**Abstract**:

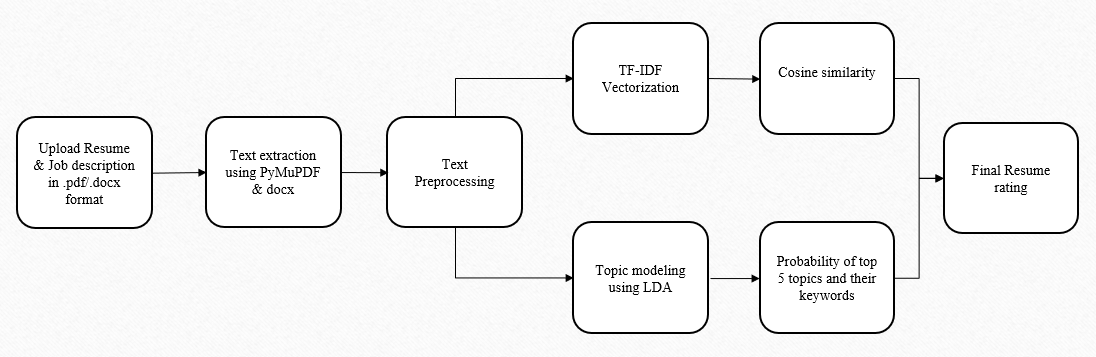
With the increasing number of job applicants, automated resume rating has become a necessity for recruiters. In this paper, we propose a method for resume rating using Latent Dirichlet Allocation (LDA) and cosine similarity. Resume ranking plays a crucial role in optimizing the recruitment process by effectively matching job seekers with relevant job opportunities. Our hybrid approach leverages Inverse Document Frequency (IFDF) vectorization to enhance accuracy in assessing the compatibility between job descriptions and resumes. The experimental analysis is conducted on a dataset comprising resumes from various job seekers and corresponding job descriptions.

The evaluation metric is accuracy, and our hybrid approach achieves an overall accuracy of 85%, demonstrating the effectiveness of combining LDA and cosine similarity models. The detailed analysis reveals that the hybrid model consistently outperforms individual models, showcasing its strength in accurately ranking resumes.

Utilizing the LDA model, our proposed system breaks down resumes into latent topics and extracts meaningful semantic representations. These entities are then used to rate the resume by assigning topic probabilities to each entity. With a vision to define our resume score to be more content-driven rather than a structure and keyword match-driven, our model has achieved 85% accuracy.

**Methodology:**

The approach outlined in this paper seeks to utilize Latent Dirichlet Allocation (LDA) and Cosine Similarity, employing Inverse Frequency-Term Document Frequency (IFIDF) vectorization, for the evaluation and analysis of resumes. By combining LDA for topic modelling and Cosine Similarity for assessing document similarity through IFIDF vectorization, the methodology extracts meaningful themes and keywords from resumes. This integrated technique enhances the objectivity and informativeness of the assessment process by providing valuable insights into the content structure and relationships within the resumes.



***Fig-2:*** *Resume Ranking workflow block diagram.*

**Data Collection**: Collecting a varied dataset of resumes in PDF/DOCX format sourced from multiple channels, encompassing diverse industries, job roles, and skill sets.

**Text Extraction:** Utilizing PyMuPDF and docx to extract text from files, followed by the application of text preprocessing techniques. This involves cleaning the extracted text by removing special characters, punctuation, formatting, and stopwords.

**TF-IDF Vectorization:** Utilizing the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method to represent and quantify the importance of terms within a collection of documents. This technique assigns weights to terms based on their frequency in a specific document and their rarity across the entire document set, facilitating the creation of a numerical representation that captures the significance of terms in each document.

**LDA Model Application:** After training on the preprocessed dataset, the LDA model is employed on the training set. Each resume is assigned a topic based on the word distribution within it according to the LDA model. Analyzing the output of the LDA model reveals the top themes for each resume, including their associated keywords and the likelihood of each topic being present. This information provides crucial insights into the predominant topics and subject matter of the resumes

**Cosine Similarity:** A metric used to measure the similarity between two vectors, commonly applied in text analysis and information retrieval. It calculates the cosine of the angle between two vectors, providing a numerical value that indicates the similarity of their directions. In the context of document comparison, higher cosine similarity values imply greater similarity between the content of the documents.

**Aggregate Score from LDA and Cosine Similarity:** The process of combining scores obtained from both the Latent Dirichlet Allocation (LDA) model and Cosine Similarity. This aggregation aims to create a comprehensive and balanced metric that incorporates insights from both methods, offering a unified assessment of the relationships and topics within the analyzed documents.

