Robotic Automation Of Employee Onboarding Using Neural Computing

Sahil Sarthak Biswal, Ashwin Ganesh, Dr. P. Madhavan

Abstract: With routine, repetitive, labour-intensive tasks in the IT industry, there is a lot of human resources involved in handling these systems for business support and operations. The very fact that remains is if these mundane tasks were machine-driven, employees would be able to concentrate on higher-value activities with improved speed, productivity, at considerably reduced costs to the organisation. Robotic Process Automation (RPA) can do this by applying automation software to perform tasks and operations in applications and process them in the same way as a human would. It delivers direct profitability while improving accuracy across entire business functions and can be leveraged irrespective of industry and application. It is already having an impact at organizations currently deploying virtual workforces and delivered game-changing results for many organizations. The main objective of this proposed work is to bring a change in the employee onboarding process wherein the paperwork can be automated at a regular interval of time with hours of work can be saved. All the documents during the recruitment process can be a part of the onboarding documentation using automation that can be quickly finished with the probability of mistakes that might happen if performed manually can be avoided.

Keywords: Inverse Document Frequency, Natural Language Processing, Neural Computing, Onboarding of Employee, Robotic Process Automation, Similarity, Term Frequency, UiPath, Word2vec

1 INTRODUCTION

In the 1990s, companies began to their processes, or offshore and outsource some elements to take advantage of labour arbitrage whereas in the 2000s, as consistency and quality became more important, shared services went through a period of standardization, allowing businesses to focus on strategic pursuits. But now Companies are now exploring how automation and AI can cut up to 80% of the time needed to manage shared services transactions so employees can focus on value-adding activities. The present system had always been consistently a way to manage the HR operations of a business. With the advent of new technology, there was a need to produce software that handled the onboarding process of employees. Possessing the right onboarding software can change the established procedure and enhance productivity during the onboarding process by saving time and money. The sheer amount of time spent on paper-based operations and regulatory undertakings had prompted many HR managers to cautiously analyse the advantages of the product. The current legacy application has some shortcomings when it comes to this feature. Hence, there was a compulsion to mark this to make it user-friendly. Robotic Process Automation (RPA) is a proposal in which a 'robot' can take over standard and monotonous activities thatare currently offered by humans.

An RPA robot is a software that does not supplant existing HR or Payroll software. It is software that can take on human-related tasks. It is very useful in supporting data-driven processes. The primary outlook of the project corresponds to the fact that HR divisions in large organizations have an arduous task of processing records required for the onboarding of employees into the corporation and therefore seeks to replace this with software for automating paperwork at regular intervals of time with hours of work being liberated for other tasks. This automation of such procedures enhances the business analytics and allows for easy standardization of workflow to have a frictionless delivery of tasks. Growth in the automation domain will provide higher technical potential to dramatically reduce the risk of incorrect managerial reporting along with spontaneous analytics and higher accuracy of results. The project involves the deployment of RPA solutions for high volume, rule-based and repetitive tasks such as employee onboarding to reduce the complexities that needs to be checked every time.

2 RELATED WORK

To fulfil the objectives of this system and obtain an in-depth analysis of the methodologies used to bolster the scope and outlook of the concept of RPA in employee onboarding systems, an overview of related works based on automation technologies and neural networks were required for the development of the description of the particular system. This was done through careful analysis of the research work by Chaithra K & Vinod Kumar HP (2019) [1] which introduces the idea of RPA Automation to perform highly repetitive, mundane tasks so that HR can focus on its strategic and value-added work. The work by Audrey Bourgouin et al. (2018) [3] allowed the analysis of the RPA relevance of each process activity and the business process as a whole. The research proposed by Anusha N D et al. (2019) [2] presents how IA can be used to enhance systems to decide how to allocate effort in an organization and decrease operation costs and improve regulatory compliance. The enabling of RPA with traditional workflow practices through Henrik Leopold et al. (2018) [8] which employs supervised machine learning to identify textual process descriptions. It validates the process using a

