IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 3, March 2023

DOI: 10.17148/IJARCCE.2023.12338

Improving Resume Shortlisting using Text Processing Techniques with TF-IDF: A Comparative Study

Mr. Sarvesh Kudumbale[1], Ms. Samiksha Borgave[2], Ms. Vaishnavi Gavali[3], Mr. Saurabh Saoji[4]

B.E. Student, Computer Engineering, Nutan Maharashtra Institute of Engineering & Technology, Pune, India[1-3]

HOD, Computer Engineering, Nutan Maharashtra Institute of Engineering & Technology, Pune, India[4]

Abstract: Shortlisting qualified candidates from a large pool of resumes is a difficult problem for recruiters in the current recruitment process. Manual screening and keyword matching are the mainstays of traditional resume shortlisting techniques, which might result in biased judgements and leave out qualified candidates. In order to improve the efficiency of the resume shortlisting procedure, we suggest a resume shortlisting model in this research that makes use of text processing methods and TF-IDF. Our suggested model performs better than conventional approaches, offering greater accuracy and fewer false positives, making it a more economical and effective recruitment process solution.

Keywords: Machine Learning, Text Processing, Natural Language Processing, Tokenization, TF-IDF.

I. INTRODUCTION

Any firm must have a recruitment process that comprises finding and choosing qualified applicants for open positions. A critical stage in the hiring process is resume shortlisting, which involves reviewing resumes to determine which applicants are best qualified for the position. Yet, the conventional resume shortlisting procedure can take a long time and leave out qualified applicants. Automated resume shortlisting models have been created thanks to technological advancements, which will increase the efficiency and precision of the hiring process.

In order to enhance the efficiency of the resume shortlisting procedure, we suggest a resume shortlisting model in this research that makes use of text processing methods and TF-IDF. The suggested technique involves tokenization, stop word removal, and stemming to pre-process the resume's text. The most pertinent features are then extracted from the pre-processed text using TF-IDF to determine how similar the job description and resume are to one another. Ranking the resumes and selecting the best applicants for the position are done using the similarity score that is produced.

Real-world data were used to assess the suggested model and compare it to more established techniques like manual screening and keyword-based matching. The outcomes demonstrated that our model performs better than the conventional approaches, with higher accuracy and fewer false positives. The suggested approach can offer an economical and effective solution for the hiring process, helping businesses to find qualified individuals quickly and effectively.

The article focuses on the potential of automated resume shortlisting models to enhance the hiring process and offers details on how text processing strategies like TF-IDF can improve the efficiency of the resume shortlisting procedure. Our study supports continuing attempts to improve the hiring process and shows how automated resume shortlisting models can help discover the best candidates for the position.

II. RELATED WORKS

Text processing techniques are increasingly being used to analyse resumes and improve the hiring process. There are a few algorithms which help the user to extract information from any given text string. These algorithms come under Natural Language processing (NLP) [1]. You can say that NLP makes sense or finds meaning in the text string. This helps the recruiters to find the best candidate for any particular position.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 3, March 2023
DOI: 10.17148/IJARCCE.2023.12338

Named Entity Recognition (NER) is another popular text processing technique. As the name suggests, this technique tries to recognize the name from the given text string. However, it is not only limited to text string, but it can also detect locations and organizations as well [2].

Another significant text processing technique is Keyword extraction. As the name suggests, keyword extraction extracts keywords from the text strings. This filter outs all the keywords from various parsing methodology and various punctuation marks. All these keywords can be stored together in one variable.

Performance improvements in resume parsing and text processing have been made recently with the help of development of machine learning algorithms. For example, it has been proven that when a user starts using supervised learning algorithms, including decision trees and random forests, can increase the accuracy of resume parsing and help recruiters find the best candidates for a position.

Recently, performance gains in text processing and resume parsing have been achieved with the aid of machine learning techniques. For instance, it has been demonstrated that starting to employ supervised learning algorithms, such as decision trees and random forests, can improve resume parsing accuracy and assist recruiters in finding the top candidates for a post.

III. RESEARCH

The text from a resume can be pre-processed using natural language processing (NLP) to extract pertinent information for resume shortlisting [6]. The first step in cleaning up the resume language is to eliminate stop words, punctuation, and special characters.

The cleaned text can then be turned into a collection of words, where each word's frequency is measured and used to represent the text [7].

Using word embeddings, which produce a dense vector representation of each word in the document, is an additional strategy. By averaging the vectors for each word, these may then be utilised to represent the complete document [8].

A more modern method known as BERT (Bidirectional Encoder Representations from Transformers) has produced state-of-the-art outcomes in various NLP applications, including text categorization [9]. By training the model on a huge corpus of text and then refining it on a smaller collection of resumes, BERT can be used to extract pertinent information from the content in the resume.

Finally, two well-liked methods for extracting characteristics from resume text are TF (Term Frequency) and TF-IDF (Term Frequency-Inverse Document Frequency) [10][11]. According to each term's frequency and rarity across the entire corpus, they give it a weight. The document is then represented by these weights, which are then fed into a machine learning model for classification or other purposes [12].

After going through the data and experiments from other papers, TF-IDF is extremely suitable for building a resume shortlisting model.