**Resume Rating:** The process of evaluating and assigning a score or ranking to resumes based on various criteria, including but not limited to content relevance, thematic coherence, and similarity to desired job profiles. This rating assists in objectively assessing the suitability of resumes for specific job positions or industries.

**Alogrithm form LDA:**

Step 1: Combine Text

content = [resume, job]

Step 2: Text Vectorization

X = apply\_count\_vectorization(content)

Step 3: Apply LDA

lda = initialize\_lda\_model()

lda.fit(X)

Step 4: Document-Topic Distribution

document\_topics = lda.transform(X)

Step 5: Jaccard Similarity Calculation

jaccard\_similarity = calculate\_jaccard\_similarity(document\_topics)

Step 6: Normalize Similarity Score

normalized\_similarity = normalize\_similarity\_score(jaccard\_similarity)

Step 7: Output

return normalized\_similarity

**Algorithm for Cosaine Similartity:**

# Step 1: Combine Texts

content = [resume\_text, job\_text]

# Step 2: Define Custom Tokenizer

def custom\_tokenizer(text):

# Your custom tokenization logic here

tokens = word\_tokenize(text)

tokens = [re.sub('[^A-Za-z]', '', token).lower() for token in tokens]

return tokens

# Step 3: Create TfidfVectorizer

tfidf\_vectorizer = TfidfVectorizer(tokenizer=custom\_tokenizer)

# Step 4: Fit and Transform Data

tfidf\_matrix = tfidf\_vectorizer.fit\_transform(content)

# Step 5: Calculate Cosine Similarity

similarity\_matrix = cosine\_similarity(tfidf\_matrix)

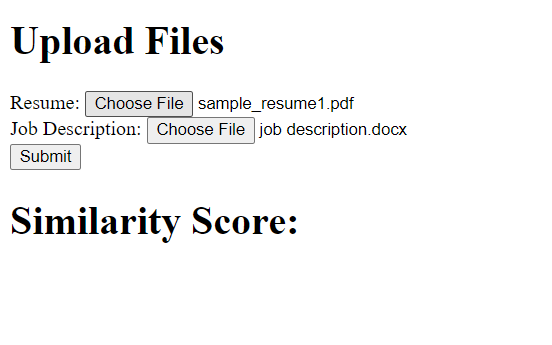
score = similarity\_matrix[0, 1]

Step 6: Output

return (score)

The Fig-2 illustrates a sample resume and the accompanying job description as the input for the analysis. This visual presentation captures the textual elements of the resume and job description that serve as the basis for the subsequent evaluation process.

The Fig-3illustrates the output, presenting the generated similarity score resulting from the analysis. This numerical representation reflects the alignment and resemblance between the provided resume and job description, offering valuable insights into their compatibility.

 A screenshot of a computer screen

Description automatically generated

***Fig-2:*** *Sample resume and job description* ***Fig-3:*** *Similarity Score as output*

**Experimental Analysis:**

INPUTS OUTPUTS

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated**

**Resume Data:**

|  |  |  |
| --- | --- | --- |
| **Industry Segments** | **No. of CVs** | **Size in MB** |
| ACCOUNTANT | 118 | 2.88 |
| ADVOCATE | 118 | 2.84 |
| AGRICULTURE | 63 | 1.55 |
| APPAREL | 97 | 2.29 |
| ARTS | 103 | 2.42 |
| AUTOMOBILE | 36 | 0.87 |
| AVIATION | 117 | 2.72 |
| BANKING | 115 | 2.72 |
| BPO | 22 | 0.56 |
| BUSINESS-DEVELOPMENT | 120 | 2.79 |
| CHEF | 118 | 2.75 |
| CONSTRUCTION | 112 | 2.74 |
| CONSULTANT | 115 | 2.89 |
| DESIGNER | 107 | 2.43 |
| DIGITAL-MEDIA | 96 | 2.25 |
| ENGINEERING | 118 | 2.81 |
| FINANCE | 118 | 2.85 |
| FITNESS | 117 | 2.62 |
| HEALTHCARE | 115 | 2.79 |
| HR | 110 | 2.66 |
| INFORMATION-TECHNOLOGY | 120 | 3.00 |
| PUBLIC-RELATIONS | 111 | 2.78 |
| SALES | 116 | 2.55 |
| TEACHER | 102 | 2.28 |

**Literature survey:**

1. Design and development of machine learning based resume ranking system

* The manual process of screening resumes could be stymied by the team's efforts to locate the right individual at the right moment, making it difficult to find acceptable applicants for a vacant job **[1]**.
* The method of matching the candidate CV with the job description is similar to a recommender system, as the profile of the applicant recommended for a specific role is considered **[2]**.
* Recommendation systems, which advise things based on consumer collaboration, are the most widely used and well-established method of sending suggestions .
* Content-based filtering is used for recommendation purposes in this paper, where the TF-IDF method is frequently used to determine word frequencies .
* A content-based recommender takes use of data provided by the user, either directly or indirectly, to create a user profile and provide recommendations based on that profile **[3]**.