- *Sahil Sarthak Biswal, B.Tech., Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, E-Mail: sahil.biswal111@gmail.com*
- *Ashwin Ganesh, B.Tech., Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, E-Mail: tag2998@gmail.com*
- *Dr. P. Madhavan, Associate Professor, Department of Computer Science, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, E-Mail: madhavap@srmist.edu.in*

collection of 424 activities from 47 textual process descriptions to see if the task is manual or automated. This enables it to consider the potential of leveraging RPA upon certain tasks. Solomiya Yatskiv et al. (2019) [4] put forward their work which allows for insights about how RPA for automation is investigated. Their experiments describe the proposed system that allows executing tests in a faster manner for software test automation with increased reliability. Ruchi Issac et al. (2018) [5] promote the relative study of the numerous industry-specific RPA tools like UIPath, Automation Anywhere and BluePrism based on the technical aspects and the implementation involved. KP Naveen Reddy and Undavalli Harichandana (2019) [6] and Anagnoste S. (2018) [7] provide works that are particularly helpful with the use of the knowledge gained and subsequent integration with modern technologies such as AI and text extraction. These works present some data related to automation for different business areas and provide relevant case studies that show how RPA is integrated with technologies. These allow for the establishment of key parameters for evaluation in those applications. Ammar Ismael Kadhim (2019) [10] has presented a contrast between TF-IDF and BM-25 for feature extraction utilized in term weighting. His work demonstrates the superiority of the latter technique which enhances the maximum value of F1-measure increased by 0.61 indicating the improved performance of feature extraction. Cai-zhi Liu et al. (2018) [9] propounded their work of text classification which utilizes the deep learning tool Word2vec and is used to get a vector representation of feature words with the enhanced TF-IDF algorithm being used to calculate the weights of the words. Haoying Wu,Na Yuan (2018) [11] have proposed an improved weighting algorithm that utilizes word frequency distribution and class distribution to augment the classification results. Their experimental results show that it can accurately reflect the differences between different text categories.

3 DATASET

For this system, we make use of three datasets, the primary of them being the StackExchange Network Posts which was required for training the word2vec model. It possessed a collection of 266529 types of words that were taken from a corpus of 84120716 raw words and 5720485 sentences. The testing dataset was a Kaggle dataset containing Job Descriptions (JDs) for several job openings with over 5000+ JDs including JDs for positions like Software developer, Web developer, Embedded Software Engineer, etc and was consequently savedThere was no open-source dataset for resumes found in text format. Indeed.com was the only site that displayed the resumes openly. Around 300 resumes of candidates for positions such as Data Scientist, Software Developer, Web Developer were collected. Finally, the resumes attached to the emails were also used for testing the model.

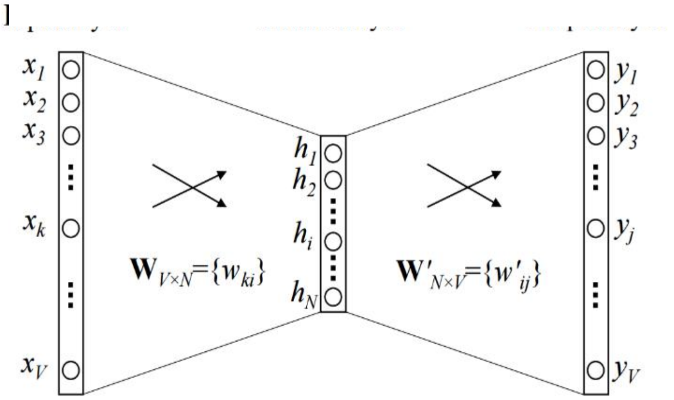
4 PROPOSED METHODOLOGY

The proposed methodology involves the deployment of RPA solutions for high volume, rule-based and repetitive tasks such as employee onboarding to reduce the complexities that need to be checked every time.

