Optimizing Candidate Selection: An Approach for CV-Job Description Matching and Academic Transcript Analysis

Manushika Maldeniya
Department of Information Technology
Specializes in Data Science
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
manushikamaldeniya1999@gmail.com

Shanali de Silva
Department of Information Technology
Specializes in Data Science
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
shanalids6@gmail.com

Maleesha De Silva
Department of Information Technology
Specializes in Data Science
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
maleeshadesilva21@gmail.com

Sandani Zoysa
Department of Information Technology
Specializes in Data Science
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
sandanikitsune@gmail.com

Dr. Anuradha Karunasena
Department of Information Technology
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
anuradha.k@sliit.lk

Dr. Lakmini Abeywardhana
Department of Information Technology
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
lakmini.d@sliit.lk

***Abstract*— As the IT industry rapidly evolved, selecting qualified candidates with the right blend of technical skills, educational achievements, and professional experiences became increasingly challenging for employers. The curriculum vitae (CV) and academic transcripts of applicants play vital roles in the recruitment process. This research paper presents an analysis of the effectiveness of CVs and academic transcripts as screening tools in the IT industry. The study developed a customized RASA-AI chatbot for automating job description generation and implemented a prediction model using a stack of machine-learning algorithms to improve performance. Matching percentages between candidate resumes and job descriptions were computed using TF-IDF and BOW vectorization techniques, with TF-IDF achieving the highest accuracy. The academic transcript analysis employed custom and pre-trained NER models, with the custom NER model showing superior accuracy in identifying module titles and grades. Skill area prediction utilized supervised machine learning models, and from those models, the SVM model showed the highest accuracy. The study visually represented skill areas and assessed technical and professional skills against specific job role requirements. The approach further assessed technical and professional skills against specific job role requirements. Leveraging advanced machine learning algorithms and natural language processing techniques, the proposed solution aimed to streamline the recruitment process, reduce manual efforts, and enhance the accuracy and efficiency of candidate evaluation.**

Keywords*— *Recruitment, CV, Job Description, Skills, Academic Transcript, Natural language processing, Machine Learning

I. Introduction

Recruitment was a critical process for any organization that aimed to hire the right talent to achieve its goals. However, the traditional recruitment process was often time-consuming, inefficient, and costly, leading to suboptimal recruitment outcomes. As per the Sri Lanka Information and Communication Technology Agency's (ICTA) records [1], the country had set a goal of attaining 200,000 employees in the IT industry alone for the year 2022. This emphasized the pressing need for a recruitment system that facilitated more efficient and less time-consuming hiring practices for IT-related job positions. Despite the pressing need for efficient systems to recruit candidates, the IT industry faced several major challenges in hiring the right talent. Two of the major challenges faced by the IT industry were eliminating bias in recruitment and reducing the time-to-hire.[2] Bias could take different forms, such as gender, race, and age, and could limit diversity within the IT industry, preventing it from benefiting from a wide range of perspectives and skills. Similarly, lengthy recruitment processes could lead to the loss of qualified candidates for other job offers or cause them to lose interest. Another major issue faced by recruiters was the inability to use candidate data effectively to identify the most suitable candidate for a job position[3]. While many recruiters were familiar with modern technology tools and their benefits, not all of them effectively utilize technology to analyze candidate data and pinpoint the most fitting individuals for specific roles.

Analyzing data from sources such as CVs along with academic transcripts can provide valuable insights into a candidate's qualifications, skills, and potential fit for a particular job position. By examining these sources, recruiters can gain a deeper understanding of an applicant's educational background, relevant experience, and achievements. This information allows them to assess a candidate's alignment with the job requirements and make informed decisions regarding their suitability for the role. However, the process of analyzing candidate data from these sources can be time-consuming and resource-intensive. The outcomes of this research hold significant implications for the field of HR technology. By integrating machine learning and natural language processing techniques, organizations operating in the IT industry can experience notable improvements in candidate selection, reduced time-to-hire, and enhanced overall recruitment outcomes. The utilization of advanced technologies not only expedites the evaluation process but also ensures that comprehensive and objective assessments are made.

II. Literature Review

Extensive research has been conducted in the field of candidate selection processes, aiming to identify the most qualified candidate from a pool of applicants through a comprehensive assessment of their skills and qualifications.

