World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 5.924

RESUME SCREENING USING WEB SCRAPING, DEEP LEARNING AND NLP

Bhushan Kinge*[1], Shrinivas Mandhare[2], Pranali Chavan[3] and S. M. Chaware[4]

[1,2,3]UG Student, IT Department, RSCOE, SPPU, Pune.

[4]Professor, IT Department, RSCOE, SPPU, Pune.

Article Received on 01/10/2022 **Article Revised on 21/10/2022** **Article Accepted on 11/11/2022**

***Corresponding Author**
Bhushan Kinge
UG Student, IT Department, RSCOE, SPPU, Pune.

ABSTRACT

The Indian Recruitment market has grown substantially over the last half-decade as the need for cheap labor grows the number of job openings is increasing. And as the job market increases so does the recruitment industry which is a new way of hiring people by outsourcing the hiring process itself to other companies whose sole purpose is to give the correct talent required for the company. This is done because these companies are hiring in bulk and doing such a thing in-house will require a lot of company resources which will hamper productivity. As such companies emerge even for them manually going through all of the Resume of candidates is very time-consuming and tedious so these Talent Acquisition Companies use various ML models to filter out top resumes according to the job roles, which reduces the efforts for the HR team.

1 INTROZDUCTION

Machine learning is a field where we train a model with a dataset to predict the desired output when given new data. Screening the resumes is mostly done using Natural Language Processing (NLP), Natural language refers to the way we humans communicate with each other. NLP is concerned with giving computers the ability to understand the text and spoken words in much the same way human beings can. NLP combines computational linguistics-rule-based modeling of human language with statistical, machine learning, and deep learning models. Together combining these technologies helps computers process the way human language works in the form of texts or voice data and 'understand' its full meaning. As the

job market is growing in India, millions of new job seekers are joining the workforce every year, as per LinkedIn.[7] Around 1.3 million new jobs were created as per the 2021 Employees Provident Fund Organization (EPFO).[8] As of this year, the unemployment rate of India is around 7.74%.[6] where the urban area has an unemployment rate of 9.06% and the rural area is 7.13%. The number of job seats available is not enough to cover the staggering amount of applications the companies will receive.

Hence, if the companies hire in bulk there are many applications to find the talent that they need which will require a considerable amount of resources and time, this problem Talent acquisition Companies arise as solutions for this problem who fill in the spot and get the job done with less amount of resources costing to the company with an acceptable timeline. Even here the applications are in millions which is a tedious task to go through them hence these companies use various Machine learning models which will rank out the top resumes which are the best fit for the job role.

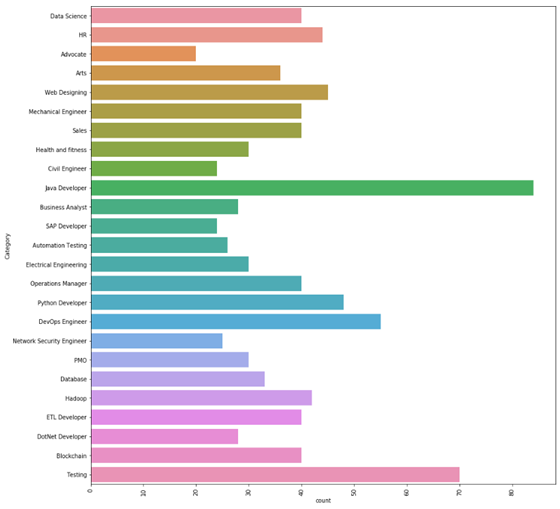
2 METHODOLOGY

The aim of the system is to predict the right job role for the given resume with the help of a trained deep learning model over the acquired dataset. The process will be divided into the following parts

1) Data Preprocessing
2) NLP pipeline
3) Training the Model
4) Testing

Dataset Description

The dataset is publicly available on Kaggle. The Dataset contains 3 fields ID, Category, and resume. There are a total of 25 different categories and a total of 962 observations.



2.1 Data Preprocessing

The Resume data is cleaned using the regular expression library in python.