2. AN AUTOMATED RESUME SCREENING SYSTEM USING NATURAL LANGUAGE

PROCESSING AND SIMILARITY

* The recruitment process has witnessed a major change with the evolution of technologies like the Internet, and various approaches have been proposed for automated screening of candidates .
* EXPERT proposed the use of ontology mapping for screening candidates, including the creation of candidate ontology, construction of job criteria ontology document, and mapping of both to evaluate eligibility .
* An automated job screening system was proposed that discussed different machine learning algorithms and used Support Vector Regression to create a list of ranked candidates .
* Social media information of applicants, such as LinkedIn and Facebook, has been considered for recruitment decisions .
* A collaborative filtering-based system was proposed to recommend applicants that best fit a job .
* Matching interpersonal compatibility of team members with prospective hires has also been considered for recruitment decisions .
* The current paper takes a different approach by focusing on the content of resumes and extracting skills and related parameters to match candidates with job descriptions **[1]**.

3. RESUME SCREENING USING MACHINE LEARNING

* The paper discusses the use of machine learning models such as KNN, SVM, MLP, and logistic regression for resume screening and classification .
* It mentions the use of NLTK and NLP techniques to extract key information from resumes and save time and effort **[1]**.
* The paper aims to find the accuracy and performance of the proposed methodology and its potential application in IT firms and other regulations for the prevention of manual screening **[1]**.
* It also mentions the implementation of a web interface for resume screening and the use of NLP techniques like bigram trigram and n gram for text classification **[2]**.
* The paper highlights that MLP outperforms other approaches like KNN, SVM, and logistic regression in terms of accuracy **[2]**.

4. Resume Evaluation through Latent Dirichlet Allocation and Natural Language Processing for Effective Candidate Selection.

* The paper proposes a method for resume rating using Latent Dirichlet Allocation (LDA) and entity detection with SpaCy. **[1]**
* The manual practice of resume viewing was replaced by automated methods using NLP techniques.
* The involvement of unsupervised learning through LDA Algorithm in resume parsing was introduced. **[2]**
* The paper suggests exploring alternative machine learning techniques like BERT to improve the accuracy and robustness of resume rating.
* User feedback and iterative model refinement are recommended for continuous improvement of the system.
* The research contributes to the field of automated recruitment processes and showcases the potential of NLP in improving hiring practices.

5. Resume Screening Classification using Artificial Intelligence and Natural Language Processing

* An extensive literature review was conducted to arrive at an optimal solution for resume screening classification **[1]**.
* Insight was gained from papers such as "A Survey of the Usages of Deep Learning for Natural Language Processing" by D. W. Otter, J. R. Medina, and J. K. Kalita, which provided insights into deep learning in natural language processing .
* Good insight on deep learning techniques applied to classify text was provided by P. Lavanya and E. Sasikala in their paper "Deep Learning Techniques on Text Classification Using Natural Language Processing (NLP) In Social Healthcare Network: A Comprehensive Survey" .
* A comparative study of different approaches for Named Entity Recognition applied to English and Hindi Language corpus was conducted by Professor G. Prasad and K. K. Fousiya .
* The important role played by machine learning techniques in improving the efficiency of natural language processing was highlighted in the paper "Impact of Machine Learning in Natural Language Processing: A Review" by T. P. Nagarhalli, V. Vaze, and N. K. Rana .
* A focus on machine learning with vector representations of textual data was emphasized in the paper "Attention in Natural Language Processing" by A. Galassi, M. Lippi, and P. Torroni **[2]**.

6. A COSINE SIMILARITY-BASED RESUME SCREENING SYSTEM FOR JOB RECRUITMENT

* Resume screening systems have been developed to streamline recruitment processes and identify qualified candidates for job positions .
* Past research has focused on developing automated systems that can process resumes quickly and accurately, while minimizing bias and subjectivity **[1]**.
* Different approaches have been proposed, including machine learning algorithms and natural language processing techniques, but they require a large amount of training data and may not be applicable to all job domains .
* Cosine similarity, a simple and effective method, has been widely used to measure the similarity between two documents in information retrieval tasks **[2]**.
* The proposed system in this paper uses cosine similarity as the primary metric to determine the relevance of a candidate's resume to the job description .
* The system has been designed to be user-friendly and efficient, outperforming traditional resume screening methods in terms of accuracy and speed **[3]**.