In a study conducted by Ransing et al.[4] , advanced machine learning techniques were employed to conduct the research. The primary objective of this study was to cater to the specific requirements of recruiters by effectively screening resumes based on their predefined criteria. The data utilized in this study was in the form of a CSV file, with carefully assigned labels corresponding to various roles within the IT industry. A sophisticated stacked classification approach was implemented to systematically rank resumes based on their suitability for different job roles. As a result of this comprehensive research effort, the study successfully obtained a good understanding of the relevance of specific job roles in the IT industry. The findings are more focused on crucial insights that can significantly aid recruiters in making well-informed decisions when evaluating applicants' resumes.

According to Muntaha Mehboob [5], the research utilized a novel approach to extract relevant information from resumes and effectively organize it into distinct categories, such as education, experience, and talents, based on values. This specialized solution was specifically designed to tackle the challenges of unstructured CVs, which often exhibited varying formats and structures. The researchers developed a sophisticated tool capable of automatically shortlisting and ranking candidates based on their job profiles. By employing cosine similarity, the tool efficiently evaluated each resume against the corresponding job description. The implementation of this novel method held considerable potential in enhancing the recruitment process, saving valuable time, and ultimately improving the quality of candidate selection for diverse job positions.

Yan Wang et al. implemented a search engine for recruitment and staffing, employing knowledge graphs and Bidirectional Encoder Representations from Transformers (BERT)[6]. In this research, the utilization of the Machine Learning Model (MLM) and Named Entity Recognition (NER) of pre-trained BERT allowed for the identification of competence keywords from the corpus of CVs and Natural Language Processing (NLP). The knowledge graph, composed of Concept Map (CMAP) and competency keywords, provided effective recommendations in the neighborhood domain, yielding positive outcomes.

The authors [6] explored the use of elite keywords, for selecting significant keywords individually for each class in the context of resume classification. The method involved shortlisting class-specific keywords, removing redundancies, and concatenating them across classes to ensure that relevant class-specific keywords were included in the feature columns. The stability of the method was ensured by the entropy partitioning method. The features were trained on a random forest classifier using a grid search for hyperparameter optimization. After conducting a thorough evaluation of various models, it was found that the elite keyword subset demonstrated the highest reliability and accuracy for classifying different categories of resumes in the benchmark dataset.

According to the study done [7] in 2020 which focuses on candidate selection for an interview for the position of Software engineers, the analysis of the academic transcript played a pivotal role in the candidate selection process for software engineer positions. The researchers recognized the importance of assessing both technical skills and soft skills derived from the academic curriculum. To achieve this, they categorized the modules within the academic transcript into these skill categories. By doing so, they gained a comprehensive understanding of the candidate's capabilities in technical areas, such as programming, as well as their proficiency in soft skills like communication and problem-solving. Furthermore, the researchers devised a grading criterion to evaluate the candidate's performance in each module. Candidates with grades above C+ were deemed eligible for further consideration, as this threshold indicated a satisfactory level of achievement and competence in the respective subject matter. On the other hand, candidates with grades below this threshold were not considered suitable for the job role.

The research paper titled "Analyzing Competences in Software Testing: Combining Thematic Analysis with Natural Language Processing (NLP)"[8] addresses the lack of analysis of competencies required in software testing education for fresh graduates in computer science. The authors use NLP techniques to analyze job descriptions and course 5 syllabi to identify the competencies required in software testing. The findings suggest a gap between the competencies required and the current software testing curriculum, as soft skills such as teamwork, communication, and leadership are not always taught in software testing courses. This paper provides valuable insights for designing software testing curricula that equip students with the necessary competencies to succeed in their careers.

The review of existing literature indicates the emergence of a novel study proposing a holistic approach to candidate evaluation. This methodology encompasses the creation of job descriptions, resume-to-job description matching, and graphical illustrations of skill proficiency. Additionally, the candidate's skills acquired during their academic program are classified by module name, grade, and module description, accompanied by graphical representations. As a result, this approach leads to a streamlined and more effective recruitment process.