4.1 Preprocessing

Word2Vec refers to the predictive models used in the pre-processing stage consisting of two-layer neural networks used to restore linguistic contexts of words. A large corpus of text is fed as input to the deep-learning model resulting in a vector space of several dimensions. A vector in this space is a word in the corpus and the proximity of vectors depends on the context of the words in the corpus. The dimensionality of this dense vector space is a lot lower than those constructed using customary models. The CBOW (Count Bag Of Word) model is an architecture originally introduced by Mikolov et al. that can be utilized by Word2Vec to construct representations for word-embedding applications. The portrayals of these words are based on pairs of (context_window, target_word) where the model predicts the target_word depending on the context_window words. The Word2Vec models, based on unsupervised learning, can develop thick word-embeddings from the corpus without specifying additional information. However, a supervised learning methodology is utilized for classifying this corpus to get to the embeddings. Figure 1 shows a basic CBOW model with just one word in the context. The input is an encoded vector of size V. The hidden layer stores N neurons and the yield is a vector of size V again. W_{VxN} refers to the weight matrix mapping the input to the hidden layer (VxN dimensionality matrix). W_{NxV} refers to the weight matrix which maps the hidden layer to the output layer (NxV dimensionality matrix) These models are accessible through gensim and spaCy packages in Python and are trained over Google News data resulting in the model not knowing about the technically aware context distinction required for this model. Therefore, a dataset consisting of StackExchange network data was required which possessed the technical knowledge and had an adequate amount of unique words as new vocabulary could not be added to the model. The Posts.xml was used to extract each Post for each site from the stackexchange/ dataset. Irrespective of whether they were a Question or an Answer, they were extracted as HTML paragraph tags (<p>) and stored as paras.txt in the corresponding subfolder of the site.

Fig. 1. A basic CBOW model with just a single word


]
x_1
x_2
x_3
x_k
x_V
h_1
h_2
h_i
h_N
$\mathbf{W}_{V\times N}=\{w_{ki}\}$
$\mathbf{W'}_{N\times V}=\{w'_{ij}\}$
y_1
y_2
y_3
y_j
y_V

The training of the Word2Vec Model required an array of sentences to be streamed from the disk. This was done by using a list to represent the sentence and each word of the sentence acting as an element of this list. The sentences were extracted from the paras.txt files and saved into sentences.txt for each site by using BeautifulSoup. This was done to remove the formatting, mathematics, code from the final result. These sentences were then streamed into the Word2Vec train method for training the model.

4.2 Algorithm

Step 1:
The first task of the system is to identify and extract the resumes sent as an attachment to the authority responsible for the onboarding of employees in the organization. A single robot is created to enable this action to be performed sequentially for each iteration of mail from a list. A UiPath workflow is employed where the robot parses through the Gmail inbox of the employer and uses the GetIMAPMailMessages activity to retrieve the attachments.

Step 2:
A Python script is utilized to convert the document attached to a text file with spaCy, an industrial-strength Natural Language Processing tool being used to detect names in the file. Another workflow allows the robot to use RegEx for extracting the phone number and email id. A customized dataset is used by the robot in a separate workflow for extracting the technical skills into a batch file. The file created is merged with the main CSV file and the values are mapped to the latter.

Step 3:
The resumes are iterated line by line and the inconsequential lines are removed by the model. Each line is categorized into one of the sections defined by a collection of such headings. This is done by calculating its similarity to the existing words. This similarity is measured against a threshold based on which the section is updated and the point is marked or the previous section is continued. This enables good enough accuracy in separating the sections.

Step 4:
Each section of the resume is stored in a .csv file after removing the stop words and performing lemmatization to get a selected few keywords. For each keyword found, similar words and their corresponding similarity.

