2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE 2024)

2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE) | 979-8-3503-8328-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICWITE59797.2024.10502523

Resume Ranking and Shortlisting with DistilBERT and XLM

Anuska Mukherjee
Department of Statistics and Data Science CHRIST (Deemed to be University)
Karnataka, India
anuskamukherjee2001@gmail.com

Umme Salma M
Department of Statistics and Data Science CHRIST (Deemed to be University)
Karnataka, India
ummesalma.m@christuniversity.in

***Abstract*—The research presented in this paper offers a solution to the time-consuming task of manual recruitment process in the field of human resources (HR). Screening resumes is a challenging and crucial responsibility for HR personnel. A single job opening can attract hundreds of applications. HR employees invest additional time in the candidate selection process to identify the most suitable candidate for the position. Shortlisting the best candidates and selecting the appropriate individual for the job can be difficult and time-consuming. The proposed study aims to streamline the process by identifying candidates who closely match the job requirements based on the skills listed in their resumes. Since it is an automated process, the candidate's individual preferences and soft skills remain unaffected by the hiring process. We leverage advanced Natural Language Processing (NLP) models to improve the recruitment process. Specifically, our emphasis lies in the utilization of the distilBERT model and the XLM (Cross-lingual Language Model). This paper explores the application of these two models in taking hundreds of resumes for the job as input and providing the ranked resumes fit for the job as output. To refine our approach further, two types of metrics for resume ranking, such as Cosine similarity score and Spatial Euclidean distance, are used, and the results are compared. Intriguingly, distilBERT and XLM result in different sets of top ten ranked resumes, highlighting the nuanced variations in their ranking approaches.**

***Keywords*—Natural Language Processing, Automatic Recruitment Process, Resume Ranking, distilBERT, XLM**

I. Introduction

Recruiting candidates for a job post is one of the most difficult tasks for human resource (HR) departments and job websites. Any job vacancy posted on the job portals receives an enormous volume of resumes, which often come in diverse and unstructured formats such as .doc, .pdf, and .rtf. Moreover, efficiently screening resumes requires the expertise of subject specialists to evaluate how well a candidate's profile aligns with a given position. Recruiters need to swiftly eliminate irrelevant profiles in the initial stages of resume screening, in order to save time and resources. Amidst the challenges posed by the current job market, the need for effective recruitment strategies is more critical than ever. Thus, many companies are now opting for automated online recruitment systems to streamline candidate selection, seeking to cut down on expenses, labour, and the time needed for sorting through resumes. Online recruiting systems use several approaches that have been established in the literature.

These techniques include machine learning methods, models based on Relevance Feed- back, Boolean Retrieval technique, Analytic Hierarchy Process and many more. Though these techniques bring good results, they fall short in capturing the semantic aspect in the process of matching resumes and JD. Semantic similarity is a feature of Natural Language Processing that enables us to find comparable pieces of text even when the sentences contain distinct words. Existing methodologies, based on term frequency and keyword matching, exhibit shortcomings when faced with phrasing variations and subtle differences in language. Sometimes, the JD demands some skill or qualification, and the candidate has the same skill or qualification written in their resume but phrased differently.

To address this issue, this study aims to bridge the existing gap in resume ranking methodologies by focusing on semantic similarity computation, offering a more nuanced and context-aware assessment of resumes against JDs. The primary objective includes implementing semantic-based similarity with the help of advanced language model, BERT (Bidirectional Encoder Representations from Transformers). We specifically employ two language models for resume ranking: distilBERT and XLM. Previous language models have only been able to interpret text input sequentially, i.e. from left to right or right to left, but not concurrently. BERT is unique since it is designed to read simultaneously in both directions. DistilBERT, a distilled version of BERT, retains the fundamental architecture of BERT while significantly reducing its size, making it computationally lighter and more suitable for resource - constrained applications. Moreover, as organizations increasingly operate in diverse linguistic environments, the demand for cross-lingual resume processing has become imperative. XLM is specifically designed to capture semantic relationships in text across multiple languages, making it a valuable asset in the development of inclusive and globally applicable resume screening systems.