III. Methodology

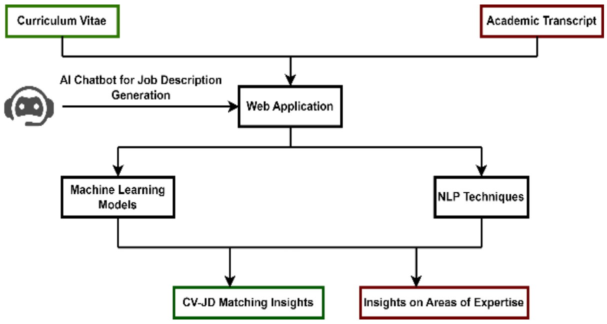


Fig. 1. Overview diagram

A. *Optimizing Candidate Selection Through CV and Job Description Matching Techniques*

This component mainly consists of three subcomponents. Firstly, an AI chatbot generates a job description prompting questions to the user and secondly, it computes matching percentages by comparing the job description with resumes. Lastly, it provides a summary representation of the skill proficiency of each candidate.

I. Automated Generation of Structured Job Descriptions

The initial phase of the project involved the development of a job description generator within an interactive environment, leveraging the capabilities of an AI chatbot. Comprehensive sets of prompts were compiled to facilitate the development of a structured job description, covering essential aspects such as job title, qualifications, and requirements of a specific job position.

The Rasa [9] framework was used to implement the chatbot, incorporating custom actions that facilitated a sequential and interactive approach for generating the job description. The chatbot prompts the user with targeted questions, and uponreceiving inputs, It stores the collected data in designated slots and acts as placeholders for the information gathered throughout the conversation. To facilitate further processing, the collected inputs are then stored in a CSV file.

II. Predicting matching percentage of resume and job description

A dataset [10] containing 90 resumes along with job descriptions and matching percentages was obtained from Kaggle for the model-building process. The resumes were originally in PDF forms and were related to IT industry job positions. As the resumes were in PDF format, a PDF2Text parser was used to extract resume data. The constructed job description was extracted from the CSV file. The length of the resumes ranged from 525 to 1539 characters, making them relatively easy to extract.

Both resumes and job descriptions underwent data-cleaning processes. This included the removal of stop words, concatenation of words, format conversion, space removal, and lemmatization. The 'NLTK' library was used to perform these steps. These steps were performed to ensure that the data extracted from the resumes were in a suitable format for further analysis.

The feature encoding of the resumes and job descriptions primarily focused on extracting numerical features such as common words, word counts in the resumes, and fuzz ratios.These features provided valuable information for subsequentanalysis.

Two methods were employed for feature vectorization:

1. Method 1: Feature Vectorization with TF-IDF vectorizer
2. Method 2: Feature Vectorization with BOW (Bags of Words)

Stacked Ensemble Model

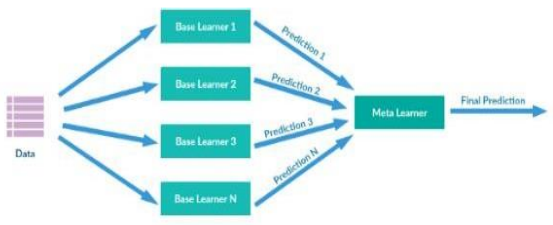


Fig. 2. Stacked model implementation diagram.

The data features were divided into separate train and test sets. Multiple individual machine learning models, including Linear Regression, KNN Regression, Decision Tree Regression, Support Vector Regression (RBF Kernel), Support Vector Regression (Linear Kernel), Random Forest, and Xgboost, were employed for modeling. To optimize their performance, hyperparameter tuning was conducted using the GridSearchCV technique, allowing for the identification of the best parameters and the selection of the most optimal model. To further enhance accuracy, the stacked method was utilized as an ensemble technique. This approach involved combining the individual models that demonstrated favorable accuracies, resulting in a single unified model with significantly improved accuracy. This methodology was applied to the data features obtained from TF-IDF vectorization and Bag-of-Words vectorization techniques.

III. Evaluating Candidate Skills Profile through a Resume Scoring Algorithm

A scoring algorithm was used to analyze the summary of skills required in the job descriptions. This algorithm enabled the creation of candidate profiles by evaluating the match between the skills listed in the resumes and the required skills in the job description. The corpus was made by collecting skills for various key areas.

B. Analyzing academic transcript for a candidate's acquired skills and knowledge.

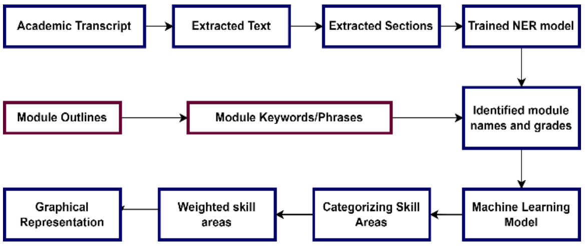


Fig. 3. Analyzing Academic Transcript methodology.

