves that all methods which rank the resumes based on comparative scores outperform the ranking based on regular scores. Also, the ranking attained by the comparative scores with the relative scores as the basis achieves the highest accuracy of 92%.

In the future, the designed model should consider the keyword matching of 13 attributes, and also work on the semantic relations and meanings of the resume language in order to take into account the writing differences in resumes. We will cover uncommon sections that some people may include in their resumes to enrich the analysis and comparison for better candidate-career matching. Also, we will consider generalizing and using the proposed methods with different specializations other than the technical domain and different languages other than English, such as Arabic.

References

[1] Jaha, E. & Nixon, M. Soft biometrics for subject identification using clothing attributes. *IEEE International Joint Conference On Biometrics*. pp. 1-6 (2014)

[2] Lee, D., Kim, M. & Na, I. Artificial intelligence based career matching. *Journal Of Intelligent & Fuzzy Systems*. **35**, 6061-6070 (2018)

[3] Russell, C. Right Person–Right Job: Guess Or Know: the Breakthrough Technologies of Performance Information. (Human Resource Development,2002)

[4] Sinha, A., Akhtar, M. & Kumar, A. Resume Screening Using Natural Language Processing and Machine Learning: A Systematic Review. *Machine Learning And Information Processing: Proceedings Of ICMLIP 2020*. **1311** pp. 207 (2021)

[5] Singh, S. Natural language processing for information extraction. *ArXiv Preprint ArXiv:1807.02383*. (2018)

[6] Martinho-Corbishley, D., Nixon, M. & Carter, J. Analysing comparative soft biometrics from crowdsourced annotations. *IET Biometrics*. **5**, 276-283 (2016)

[7] Reid, D., Nixon, M. & Stevenage, S. Soft biometrics; human identification using comparative descriptions. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **36**, 1216-1228 (2013)

[8] Parikh, D. & Grauman, K. Relative attributes. *2011 International Conference On Computer Vision*. pp. 503-510 (2011)

[9] Almudhahka, N., Nixon, M. & Hare, J. Comparative face soft biometrics for human identification. *Surveillance In Action*. pp. 25-50 (2018)

[10] Nimbekar, R., Patil, Y., Prabhu, R. & Mulla, S. Automated Resume Evaluation System using NLP. *2019 International Conference On Advances In Computing, Communication And Control (ICAC3)*. pp. 1-4 (2019)

[11] Oladipo, F. & Ayomikun, A. A MODEL FOR AUTOMATIC RESUME SUMMARIZATION. *Journal Of Information Systems & Operations Management*. **14**, 147-159 (2020)

[12] Roy, P., Chowdhary, S. & Bhatia, R. A Machine Learning approach for automation of Resume Recommendation system. *Procedia Computer Science*. **167** pp. 2318-2327 (2020)

[13] Khamker, N. Resume Match System. *International Journal Of Innovative Science And Research Technology*. (2021)

[14] Mittal, V., Mehta, P., Relan, D. & Gabrani, G. Methodology for resume parsing and job domain prediction. *Journal Of Statistics And Management Systems*. **23**, 1265-1274 (2020)

[15] Suresh, Y., Reddy, A. & Others A Contextual Model for Information Extraction in Resume Analytics Using NLP's Spacy. *Inventive Computation And Information Technologies*. pp. 395-404 (2021)

[16] Bhor, S., Shinde, H., Gupta, V., Nair, V. & Kulkarni, M. Resume parser using natural language processing techniques. (2021)

[17] Haddad, R. & Mercier-Laurent, E. Curriculum Vitae (CVs) Evaluation Using Machine Learning Approach. *IFIP International Workshop On Artificial Intelligence For Knowledge Management*. pp. 48-65 (2021)

[18] Vukadin, D., Kurdija, A., Delač, G. & Šilić, M. Information Extraction from Free-Form CV Documents in Multiple Languages. *IEEE Access*. (2021)

[19] Wosiak, A. Automated extraction of information from Polish resume documents in the IT recruitment process. *Procedia Computer Science*. **192** pp. 2432-2439 (2021)

[20] Gaur, B., Saluja, G., Sivakumar, H. & Singh, S. Semi-supervised deep learning based named entity recognition model to parse education section of resumes. *Neural Computing And Applications*. **33**, 5705-5718 (2021)

[21] Gopalakrishna, S. & Vijayaraghavan, V. Automated Tool for Resume Classification Using Sementic Analysis. *International Journal Of Artificial Intelligence And Applications (IJAIA)*. **10** (2019)

[22] Ali, I., Mughal, N., Khand, Z., Ahmed, J. & Mujtaba, G. Resume classification system using natural language processing and machine learning techniques. *Mehran University Research Journal Of Engineering & Technology*. **41**, 65-79 (2022)

[23] Sainani, A. & Reddy, P. Extracting special information to improve the efficiency of resume selection process. (Citeseer,2011)