International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 03 | Mar 2022 www.irjet.net p-ISSN: 2395-0072

Review of HR Recruitment Shortlisting

Aniket Surve[1], Abhay Tiwari[2], Jinesh Van[3], Varsha Salunkhe[4]

[1], [2],[3]Student, Department of Computer Engineering, Atharva College of Engineering, Mumbai
[4]Professor, Department of Computer Engineering, Atharva College of Engineering, Mumbai

---

Abstract - *KNN is a highly efficient machine learning algorithm which is east to comprehend, simple in implementation and highly productive. It can be applied in problems of classification as well as regression making it a versatile algorithm to use. KNN classifies data points based on the distance of those points from 'K' data points which are already classified. Thus, the implementation is an example of supervised machine learning. It uses the entire dataset for classification and thus has high space complexity. But, its simplicity of use and versatile applications and the modern computational powers make it a highly efficient algorithm. To extract useful words related to the job profiles the concept of Natural Language Processing (NLP) is used. NLP deals with analyzing and understanding human language and making it more machine readable. Tokenization of words is used for converting English language words to more machine-readable formats like integers (indices). The useful set of words can be extracted from a document and converted to tokens and these tokens can be referred to whenever necessary using indexing. In this report, we have used NLP to extract keywords from the resume document and using this dataset to train a KNN algorithm which can classify between different set of profiles for a particular job role.*

Key Words: **Natural Language Processing, K Nearest Neighbors, Tokenization, Classification.**

1.INTRODUCTION

The system proposed in this report uses KNN algorithm, which is a supervised machine learning algorithm used for classification of job profile based on candidate's skills as mentioned in the resume document. Companies receive ample of resumes for a given job posting and for shortlisting they employ a dedicated screening officer who does the process of screening manually. The challenge is to classify the candidate as a right fit for a specific role. This challenge is magnified by the high volume of applicants if the business is labor-intensive, growing, and facing high attrition rates. IT departments has ever-growing market. In a typical corporation, professionals are hired for a variety of technical skills and other roles and assigned to different projects to address customer issues. This tedious task is referred to as resume screening process. Typically, large companies do not have enough time to open each CV, so they can use machine learning algorithms for the Resume Screening task.

2. Literature Survey

KNN based Machine Learning Approach for Text and Document Mining. Vishwanath Bijalwan, Pinki Kumari, et al. [1] Text Categorization (TC), also known as Text Classification, is the task of classifying documents based on content automatically. If a document has exactly one category, it is referred to as single label classification; otherwise, it is a multi-label classification. Text Categorization makes use of various tools from Information Retrieval and Machine Learning and has received much importance in the last years from researchers in the academia as well as industry developers.

Keyword Extraction: A review of method and approaches. Slobodan Beliga [2] Keyword Extraction (KE) is defined as a task that automatically identifies a set of all terms that best describes the subject of document [2]. A document often contains key words that represent relevant information: key phrases, key segments, key terms or just keywords. These represent most of the information the document contains. The extraction of these segments can result in faster computations and decreased computation time.

Work by Korsten (2003) et al. (2006) [3] According to Korsten (2003) and Jones et al. (2006), "HR Management theories" give significance on techniques of recruitment and selection and outline the advantages of interviews, tests and psychometric examinations as candidate selection process. They further stated that recruitment process may be internal or external or may also be conducted online. Typically, this process relies on layers of recruitment policies, job postings and, marketing, job application and interviewing process, assessment, decision-making, formal selection and training (2003). This paper is reviewed to gain domain knowledge for estimating the model performance more subjectively.

Techniques for text classification" by Jindal, Rajni et al. [4]. The main emphasis is laid on various steps involved in text classification process viz. document representation methods, feature selection methods, data mining methods and the evaluation methods to resolute on a given dataset. It focuses on preprocessing steps required for textual data. The raw text in itself has various attributes which are not useful for a model, such as articles, conjunctions, prefix and suffixes, etc. The paper dives into how textual information can be converted to a format suitable for solving classification problems.

Gongde Guo et al. [5] in their paper "KNN Model-Based Approach in Classification" they have presented a novel solution for dealing with the shortcomings of kNN. KNN can perform poorly is the data which is fed to the algorithm has uneven distribution of classes. To overcome this, the paper emphasizes on automatically selecting k values for different data and making classification faster.

Rahul S. Dudhabaware et al. [6] review on "Natural language processing tasks for text documents" focuses on deciding which NLP task will be better for preprocessing of search keyword, which in turn uses for appropriate matching to desired text documents. It touches through various important NLP tasks such as Part of Speech (POS) tagging, Named Entity Recognition (NER), Word Sense Disambiguation (WSD), sentiment analysis, etc.

Multinomial Naive Bayes Classification Model for Sentiment Analysis [7] Muhammad Abbas, Anees Ahmed, et al. This paper reviews how text categorization for documents can be achieved using multinomial naive bayes model. It focuses on the adaptation of simple Multinomial Naive Bayes that is used for text classification. The model described here overcomes the performance limitations of Bernoulli model. The paper states that MNB usually performs better on text classification problems such as topic categorization, spam detection, etc.

A Fast KNN Algorithm for Text Categorization [8] Yu Wang, Zheng Ou Wang. This paper focuses on how using KNN algorithm for text categorization for large samples can have a fatal effect on time complexity. It focuses mainly on how K neighbors can be searched quickly for large sample population. The algorithm reviewed here decreases the time for computing K neighbors. The method described here makes use of Tree Fast KNN abbreviated as TFKNN, which uses a tree data structure to organize the data points according to the distances of neighbors.