Analyzing academic transcripts targets a particular university and aims to extract and analyze relevant information from academic transcripts and module outlines, which is then used to evaluate the suitability of fresh graduates for specific job roles. Within this research component, it was assumed that the module outlines were obtained from a specific university in Sri Lanka. From these outlines, keywords, and phrases related to the modules were extracted. To analyze the academic transcript, relevant data from the transcript was extracted and organized into distinct skill areas. Subsequently, these skill areas were graphically represented,presenting a visual depiction of the distribution and relationship between different skills. The following approach was used to carry out the methodology for this process:

When collecting data, the module outlines, which provide detailed information about the content and objectivesof each academic module, were obtained from a university. Additionally, academic transcripts from a selected number of students were gathered to analyze their performance. To facilitate the development of a Named Entity Recognition (NER) model and predict skills in the dataset, dummy datasets were prepared and utilized for training.

In data preprocessing, the data cleaning process involves removing unnecessary words, standardizing the format, and segmenting the text into tokens using NLTK. Filtering out insignificant words and using TF-IDF vectorizer for feature extraction. Lemmatization is applied using WordNet to improve machine processing.

I. *Extracting the keywords and phrases from the module outlines*

The module outlines were processed using N-grams to extract important keywords. This technique identified significant phrases and word combinations representing each module's key concepts. The extracted keywords were then saved into a CSV file for subsequent analysis.

II. *Analyzing the Academic Transcript*

To analyze the academic transcript, the module titles and the grades should be identified, this is done using Name Entity Recognition. Two methods were utilized to develop a NER model.

Method 1: Custom NER model
Method 2: Pre-trained NER model

Custom NER model - A Custom Named Entity Recognition (NER) model using the Spacy library was developed specifically to identify module names and their respective grades separately.A new Spacy model is initially loaded, and the custom NER component is added to the pipeline. The labels for the entities of interest, such as MODULE_CODE, MODULE_TITLE, SEMESTER, PERIOD, CREDITS, and GRADE are defined. The dataset containing relevant information is loaded, and the text and entity annotations are structured accordingly. The NER model is trained using the provided training data, with iterations and mini-batch training. The trained model is then saved.

Pre-Trained NER model - Initially, a blank Spacy model is loaded, and a DocBin object is created to store the training data. The training data, obtained from a JSON file, contains text annotations and entity labels. The text and entity annotations are processed, and the entities are added to the DocBin object. This training data is then saved to a file. Next, a configuration file is initialized for the NER pipeline, followed by training the NER model using the training data. The trained model is saved and loaded for testing purposes.

Extracting the academic transcript and skill area categorization - The text extraction process from the academic transcript was done using OCR (Optical Character Recognition) libraries such as py-tesseract and pdf2Image. These libraries enabled the extraction of text from the transcript, enabling further analysis. To extract the academic year information forthe candidate, regular expressions were applied to partition the extracted text into distinct sections based on academic years. These sections were then subjected to the trained NERmodel, which successfully identified the module titles and their corresponding grades. The module titles extracted using the NER and the module keywords obtained from the module outlines were matched and a merged data frame was created. Then it is passed to the model for categorization into relevant skill categories. The identification of the 10 skill areas for the survey was based on thorough research conducted from sources on the web[11], [12]. Furthermore, the population who was considered for this survey were the IT employees from the industry. From the survey below, we can see that the majority of the respondents have selected the first 10 categories.

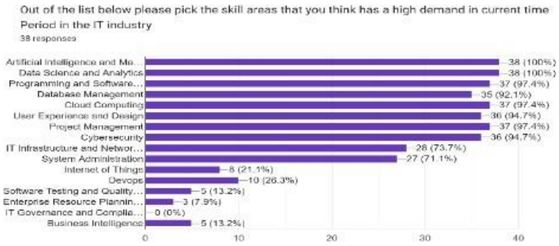


Fig. 4. Survey results of Identifying Skill Areas

Thus, the skill categories are:

- System Administration
- IT Infrastructure and Networking
- Data Science and Data Analytics
- Artificial Intelligence and Machine Learning
- Cloud Computing
- Cybersecurity
- Database Management
- Project Management
- Programming and Software Development
- User Experience, and Design

Furthermore, the grades extracted using the trained NER model were assigned scores based on a weighted scale (e.g., 10 for A+, 9 for A, 8 for A-, 7 for B+, etc.). A skill area table is created, including the category, module title, module description, and weighted grade. Then, the category totals are calculated by grouping the data frame by the "Category" column and summing the "Weighted Grade" column, and the categories are sorted in descending order of their total weighted grades.