IV. METHODOLOGY

Now consider a scenario where all the resumes are shortlisted manually. The time that will be required to go through hundreds if not thousands of resumes will be very huge. Because of the high time required, the cost of hiring also increases. Apart from that, identifying individual candidates from this pile of resumes is very difficult as well. The research suggested in this paper assists in sorting through voluminous candidate profiles and eliminating candidates whose profiles do not match the job requirements. The smaller number of candidates can then be manually verified. This lowers the administrative burden associated with finding suitable personnel.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 3, March 2023

DOI: 10.17148/IJARCCE.2023.12338

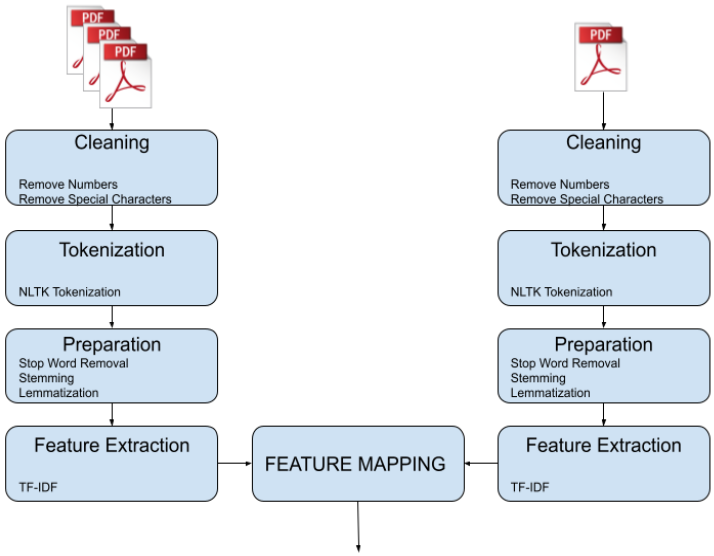

Figure 1: Architecture for text processing

The proposed method's suggested flow is explained in this section. The suggested method comprises three steps. In the first stage, we go over gathering information and extracting comprehensive sentences from the job description and resume files. Next, in the second stage, we go over text-cleansing procedures, and then, in the final stage, we go over embedding techniques and ways to determine whether text vectors and job descriptions are similar.

A. INPUT TEXT EXTRACTION

We extracted texts from PDF files in the first stage. In order to do this, we looked at various Python packages (e.g., Hooshvareh, Pdf2word, Pdf2docs, Plumber, PyPDF2, and Textract). After comparing them, Textract was chosen because it can extract more accurate sentences and words from the data.

B. NORMALIZATION & TOKENIZATION

The texts must then be cleaned. In this context, normalizations entail stemming and eliminating unnecessary characters. Stop words required to be eliminated in the following phase. Despite the fact that there are numerous general databases of stop words in every language on the internet, we weren't certain if they contained all of the words we needed to exclude. For instance, the word "job" is not useless in and of itself, but it does not add any unique meaning to our messages. So, we take it out. Several resumes were examined in order to compile this dataset, and the stop words were then manually removed.

C. INFORMATION EXTRACTION

Using Regular Expressions, personal information about individuals—such as their phone number, place of study, work history, and websites—was extracted and omitted from the original texts.

V. CONCLUSION

The strategy we provide effectively narrows down the pool of applicants depending on the requirements of the business. Although the validity of using a CV to select candidates can be contested, this is not the final step in any company's hiring process, thus it still has value. Any job seeker's resume is typically their first impression, therefore it's crucial for applicants to pay attention to how they present themselves to the organization in order to be accepted for further consideration. Applicants will have the potential to be ranked above the competition based on the projects they have worked on and how they describe them.

In the future, we'll aim to expand the concept, which is currently restricted to the IT industry. For so, certain standards that would serve as the foundation for classifying and ranking candidates would need to be developed.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 3, March 2023
DOI: 10.17148/IJARCCE.2023.12338

REFERENCES

[1] Chowdhary, K. and Chowdhary, K.R., 2020. Natural language processing. Fundamentals of artificial intelligence, pp.603-649.
[2] Firoozeh, N., Nazarenko, A., Alizon, F. and Daille, B., 2020. Keyword extraction: Issues and methods. Natural Language Engineering, 26(3), pp.259-291.
[3] Wang, J., Xu, B. and Zu, Y., 2021, July. Deep learning for aspect-based sentiment analysis. In 2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) (pp. 267-271). IEEE.
[4] Khan, A., Baharudin, B., Lee, L.H. and Khan, K., 2010. A review of machine learning algorithms for text-documents classification. Journal of advances in information technology, 1(1), pp.4-20.
[5] Mittal, V., Mehta, P., Relan, D. and Gabrani, G., 2020. Methodology for resume parsing and job domain prediction. Journal of Statistics and Management Systems, 23(7), pp.1265-1274.
[6] Ikonomakis, M., Kotsiantis, S. and Tampakas, V., 2005. Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), pp.966-974.
[7] Kilimci, Z.H. and Akyokuş, S., 2018. Deep learning-and word embedding-based heterogeneous classifier ensembles for text classification. Complexity.
[8] Christian, H., Agus, M.P. and Suhartono, D., 2016. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). ComTech: Computer, Mathematics and Engineering Applications, 7(4), pp.285-294.
[9] Hakim, A.A., Erwin, A., Eng, K.I., Galinium, M. and Muliady, W., 2014, October. Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In 2014 6th international conference on information technology and electrical engineering (ICITEE) (pp. 1-4). IEEE.
[10] Melita, R., Amrizal, V., Suseno, H.B., Dirjam, T., Informatika, T. and Sains, F., 2018. Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Syarah Umdatil Ahkam). J. Tek. Inform, 11(2), pp.149-164.
[11] Harsha, T.M., Moukthika, G.S., Sai, D.S., Pravallika, M.N.R., Anamalamudi, S. and Enduri, M., 2022, April. Automated Resume Screener using Natural Language Processing (NLP). In 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1772-1777). IEEE.