KNN with TF-IDF based Framework for Text Categorization [9] Bruno Trstenjak et al. This paper presents how KNN algorithm integrated with TF-IDF vectorizer can be effectively used for text classification purposes. It emphasizes on using KNN to group similar texts for classification of different documents or text corpus. It explains in brief about the TF-IDF method which is used to compute the weights of texts in comparison to the entire corpus of documents. Then it combines the TF-IDF method together with KNN algorithm to classify textual information.

CATEGORIZATION USING k-NEAREST NEIGHBOR CLASSIFICATION [10] Gulen Toker, Oznur Kirmemis. The concepts presented in this paper focuses on how documents can be classified using a simple and effective machine learning algorithm called KNN algorithm. KNN is widely used for classifying data points based on their proximity with each other. It classifies data based on the similarity of variables. It a very intuitive algorithm which makes it easier and very quick to implement.

An Improved k-Nearest Neighbor Algorithm for Text Categorization [11] Baoli Li, Q. Lu, et al. Although, KNN is a very effective algorithm for classification problems, it could lead to poor performance if the value of parameter "K" is not selected properly. This paper focuses on how to improve the performance of KNN algorithms by selecting value of K that satisfies the performance levels of an effective model. It presents concepts of overfitting and underfitting that results due to poor selection of K values for KNN algorithm.

A re-examination of text categorization methods [12] Yiming Yang, Xin Liu. This paper focuses on different learning algorithms that perform poorly on skewed or unevenly distributed data. It then shows how Support Vector Machines (SVM), K Nearest Neighbors (KNN), and Linear Least Squares Fit (LLSF) perform significantly better for data with skewed classes. It emphasizes the differences in model performances for unevenly distributed data for different learning algorithms.

Confusion Matrix [13] K. Ting. A confusion matrix is used to obtain information of how learning model performs for a classification problem. It contains the numbers for correct predictions for a positive class, incorrect predictions for a positive class, correct predictions for negative class, and incorrect predictions for a negative class. Using confusion matrix, we can fine tune the model to improve performance. It can be used for binary as well as multi-class classification problems.

3. CONCLUSIONS

Resume defines the chance of selection a candidate holds for a particular job profile. If a candidate wants to improve the resume, they can try out different resumes using the algorithm and use one with the optimal results. Further, companies can save ample amount of time and effort using classification model built using the concept of KNN algorithm, to shortlist candidates that fit better for the job and who can provide with all the necessary skill sets required for the position Thus, KNN algorithm is used to build the model for making this project a success and implementing the test set with a high accuracy score and one which generalizes well on data which is unseen.

ACKNOWLEDGEMENT

We owe a sincere thanks to our college Atharva College of Engineering, especially our Head of Department, Dr. Suvarna Pansambal, our guide, Prof. Varsha Salunkhe for their kind cooperation and guidance which helped us in the completion of this project which would have seemed difficult without their motivation, constant support and valuable suggestions. Moreover, the completion of this research would have been impossible without the cooperation, suggestions and help of our family and friends.

REFERENCES

[1] Bijalwan, Vishwanath & Kumari, Pinki & Espada, Jordán & Semwal, Vijay & Kumar, Vinay. KNN based Machine Learning Approach for Text and Document Mining. International Journal of Database Theory and Application. (2014).

[2] Beliga, Slobodan. "Keyword extraction: a review of methods and approaches." (2014).

[3] Work by Korsten (2003) and Jones (2006) et al., "A systematic review of literature on recruitment and selection process", Humanities and Social Sciences Reviews, (2019).

[4] Jindal Rajni, Ruchika Malhotra and Abha Jain. "Techniques for text classification: Literature review and current trends." Webology 12 (2015).

[5] Guo G., Wang H., Bell D., Bi Y., Greer K. (2003) KNN Model-Based Approach in Classification. In: Meersman R., Tari Z., Schmidt D.C. (eds) On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003.

[6] R. S. Dudhabaware and M. S. Madankar, "Review on natural language processing tasks for text documents," 2014 IEEE International Conference on Computational Intelligence and Computing Research, 2014.

[7] Abbas, Muhammad & Ali, Kamran & Memon, Saleem & Jamali, Abdul & Memon, Saleemullah & Ahmed, Anees., Multinomial Naive Bayes Classification Model for Sentiment Analysis (2019).

[8] Y. Wang and Z. -O. Wang, "A Fast KNN Algorithm for Text Categorization," 2007 International Conference on Machine Learning and Cybernetics, 2007.

[9] Bruno Trstenjak, Sasa Mikac, Dzenana Donko, "KNN with TF-IDF based Framework for Text Categorization", Procedia Engineering, Volume 69, 2014.

[10] Toker, Gülen and Öznur Kırmemiş. "CATEGORIZATION USING k-NEAREST NEIGHBOR CLASSIFICATION." (2017). https://semanticscholar.org/paper/

[11] Li, B., Yu, S., and Lu, Q., "An Improved k-Nearest Neighbor Algorithm for Text Categorization", arXiv e-prints, 2003.

[12] Yiming Yang and Xin Liu. 1999, "A re-examination of text categorization methods". In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). Association for Computing Machinery, New York, NY, USA, 42–4.

[13] Ting, Kai Ming. "Confusion Matrix." Encyclopedia of Machine Learning (2010). https://semanticscholar.org/paper