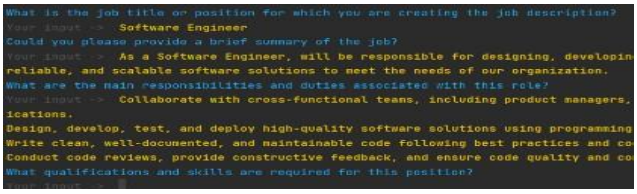
Skill area prediction – Various supervised machine learning models for text classification models were employed to determine the most accurate model for predicting skill categories. First, Word2Vec embeddings are generated using the training text data. Logistic Regression, Naive Bayes, XGBoost, Support Vector Machine (SVM), and Random Forest algorithms are applied to train separate models. For Logistic Regression, the Word2Vec embeddings are used as document vectors, while for Naive Bayes and XGBoost, CountVectorizer, and Tf-idf Transformer are utilized to preprocess the text data. In the case of Random Forest, CountVectorizer is used to convert text data into numerical features. For SVM, CountVectorizer is applied to vectorize the text data, and the resulting feature matrix is then resampled using the Synthetic Minority Over-sampling Technique (SMOTE). The resampled data is used to train the SVM model, which is then evaluated using accuracy. Performance evaluation of these models was conducted through the generation of classification reports and confusion matrices.

IV. Results and Discussion

1) *Optimizing Candidate Selection Through CV and Job Description Matching Techniques*

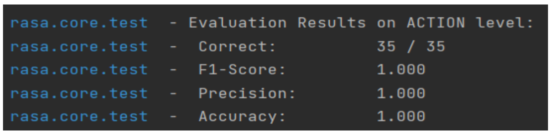
a) *Job Description Bot Implementation*

During the development of the chatbot, a crucial step involves prompting users with questions to gather their inputs, which are then validated based on the length of the provided responses. The implementation of this methodology underscores the chatbot's commitment to quality user interactions and accurate information retrieval. By following a sequential questioning strategy and employing length-based validation, the chatbot can better comprehend user requirements and provide more effective, relevant, and personalized responses.



```
What is the job title or position for which you are creating the job description?
Your input ->  Software Engineer
Could you please provide a brief summary of the job?
Your input ->  As a Software Engineer, will be responsible for designing, developin
reliable, and scalable software solutions to meet the needs of our organization.
What are the main responsibilities and duties associated with this role?
Your input ->  Collaborate with cross-functional teams, including product managers,
ications.
Design, develop, test, and deploy high-quality software solutions using programming
Write clean, well-documented, and maintainable code following best practices and co
Conduct code reviews, provide constructive feedback, and ensure code quality and co
What qualifications and skills are required for this position?
Your input ->
```

Fig. 5. Job description generation through the chatbot



```
rasa.core.test  - Evaluation Results on ACTION level:
rasa.core.test  -  Correct:          35 / 35
rasa.core.test  -  F1-Score:         1.000
rasa.core.test  -  Precision:        1.000
rasa.core.test  -  Accuracy:         1.000
```

Fig. 6. Accuracy of questions used to create a job description.

The test involved a meticulously designed set of 35 questions, strategically arranged in a sequential order to simulate a job-related inquiry. Remarkably, the chatbot demonstrated exceptional proficiency, as it accurately responded to all questions with a perfect accuracy score of 1.00.

b) *Predicting matching percentage of resume and job description*

In the initial attempts to match resumes with job descriptions using machine learning models, the achieved accuracies fell short of expectations. Consequently, a novel approach was used by developing a stacked ensemble model that combined multiple base models for both TD-IDF and BOW vectors. Through the implementation of hyperparameter tuning and advanced cross-validation techniques, remarkable performance were accomplished..

The models and their accuracies for both TD-IDF vectors and BOW are in the table below.