Step 5:
In order to calculate the score, the tf-idf weight is calculated. TF refers to Term Frequency that measures how often a word occurs in an archive.

$$TF(w) = \frac{Frequency\ of\ word\ w\ in\ a\ document}{Total\ number\ of\ words\ in\ the\ document} \quad (1)$$

IDF refers to Inverse Document Frequency which quantifies the significance of the word.

$$IDF(w) = log_e\left(\frac{Total\ number\ of\ documents}{Frequency\ of\ Documents\ with\ word\ w}\right)$$

(2)

Step 6:
The score of the resume is given by:

$$Score = \sum_{i=1}^{n} TF(w_i) * IDF\ (w_i) * Similarity(w_i)$$

(3)

where, w_i refers to the words present in the resume

Step 7:
Based on the score, the robot either enters the data into the employee database portal or sends a response back to the user stating that his resume is rejected.

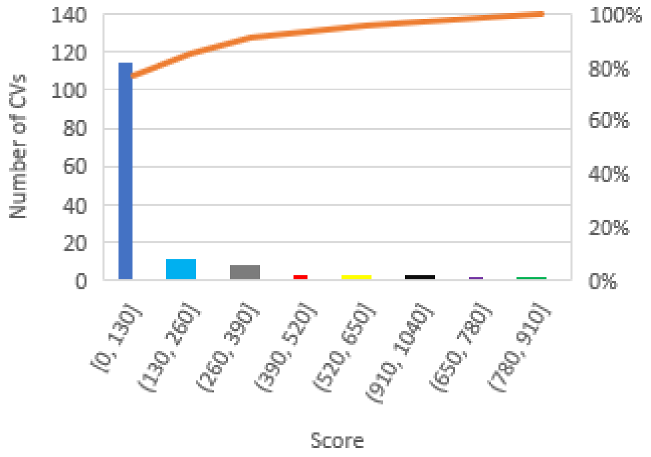


Fig. 2. Score distribution for different CVs

5 RESULTS AND DISCUSSION

This proposed model was built on UiPath and Python with packages like spaCy used to detect names and gensim used for the scoring algorithm. To assess the performance, the distribution of the scores of different resumes from Indeed.com was compared and analysed. This was done after fine-tuning the model to give the highest possible accuracy while testing.Figure 2 illustrates the distribution of scores for different resumes. The graph shows a peak for cv 109, 122 and 135, which contains the maximum number of words and possesses a high similarity for the job description required. These metrics showcase that these resumes performed better than the others. This finding indicates the supremacy of the training methodology over the sufficiently large dataset. The results verify the proposal that features learned by pre-trained models help to learn features for a completely different domain dataset.

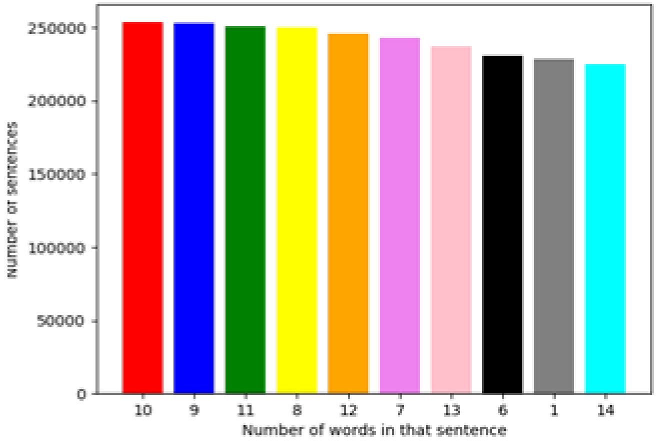


Fig. 3. Identification of proper context for word2vec model

Figure 3 showcases the frequency of words in a sentence to the number of sentences. It was observed that the mean of these values was at least 9 words in a sentence. It is used to identify the proper context for the word2vec training model as they capture semantic similarity. These generated dense word-embeddings in the vector space in the model.