The following sections delve into the methodologies employed, the dataset used for experimentation, and the evaluation metrics applied. The findings and their implications are discussed, followed by a conclusion that summarizes the contributions of the study and outlines potential future scopes.

II. Related Works

Similar to a recommender system, the resume matching process matches the candidate's profile with the job description for a certain position. The methodologies of recommender systems were first introduced by Resnick and Varian [1]. In literature, recommendation systems have been widely employed in many other fields, such as e-commerce portal product recommendations [2], news suggestions [3], personalized book recommendations [4], curated movie recommendations [5], and music recommendations [6], among many others.

The various kinds of recommendation algorithms and their workings were thoroughly covered by Wei et al. [7]. Otaibi et al. [8] conducted a thorough investigation into the use of employment referral services and discussed the precautions any organisation should take during the hiring process. An Expectation-Maximization (EM) algorithm was used by Malinowski et al. [9] to generate employment suggestions, taking into account the job description provided by the organisation as well as the candidate's resume. A fuzzy-based approach was proposed by Golec and Kahya [10] to assess a candidate's relevance to the job description. Roy et al. [11] and Tejaswini et al. [12] used Tf-Idf vectorizer to process the collected resumes and used cosine similarity score and KNN to rank the candidates' resumes.

As technology progressed, researchers have turned to advanced NLP models to enhance the accuracy and efficiency of resume screening. Transformer models, introduced by Vaswani et al. [13], have become the backbone of recent advancements in NLP tasks. These models, based on attention mechanisms, enable the capture of long-range dependencies and semantic relationships in text. DistilBERT, a distilled version of BERT (Debut et al. [14]), and XLM (Lample et al., [15]), a model designed for cross-lingual language understanding, have emerged as powerful choices for various NLP applications. Recently James et al. [16] suggested using Sentence-BERT (SBERT), which generates semantically significant sentence embeddings, to shortlist and rank resumes. They have also compared performance of SBERT with BERT and showed that SBERT shows superior results in this domain.

Our approach differs from that of previous proposed mechanisms since most of the current systems recommend candidate resumes to HR personnel based on the content of their resumes, which results in low classification accuracy. To make it better, we suggested a two-step process that involves categorising resumes into relevant categories and then evaluating the contents of each resume according to the job description. Further, we used distilBERT and XLM. With its lighter architecture, distilBERT offers computational efficiency without compromising on performance. The distilled nature of DistilBERT makes it particularly suitable for real-time resume screening applications. Cross-lingual resume processing is essential in a globalized job market. Thus, by learning representations that are invariant across languages, XLM facilitates the development of multilingual resume screening systems, ensuring fair and inclusive hiring practices.

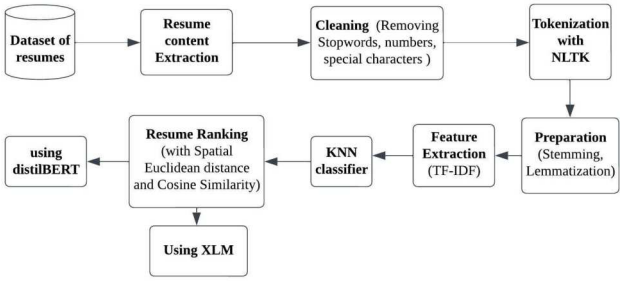

Fig. 1. A complete framework of the proposed method

III. Proposed Methodology

Our proposed method consists of two phases: first, classifying different resumes in the correct job category and then evaluating the content of resumes using distilBERT and XLM, and ranking the resumes based on a sample JD, with the help of cosine similarity score and spatial Euclidean distance. The complete framework of our work is shown in Fig. 1

A. Dataset Description

The dataset consists of candidate ID, their job category, and their resumes. Most of the resumes from different job categories were obtained from Kaggle in .csv format. However, for job positions like data analysts and data scientists, resumes were collected from senior students in our department, obtained in .pdf format, and later extracted into a csv file. Merging all of these available resumes, we have data of a total 1008 candidates in 26 job categories. Since the extracted resumes are not fit for further analysis, our next step would be to clean them.