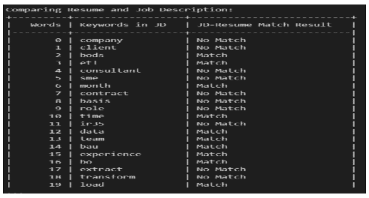
TABLE I. ACCURACIES OF THE STACKED MODELS

Model	Accuracy(Before Hyperparameters)	Accuracy (After Hyperparameters)
The stacked model with BOW Vectors	0.6035	0.8050
The stacked model with TF-IDF Vectors	0.3591	0.5173

TABLE II. HYPERPARAMETERS USED IN STACKED MODELS

Models	Hyperparameters at the Initial Stage	Hyperparameters After Tuning
Support Vector Regressor -linear	kernel=linear	C=1, epsilon=0.01,kernel='linear'
Random Forest Regressor	n_estimators =70	max_depth= 10, min_samples_split=3, n_estimators =70
Decision Tree Regressor	min_samples_split=10	min_samples_leaf=4, min_samples_split=10
Support Vector Regressor -RBF	kernel='rbf'	C=10, epsilon=1, gamma=0.1, kernel='rbf'
KNeighborsRegressor	n_neighbors=9	metric='manhattan', n_neighbors=9, p=1, weights='distance'

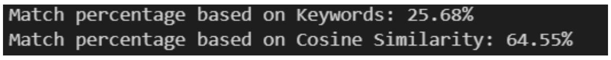
Furthermore, the resumes were matched with job descriptions based on keywords that persisted in them. The results were categorized as **'Match'** and **'No Match'**.



```
Comparing Resume and Job Description:
+----------+-----------------+-------------------------+
|   Words  | Keywords in JD  | JD-Resume Match Result  |
|----------+-----------------+-------------------------|
|        0 | company         | No Match                |
|        1 | client          | No Match                |
|        2 | bods            | Match                   |
|        3 | etl             | Match                   |
|        4 | consultant      | No Match                |
|        5 | sme             | No Match                |
|        6 | month           | Match                   |
|        7 | contract        | No Match                |
|        8 | basis           | No Match                |
|        9 | role            | No Match                |
|       10 | time            | Match                   |
|       11 | ir35            | No Match                |
|       12 | data            | Match                   |
|       13 | team            | Match                   |
|       14 | bau             | Match                   |
|       15 | experience      | Match                   |
|       16 | bo              | Match                   |
|       17 | extract         | No Match                |
|       18 | transform       | No Match                |
|       19 | load            | Match                   |
```

Fig. 7. Results of matching job description with resume.

As a result of this approach, matching percentages based on keywords and cosine similarity were obtained for each job description.



```
Match percentage based on Keywords: 25.68%
Match percentage based on Cosine Similarity: 64.55%
```

Fig. 8. Matching percentages

c) *Candidate Technical Skills Proficiency Summarization through Resume*

Applications of the technical skills scoring approach, which involves categorizing keywords found within a resume, yielded promising results in assessing professional qualifications. Here skill proficiency comparisons of candidates are graphical and present with categories with a keyword count. Different colors were used for each categorization.

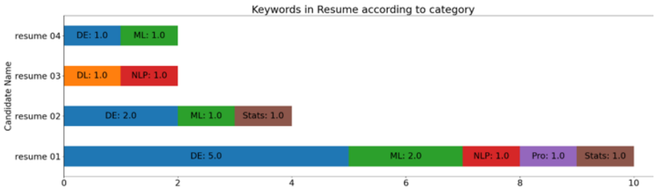
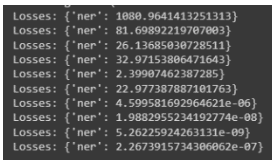


Fig. 9. Resume Skill proficiency with the job description

2) *Analyzing academic transcript for candidate's acquiredskills and knowledge.*

a) *Custom NER model*

The custom NER model was trained with 10 epochs and had a loss of 2.26739157343060620e-07.

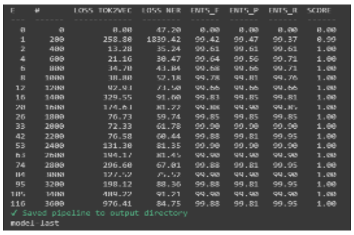


```
Losses: {'ner': 1080.9641413251313}
Losses: {'ner': 81.69892219707003}
Losses: {'ner': 26.13685030728511}
Losses: {'ner': 32.97153806471643}
Losses: {'ner': 2.39907462387285}
Losses: {'ner': 22.977387887101763}
Losses: {'ner': 4.599581692964621e-06}
Losses: {'ner': 1.9882955234192774e-08}
Losses: {'ner': 5.26225924263131e-09}
Losses: {'ner': 2.2673915734306062e-07}
```