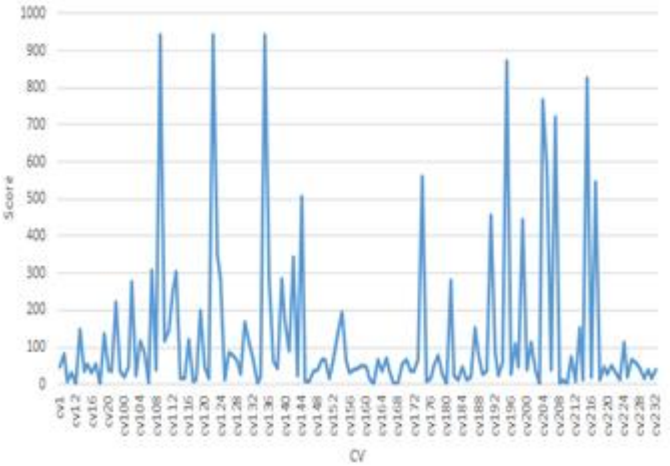


Fig. 4*. Categorizing different job titles based on description*

Figure 4 elucidates the classification of different job titles into 3 categories based on the description. The dataset retrieved from Kaggle shows a higher frequency of resumes for the designation of a web developer. The TF and IDF are statistical measures used here to evaluate how important a word is to a document in the corpus by assigning a particular weight to it.

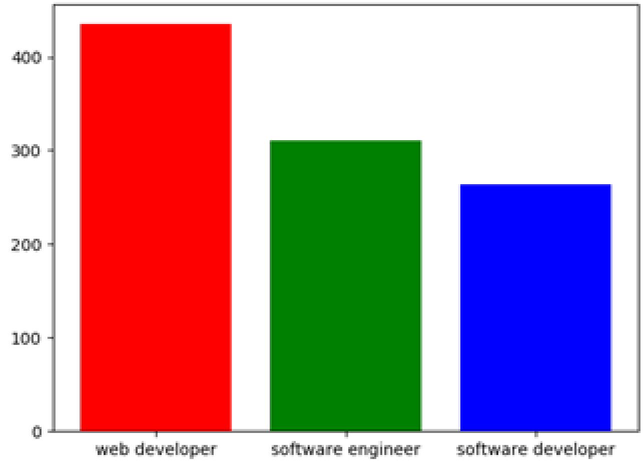


Fig. 5. *Number of CVs present in various score categories*

Figure 5 represents the number of CVs that fall in different score categories. It can be observed that most of the CVs taken as test data from Indeed.com fall under the category 0-130 score. Considering number of CVs that fall in the category of 0-130, there are 115 CVs. Considering the average of scores obtained by individual CVs, the trend can be devised.

TABLE 1
THE RESPONSE BY BOT BASED ON THE SCORE

Score for the CV	Result
Score < 40	Rejected by bot
Score>=40 and Score<130	Under Review – Send to User

Table 1 shows the action to be performed by the bot based on the score generated by the model. If the CV gets a score less than 40, then the bot will automatically reject the candidate’s CV. If the score is greater than or equal to 130, then the bot will accept the candidate’s CV. If the score lies in between the range of 40 and 130, then the bot will allow the user to choose whether to accept or reject the candidate’s CV after reviewing it.