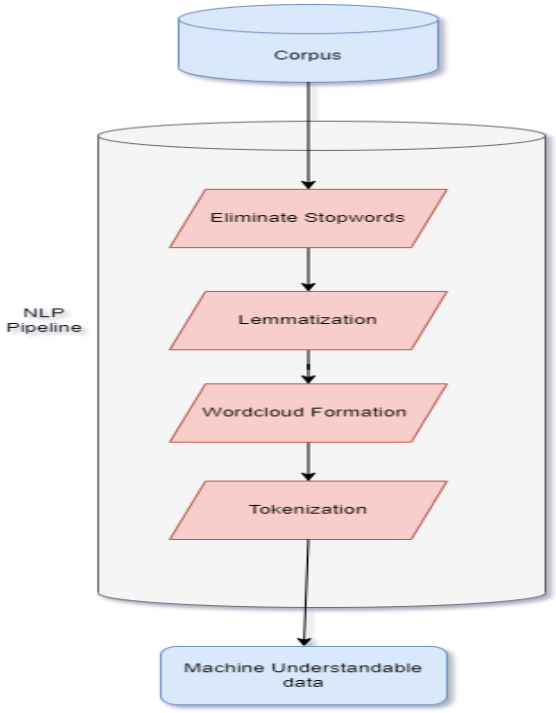
All the special characters such as @, $,%,#.& and * are removed, other characters like /,”,.,:,, = and _ is also removed in this process. Once the data is cleared of all these characters which will not add any strength to the model the processed data will be sent to the NLP pipeline

	Category	Resume	cleaned_resume
0	Data Science	Skills * Programming Languages: Python (pandas...	Skills Programming Languages Python pandas num...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	Education Details May 2013 to May 2017 B E UIT...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control System...
3	Data Science	Skills â¯ R â¯ Python â¯ SAP HANA â¯ Table...	Skills R Python SAP HANA Tableau SAP HANA SQL ...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	Education Details MCA YMCAUST Faridabad Haryan...

2.2 NLP Pipeline

Natural Language processing is basically a process where we help machines understand human language. As we all know machine’s language is binary it understands 0’s and 1’s. So to make the machine understand human sentences and words we use the NLP.

Below given is the NLP pipeline used.



When we extract the corpus from the dataset and the cleaning process is completed, the data is sent to the NLP pipeline to perform the following operations.

2.2.1 Eliminating the stop words

A stop word is a commonly used word like (the, and, or and so on). These words are not important for the model as they don't make any meaning and will contribute the least while prediction.

We use the NLTK library for the removal of stop words, below given are the stop words it consists.

{'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over', 'why', 'so', 'can', 'did', 'not', 'now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only', 'myself', 'which', 'those', 'i', 'after', 'few',

'whom', 't', 'being', 'if', 'theirs', 'my', 'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'}

2.2.2 Lemmatization

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meanings to one word.

Text preprocessing includes both Stemming as well as Lemmatization. Many times people find these two terms confusing. Some treat these two as the same. Actually, lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

2.2.3 Word Cloud Formation

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

2.2.4 Tokenization

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units is called a token.

The tokens could be words, numbers, or punctuation marks. In tokenization, smaller units are created by locating word boundaries.

These are the ending point of a word and the beginning of the next word**.**

2.3 Training Method

For the training purpose in deep learning, we have used Keras layers to form a neural network.

We've used 3 different layers.

2.3.1 Embedding layer

The embedding layer is one of the available layers in Keras. This is mainly used in Natural Language Processing related applications such as language modeling, but it can also be used with other tasks that involve neural networks. While dealing with NLP problems, we can use pre-trained word embeddings such as GloVe. Alternatively, we can also train our own embeddings using the Keras embedding layer.

2.3.2 Dense Layer

The dense layer is the regular deeply connected neural network layer. It is the most common and frequently used layer. The dense layer does the below operation on the input and returns the output.

2.3.3 Bidirectional LSTM

Bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification problems.

We've used 1 embedding layer and Bi-directional layer and 3 the dense layer, with the dense layer we use 2 different activation functions (SoftMax and relu).

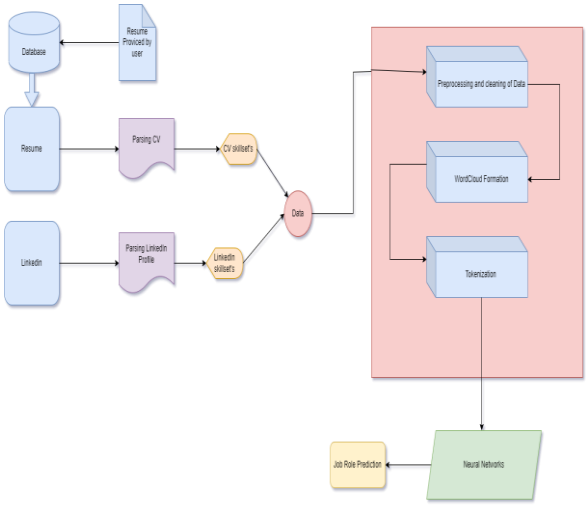
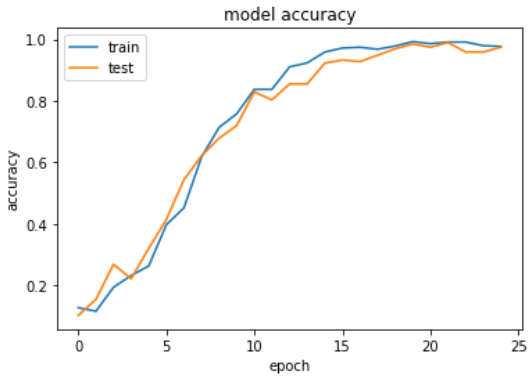


Fig:1

2.4 Testing

The dataset was distributed in the ratio of 80:20 before training the model. Once the model is trained the remaining 20% of the dataset is used to measure the accuracy of the trained model in our case the model is giving out accuracy of 97%.