B. Data pre-processing

The resumes that are being submitted as input are cleaned during this procedure to get rid of any unusual or unnecessary characters. Digits, special characters such as hashtags and mentions, and single-letter words are all eliminated, and contact numbers and email addresses are masked during cleaning. Additionally, pre-processing included stop word removal, stemming, and lemmatization. After completing these procedures, we had a clean dataset devoid of single letters, digits, or special characters. The resumes and JD are then tokenized using the NLTK library [17].

C. Classification of job categories

After resume cleaning, we move to the job category classification. Classification of job categories was done using k- nearest neighbour algorithm. KNN is a non-parametric supervised learning classifier. This algorithm identifies or predicts how a single data item will be categorised based on proximity. Although it can be used to solve both regression and classification problems, it is most commonly used as a

classification technique based on the premise that comparable points can be adjacent.

D. Resume Ranking Metrics

This study used two ranking metrics that yielded optimal results: cosine similarity and spatial Euclidean distance.

1) *Cosine Similarity:* Cosine similarity serves as a metric to gauge the similarity between two non-zero vectors within an inner product space. Specifically, it quantifies the cosine of the angle formed between the vectors, akin to the inner product of those vectors, after normalization to ensure equal length. To rank resumes based on cosine similarity score, we need two vectors $\vec{a}$ and $\vec{b}$,where $\vec{a}$ is the vector of words present in the resume and $\vec{b}$ is the vector of words present in the sample JD we shall consider for our analysis.

The similarity score between these two vectors is defined as

$$s(\vec{a}, \vec{b}) = \frac{\vec{a} . \vec{b}}{|\vec{a}| * |\vec{b}|} \quad (1)$$

2) *Spatial Euclidean distance:* The Euclidean distance between two points in Euclidean space is the length of a line segment connecting those two points. Spatial Euclidean distance, in the context of our research, calculates the Euclidean distance between two vectors. The Euclidean distance between vectors $\vec{a}$ and $\vec{b}$ is defined as

$$||a - b||_2 = (\sum (a_i - b_i)^2)^{\frac{1}{2}} \quad (2)$$

E. Transformer Models

1) *distilBERT:* The BERT model, which was quite revolutionary in the NLP domain, was first released in 2018 by GoogleAI researchers. However, the BERT model had some drawbacks. Its substantial size led to operational challenges, including slower processing speeds. In response to these limitations, researchers from Hugging Face launched distilBERT [14], which is a distilled form of the BERT model. DistilBERT offers a notable advantage over its predecessor by reducing its size by 40% through knowledge distillation during the pre-training phase while it could retain 97% of its language understanding abilities, and it is 60%faster compared to BERT.

2) *XLM:* XLM is a transformer-based model [15] trained in several languages. XLM was pre-trained through a diverse range of language modeling objectives, including Masked Language Modeling, akin to the approach employed by BERT; Causal Language Modeling, where the model predicts the likelihood of a word given the previous words in a sentence; and Translation Language Modeling, a new translation language modelling objective for improving cross-lingual pre-training.

IV. Results and Discussion

There are 26 different categories of jobs present in our data. Fig.2 suggests that job categories such as Java developer, data analyst, and data scientist have a significant amount of resumes under them. For classifying resumes into correct job categories, we split the whole data into train-test sets and used KNN classification algorithm.