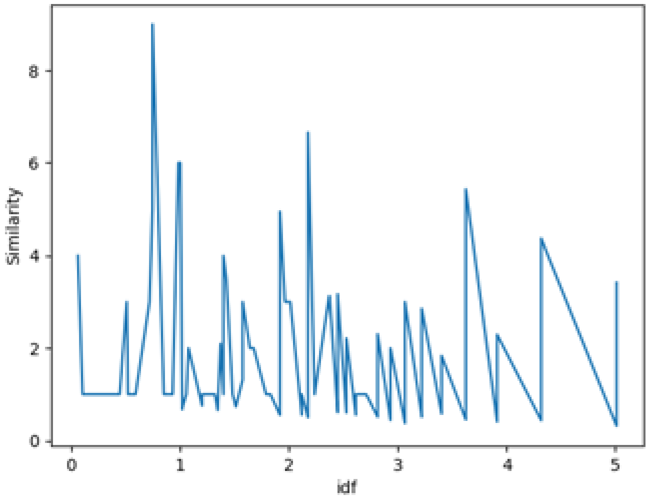


Fig. 6*. Relationship between idf and Similarity for different words*

Figure 6 tells us about the relationship between the idf of the word with respect to the Similarity of the word. According to the graph, there are few words that share the same similarity value for increasing idf value for the words. Also, the graph shows a high peak in similarity when the idf for a word is close to 1.

6 CONCLUSION

CV filtering has always been a subjective process, though the use of machine learning can reduce unnecessary human effort significantly. Owing to the fact that manual procedure is a complex and time-consuming task, it would be more prudent to use automation tools and procedures like UIPath which enables RPA thus saving time and improving productivity, accuracy, and consistency while avoiding human errors and reducing human efforts. Another conclusion drawn is the automation of such procedures enhances the business analytics and allows for easy standardization of workflow to have a frictionless delivery of tasks. RPA technology is going to have a high implementation and a greater scope soon and the odds of job opportunities in this field are growing high day by day. After all, who would like to waste time, money and the resources for a job that a single robotics automated software can do in a fraction of seconds.

VII. REFERENCES

[1] Chaithra K, Vinod Kumar H P, 2019, ‘Robotic Process Automation: Strategic Technology Solutions for IT’ NCARES 2019 Volume 7, Issue 10.

[2] Anusha N D, Baishali Rawat, Renuka J, Sahana S, Vijayshree H P, 2019,’RPA for Human Resource Operation’ International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 8 Issue 04.

[3] Audrey Bourgouin, Abderrahmane Leshob, and Laurent Renard, 2018, ‘Towards a Process Analysis Approach to Adopt Robotic Process Automation’ IEEE 15th International Conference on e-Business Engineering (ICEBE)

[4] Solomiya Yatskiv, Iryna Voytyuk, Nataliia Yatskiv, Oksana Kushnir, Yuliia Trufanova and Valentyna Panasyuk, 2019, ‘Improved Method of Software Automation Testing Based on Robotic Process Automation Technology’, 9th International Conference on Advanced Computer Information Technologies (ACIT)

[5] Ruchi Issac, Riya Muni and Kenali Desai, 2018, ‘Delineated Analysis of Robotic Process Automation Tools’, Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)

[6] KP Naveen Reddy and Undavalli Harichandana, 2019, ’A Study of Robotic Process Automation Among Artificial Intelligence’ International Journal of Scientific and Research Publications (IJSRP), volume 9, issue 2

[7] Anagnoste S., 2018, ‘Robotic Automation Process – The operating system for the digital enterprise’ Proceedings of the 12th International Conference on Business Excellence 2018, DOI:10.2478/picbe-2018-0007, pp. 54-69, ISSN 2558-9652

[8] Henrik Leopold, Han van der Aa, and Hajo A. Reijers (2018). Identifying Candidate Tasks for Robotic Process Automation in Textual Process Descriptions. 19th International Conference, BPMDS 2018, 23rd International Conference, EMMSAD 2018.

[9] Cai-zhi Liu,Yan-xiu Sheng ,Zhi-qiang Wei , Yong-Quan Yang,2018,’Research of Text Classification Based on Improved TF-IDF Algorithm’ , 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)

[10] Ammar Ismael Kadhim,2019, Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF, 2019 International Conference on Advanced Science and Engineering (ICOASE)

[11] Haoying Wu,Na Yuan,2018, An Improved TF-IDF algorithm based on word frequency distribution information and category distribution information, ICIIP '18: Proceedings of the 3rd International Conference on Intelligent Information Processing