CONCLUSION

This Paper deals with multiple methods to detect, identify and classify various resumes using multiple machine learning and Neural Network models. The accuracy of the models varies based on the datasets used the complexity of the learning methods and the size of the dataset, the results range from 78% - 98%. We conclude that with a proper dataset and the right algorithm we can get good accuracy and desired output for a large variety of purpose.

REFERENCES

1. Pradeep Kumar Roy, Vellore Institute of Technology. A Machine learning approach for automation of resume recommendation system, ICCIDS, 2019. 10.1016/j.procs. 2020.03.284.
2. Thimma Reddy Kalva, Utah State University, 2013. Skill-Finder: Automated Job-Resume Matching system.
3. Yong Luo, Nanyang Technological University, 2018. A Learning-Based Framework for automatic resume quality assessment, arXiv: 1810.02832v1 [cs.IR].
4. Suhjit Amin, Fr.Conceicao Rodrigues Institute of Technology, 2019. Web Application for Screening resume, IEEE DOI: 10.1109/ICNTE44896.2019.8945869
5. Tejaswini K, Umadevi V, Shashank M Kadiwal, Sanjay Revanna, Design and Development of Machine Learning based Resume Ranking System DOI: https://doi.org/10.1016/j.gltp.2021.10.002, 2021.
6. Riza tana Fareed, rajah V, and Sharadadevi kaganumath, "Resume Classification and Ranking using KNN and Cosine Similarity" In International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, 2021; 10.
7. Suhas Tangadle Gopalakrishna, Vijayaraghavan Varadharajan, "Automated Tool for Resume Classification Using Semantic Analysis", International Journal of Artificial Intelligence and Applications (IJAIA), January 2019; 10(1).
8. Suhas H E, Manjunath AE, "Differential Hiring using a Combination of NER and Word Embedding", In International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, 2020; 9.
9. Centre for Monitoring Indian Economy Pvt Ltd. (CMIE). The unemployment rate in India, 2022.
10. Howard, J.L., Ferris, G.R. The employment interview context: Social and situational influences on interviewer decisions 1. Journal of applied social psychology, 1996; 26: 112-136.
11. Mudit Kapoor, Business Today, India's formal job creation numbers beat pandemic blues, 2021.
12. M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," Journal of Machine Learning Research, 2006; 7: 2399–2434.

A. Zaroor, M. Maree, and M. Sabha, "A Hybrid Approach to Conceptual Classification and Ranking of Resumes" In Czarnowski I., Howlett R. (eds) Intelligent Decision Technologies. IDT. Smart Innovation, Systems and Technologies, 2017; 72.

13. Jabri, Siham, Azzeddine Dahbi, Taoufiq Gadi, and Abdelhak Bassir. "Ranking of text documents using TF-IDF weighting and association rules mining." In 2018 4th international conference on optimization and applications (ICOA), pp. 1-6. IEEE, 2018.
14. The data source for the skills used in the NER train.
15. Jagan Mohan Reddy D, Sirisha Regella., "Recruitment Prediction using Machine Learning", IEEE Xplore, 2020.
16. Resnick, P, Varian, H.R., Recommender Systems Communications of the ACM40, 1997; 56–59.
17. Xavier Schmitt, Sylvain Kubler, Jer my Robert, Mike Papadakis, Yves LeTraon University of Luxembourg, Luxembourg Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate.
18. Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring knowledge fragments for learning distance metric from a heterogeneous domain," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.
19. Mikheev, Andrei; Moens, Marc; Glover, Claire. "Named Entity Recognition without Gazetteers." Proceedings of EACL '99. HCRC Language Technology Group, University of Edinburgh. http://acl.ldc.upenn.edu/E/E99/E99-1001.pdf., 1999.
20. Zhou, GuoDong; Su, Jian. "Named Entity Recognition using an HMM-based Chunk Tagger." Proceedings of the Association for Computational Linguistics (ACL), Philadelphia, July 2002. Laboratories for Information Technology, 2002.
21. Zhang, L., Fei, W., Wang, L., Pjmatchingmodelofknowledgeworkers. Procediacomputerscience, 2015; 60: 1128–1137.
22. http://www.indeed.com/isp/apiinfo.jsp.