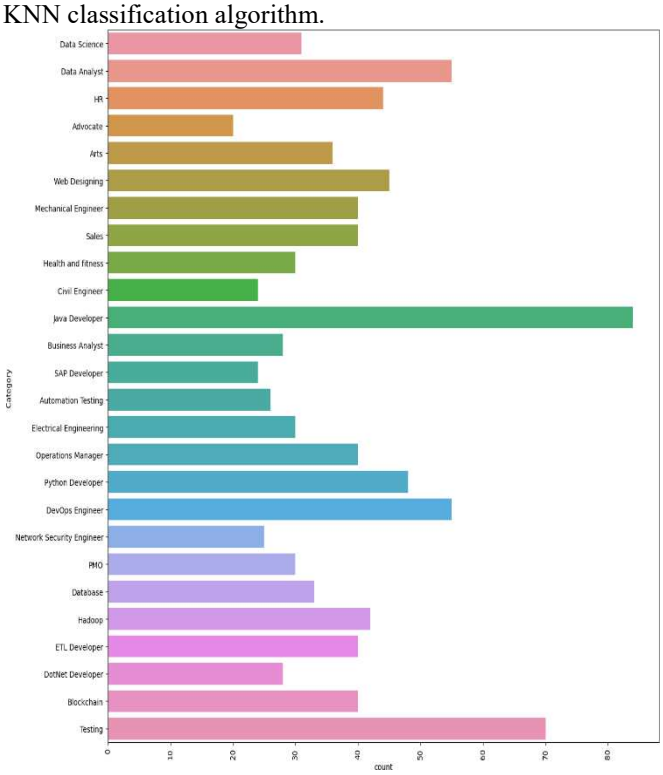

Fig. 2 Barplot of Job Categories

Among different train-test split ratios, the 80-20 train-test split produced the optimal result, yielding a 98% train accuracy and 95% test accuracy. This suggests that the model performed well on both the training and testing datasets. The small difference in the training and testing accuracy indicates that the model does not overfit the training data excessively. Next, the sample JD is matched against the content of resumes in our dataset and the top ten matching resumes are recommended using the following four ways. With Cosine similarity score as a metric to rank resumes, the less the angle between resume vector and JD vector and hence higher the similarity score, higher is the rank. With Spatial Euclidean distance as a metric, the less the Euclidean distance between resume and JD, higher the rank. distilBERT, the model successfully identified distinct candidates.

Table. III and Table. IV shows ranked resumes based on Spatial Euclidean distance using distlBERT and XLM, respectively. XLM model recommends a different set of top ten candidates than distilBERT for both cosine similarity and spatial Euclidean distance. This is due to the fact that distilBERT and XLM have different architectures. XLM (Cross- Lingual Language Model) is explicitly designed to understand multiple languages, whereas distilBERT is often used for English-focused tasks. If the dataset contains resumes in different languages or has a diverse set of writing

styles, XLM might capture nuances that distilBERT doesn't, and vice versa. Differences in architecture can lead to variations in how they capture and represent information.

TABLE I
RANKING BASED ON COSINE SIMILARITY SCORE USING DISTILBERT

Rank	Candidate ID	Similarity score
1	10	0.886616
2	20	0.886616
3	30	0.886616
4	40	0.886616
5	4	0.850444
6	14	0.850444
7	24	0.850444
8	34	0.850444
9	7	0.849833
10	17	0.849833

TABLE II
RANKING BASED ON COSINE SIMILARITY SCORE USING XLM

Rank	Candidate ID	Similarity score
1	1008	0.774167
2	744	0.751922
3	751	0.751922
4	758	0.751922
5	765	0.751922
6	772	0.751922
7	779	0.751922
8	680	0.746774
9	683	0.746774
10	686	0.746774

TABLE III
RANKING BASED ON SPATIAL EUCLIDEAN DISTANCE USING DISTILBERT

Rank	Candidate ID	Distance
1	40	5.1426
2	20	5.1426
3	30	5.1426
4	10	5.1426
5	27	5.8738
6	17	5.8738
7	7	5.8738
8	37	5.8738
9	34	6.0238
10	14	6.0238