Fig. 10. Custom NER losses

b) *Pre-Trained Custom NER model*

The pre-trained Named Entity Recognition (NER) model that was built using the Spacy library was trained with 200 epochs, 128 as the batch size, a learning rate of 0.001, and a loss of 84.75 with a score of 1.00.



```
E    #       LOSS TOK2VEC  LOSS NER  ENTS_F  ENTS_P  ENTS_R  SCORE
---  ------  ------------  --------  ------  ------  ------  ------
  0       0          0.00     47.20    0.00    0.00    0.00    0.00
  1     200        258.88   1830.42   99.42   99.47   99.37    0.99
  2     400         13.28     35.24   99.61   99.61   99.61    1.00
  4     600         21.16     30.47   99.64   99.56   99.71    1.00
  6     800         34.70     43.84   99.68   99.66   99.71    1.00
  8    1000         38.88     52.18   99.78   99.81   99.76    1.00
 12    1200         92.93     71.50   99.66   99.66   99.66    1.00
 16    1400        329.55     91.60   99.83   99.85   99.81    1.00
 20    1600        174.61     81.77   99.88   99.90   99.85    1.00
 26    1800         76.73     59.74   99.85   99.85   99.85    1.00
 33    2000         72.33     61.78   99.90   99.90   99.90    1.00
 42    2200         76.58     60.44   99.88   99.81   99.95    1.00
 53    2400        131.30     81.35   99.90   99.90   99.90    1.00
 63    2600        194.17     81.45   99.90   99.90   99.90    1.00
 74    2800        296.60     67.01   99.88   99.81   99.95    1.00
 84    3000        127.52     75.57   99.90   99.90   99.90    1.00
 95    3200        198.12     88.36   99.88   99.81   99.95    1.00
105    3400        489.22     91.71   99.90   99.90   99.90    1.00
116    3600        976.41     84.75   99.88   99.81   99.95    1.00
✓ Saved pipeline to output directory
model-last
```

Fig. 11. Pre-Trained Custom NER accuracy

Thus, the custom NER model was chosen over the pre-trained custom NER model as the loss that was obtained after running 10 epochs was low in the custom NER model.

c) *Skill area prediction*

The SVM model demonstrated the highest accuracy among the evaluated models, achieving an impressive accuracy rate of 0.8810. The Random Forest model closely followed with an accuracy of 0.8784, while the XGBoost model achieved an accuracy of 0.8623. Additionally, the Logistic Regression model yielded an accuracy of 0.7524, and the Naïve Bayes model achieved an accuracy of 0.7980. The hyper-parameters were tuned for a better result. Based on these results, the SVM model was determined to be the most effective and reliable model for skill area prediction.

TABLE III. ACCURACY TABLE OF THE SKILL AREA PREDICTION MODEL

Text Classification model	SVM	Random Forest	XGBooster	Naïve Bayes	Logistic Regression
Accuracy	0.8810	0.8784	0.8623	0.7980	0.7524

The below figures show the confusion matrix and the classification report which validates the model performance.

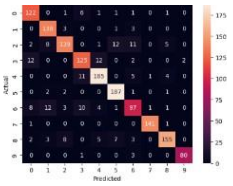
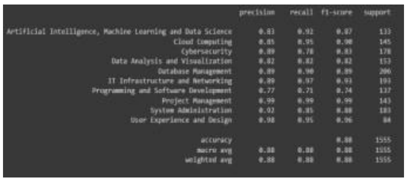


Fig. 12 : Confusion Matrix



```
                                                     precision    recall  f1-score   support

Artificial Intelligence, Machine Learning and Data Science       0.83      0.92      0.87       133
                                      Cloud Computing       0.85      0.95      0.90       145
                                        Cybersecurity       0.89      0.78      0.83       178
                       Data Analysis and Visualization       0.82      0.82      0.82       153
                                  Database Management       0.89      0.90      0.89       206
                      IT Infrastructure and Networking       0.89      0.97      0.93       193
                   Programming and Software Development       0.77      0.71      0.74       137
                                   Project Management       0.99      0.99      0.99       143
                                System Administration       0.92      0.85      0.88       183
                           User Experience and Design       0.98      0.95      0.96        84

                                             accuracy                           0.88      1555
                                            macro avg       0.88      0.88      0.88      1555
                                         weighted avg       0.88      0.88      0.88      1555
```

Fig. 13. Classification Report

After categorizing the skills of a candidate, they are visually presented using a graphical representation.

V. Conclusion and Future Work

The research centered around comprehensive automation of the recruitment process, catering to young professionals and recent graduates within the dynamic IT industry. The study aimed at automated processes that revolutionize hiring practices specifically designed for job market entrants. The proposed solution primarily focused on leveraging candidate resumes and academic transcripts to select the most suitable candidates for specific job positions. By combining these elements, organizations could gain a holistic understanding of candidates' qualifications, experiences, and skills. When analyzing resumes, the AI model became more capable of making detailed and personalized job descriptions. Although there were some limitations in accurately matching candidates in the IT industry at its current stage, the research showed promise in expanding the technology to work for candidates in all industries. In terms of future work in analyzing the academic transcript, currently designed for one specific university, the system could be expanded to analyze academic transcripts from all universities in Sri Lanka. These advancements would enhance the accuracy and depth of candidate assessment, contributing to more effective and tailored recruitment processes in the future.

VI. References

[1] "Sri Lanka aiming 200,000 ICT workforce by 2022," Aug. 08, 2019. https://www.icta.lk/news/sri-lanka-aiming-200000-ict-workforce-by-2022/

[2] A. Ghodasara, "11 Biggest Recruitment Challenges Faced by Recruiters in 2023." https://www.ismartrecruit.com/blog-recruitment-challenges-faced-by-recruiters#10.-implementing-data-driven-recruitment-11

[3] S. Karpe, "The Top Recruitment Challenges of IT Services Companies," Jun. 02, 2020. https://blog.imocha.io/the-top-recruitment-challenges-of-it-services-companies

[4] R. Ransing, A. Mohan, N. B. Emberi, and K. Mahavarkar, "Screening and Ranking Resumes using Stacked Model," in *2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, Mysuru, India: IEEE, Dec. 2021, pp. 643–648. doi: 10.1109/ICEECCOT52851.2021.9707977.

[5] M. Mehboob, M. S. Ali, S. Ul Islam, and S. Sarmad Ali, "Evaluating Automatic CV Shortlisting Tool For Job Recruitment Based On Machine Learning Techniques," in *2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*, Karachi, Pakistan: IEEE, Oct. 2022, pp. 1–4. doi: 10.1109/MAJICC56935.2022.9994112.

[6] M. Sharma, G. Choudhary, and S. Susan, "Resume Classification using Elite Bag-of-Words Approach," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India: IEEE, Jan. 2023, pp. 1409–1413. doi: 10.1109/ICSSIT55814.2023.10061036.

[7] R. G. U. S. Gajanayake, M. H. M. Hiras, P. I. N. Gunathunga, E. G. Janith Supun, A. Karunasenna, and P. Bandara, "Candidate Selection for the Interview using GitHub Profile and User Analysis for the Position of Software Engineer," in *2020 2nd International Conference on Advancements in Computing (ICAC)*, Malabe, Sri Lanka: IEEE, Dec. 2020, pp. 168–173. doi: 10.1109/ICAC51239.2020.9357279.

[8] T. Rahman, J. Nwokeji, R. Matovu, S. Frezza, H. Sugnanam, and A. Pisolkar, "Analyzing Competences in Software Testing: Combining Thematic Analysis with Natural Language Processing (NLP)," in *2021 IEEE Frontiers in Education Conference (FIE)*, Lincoln, NE, USA: IEEE, Oct. 2021, pp. 1–9. doi 10.1109/FIE49875.2021.9637220.

[9] "Introduction to Rasa," *Rasa*. https://rasa.com/docs/rasa-enterprise/

[10] "Perfect Fit," *Perfect Fit*. https://www.kaggle.com/datasets/mukund23/a-perfect-fit

[11] R. Day, "THE 10 MOST IMPORTANT IT SKILLS FOR 2020," Aug. 17, 2020. https://www.globalknowledge.com/us-en/resources/resource-library/articles/the-10-most-important-it-skills-for-2020/#gref

[12] L. Stevens-Huffman, "Top 10 IT Skills In-Demand for 2023," May 18, 2023. https://www.itcareerfinder.com/brain-food/blog/entry/top-10-it-skills-in-demand-for-2023.html