TABLE IV
RANKING BASED ON SPATIAL EUCLIDEAN DISTANCE USING XLM

Rank	Candidate ID	Distance
1	1008	8.9325
2	744	9.6251
3	779	9.6251
4	751	9.6251
5	772	9.6251
6	758	9.6251
7	765	9.6251
8	708	9.7702
9	702	9.7702
10	669	9.7702

V. CONCLUSION AND FUTURE WORK

In this research article, inefficient manual screening of resumes is replaced with the Automated Resume Screening System, backed by NLP and machine learning techniques. A foundation for leveraging transformer models, distilBERT and XLM, in the context of resume ranking is provided. Employers and recruiters can leverage these models based on their specific preferences—whether prioritizing candidates with high cosine similarity or considering geometric relationships through spatial Euclidean distance. Again, the choice between distilBERT and XLM depends on several factors. If the dataset includes resumes in multiple languages, XLM might be more suitable. On the other hand, if computational resources are a significant concern, and we still want a powerful language model, DistilBERT might be a good choice, since distilBERT is more computationally efficient version of BERT.

In future work, the models can be evaluated on a larger and more diverse dataset. Further exploration into fine-tuning models and incorporating additional features may enhance the precision and adaptability of the ranking system.

REFERENCES

[1] P. Resnick and H.R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56-58, Mar. 1997.

[2] J.B. Schafer, J. Konstan, and J. Reidl, "Recommender systems in e-commerce," in *Proceedings of the 1st ACM conference on Electronic commerce*, Nov 1999, pp. 158-166.

[3] A.S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in *Proceedings of the 16th international conference on World Wide Web*, May 2007, pp. 271-280.

[4] R.J. Mooney, and L. Roy, "Content-based book recommending using learning for text categorization," in *Proceedings of the fifth ACM conference on Digital libraries*, Jun 2000, pp. 195-204.

[5] Q. Diao, M. Qiu, C.Y. Wu, A.J. Smola, J. Jiang, and C. Wang, "Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS)," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug 2014, pp. 193-202.

[6] O. Celma, "Music recommendation," *Music recommendation and discovery: The long tail, long fail, and long play in the digital music space* , pp. 43-85, Jun 2010.

[7] K. Wei, J. Huang, and S. Fu, "A survey of e-commerce recommender systems," in *2007 international conference on service systems and service management*, Jun 2007, pp. 1-5.

[8] S.T. Al-Otaibi, and M. Ykhlef, "A survey of job recommender systems," *International Journal of the Physical Sciences*, vol. 7, no. 29, pp. 5127-5142, Jul 2012.

[9] J. Malinowski, T. Keim, O. Wendt, and T. Weitzel, "Matching people and jobs: A bilateral recommendation approach," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Jan 2006, pp. 137c-137c.

[10] A. Golec, and E. Kahya, "A fuzzy model for competency-based employee evaluation and selection", *Computers and Industrial Engineering*, vol. 52, no. 1, pp. 143-161, Feb 2007.

[11] P.K. Roy, S.S. Chowdhary, and R. Bhatia, "A Machine Learning approach for automation of Resume Recommendation system", *Procedia Computer Science*, vol. 167, pp. 2318-2327, Jan 2020.

[12] K. Tejaswini, V. Umadevi, S.M. Kadiwal, and S. Revanna, "Design and development of machine learning based resume ranking system", *Global Transitions Proceedings*, vol. 3, no. 2, pp. 371-375, Nov 2022.

[13] A. Vaswini, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Oct 2022, arXiv:1910.01108

[15] G. Lample, and A. Conneau, "Cross-lingual language model pretraining," Jan 2019, arXiv:1901.07291

[16] V. James, A. Kulkarni, and R. Agarwal, "Resume Shortlisting and Ranking with Transformers," in *International Conference on Intelligent Systems and Machine Learning*, Dec 2022, pp. 99-108.

[17] E. Loper, and S. Bird, "Nltk: The natural language toolkit," May 2002, arXiv preprint cs/0205028.