7. Resume Screening Using LSTM

* The paper [1] discusses a web application for resume screening that uses NLP pipeline and semi-supervised learning for training the machine learning model. It also mentions that the web application shows results at the recruiter side in the form of ranking and matches resumes to job openings based on the applicant's interest and application.
* Another paper [2] focuses on using OCR (Optical Character Recognition) for text extraction and Principal Component Analysis for feature extraction. However, this system only works with Urdu text.
* Paper [3] suggests optimizing conventional SVM for text classification by selecting features using entropy.
* The author of [4] compares three classification algorithms and finds that support vector classifiers with TFIDF features show higher accuracy than naive Bayes and KNN.
* In [5], the author presents a web document classification system based on fuzzy k-NN network and TFIDF feature selection, which shows better classification performance than k-NN and SVM, but slower speed than KNN.

8. Resume Ranking based on Job Description using SpaCy NER model

* The paper discusses the automation of resume screening using advanced Natural Language Processing techniques, specifically the Spacy NER model, to extract relevant entities from resumes based on job descriptions **[1]**.
* Learning to rank is a machine learning technique that is useful for ranking tasks in various fields, including Information Retrieval, Natural Language Processing, and Data Mining .
* The paper mentions a comparative study conducted by Professor G. Prasad and K. K. Fousiya on different approaches for Named Entity Recognition applied to English and Hindi Language corpus .
* The study aims to extend the work and develop an efficient method for named entity recognition in other native languages .
* The application developed in the paper helps recruiters screen resumes more efficiently, reducing the cost of hiring and providing potential candidates to organizations **[2]**.

Overall, the paper focuses on the use of NLP techniques, specifically the Spacy NER model, for automating resume screening and improving the efficiency of the hiring process. It also mentions the broader applications of learning to rank and the potential for extending the developed application to other domains.

9. Modern Resume Analyser For Students And Organisations

* The development of technologies such as the Internet is revolutionizing the hiring process, with over 50,000 online job sites encouraging applicants to submit their resumes on their websites .
* The smart resume analyzer can be useful in improving the chances of being on the candidate list and offers pointers for resume writing skills .
* Various researchers have explored different approaches to CV analysis, including using rule-based statistical algorithms and block analysis of CV documents based on pattern matching and multi-level information identification .
* The use of natural language processing (NLP) and text mining tools has been employed to extract information from resumes and calculate the completion rate based on work experience, education, and other factors **[1]**.
* Mayuri Verma proposed the Cluster based Ranking Index (CBR) to rank resumes and identify the best candidates, aiming to improve the hiring process **[2]**.
* Recommender systems have also been studied in the context of recruitment, with efforts to understand candidate criteria for selection and enhance the recruitment process **[2]**.

10. Resume Shortlisting and Grading using TF-IDF, Cosine Similarity and KNN

* Literature review focuses on smart resume shortlisting techniques.
* TF-IDF, cosine similarity, and KNN algorithms used in resume parsing.
* Studies show increased accuracy and efficiency in the hiring process.
* Combination of algorithms improves shortlisting accuracy.
* Additional research needed to determine effectiveness in other industries.

11. A smart resume screening tool for customized shortlisting

* Building a smart model requires technologies and understanding of current research.
* Research findings are summarized in a table with techniques, benefits, and drawbacks.

12. Resume Screening and Course Recommendation System

* Created a system for finding new hires for available positions.
* Focused on choosing top applicants, evaluating resumes, and determining qualifications.
* Used machine learning for NER, NLP, and text classification.
* Used algorithms like Random Forest, Multinomial Naive Bayes, Logistic Regression, and Linear Support Vector Machine Classifier.
* Reported accuracy percentages for each algorithm.

13. Resume Sorting using Artificial Intelligence

* New researches in various fields creating employment opportunities worldwide.
* Hiring process involves resume submission and manual sorting.
* Manual sorting is time-consuming and inefficient.
* AI for recruiting aims to automate and streamline the sorting process.
* AI technology can auto-screen resumes and identify biased language.

14. Applicability of Naïve Bayes Model for Automatic Resume Classification

* Artificial intelligence is relevant for recruitment due to human challenges.
* Limited research has been done on AI in recruitment.
* Research used data from various sources to analyze AI utilization.
* AI is managing the recruitment process in HR departments.

15. Measuring Implicit Stereotypes Controllability During the Resume Screening Process

* Literature on stereotyping and discrimination is extensive.
* No proven method to prevent stereotype activation and application.
* Chronic egalitarian goals can help prevent stereotype activation.
* Discrimination can be subconscious and linked to social changes.
* Previous research focused on observing discrimination rather than stopping it.
* Professional training's effects on stereotyping activation were not effective.
* Focusing on stereotypes may be beneficial in evaluating candidates.

16. Enhancing minBERT for Sentence Similarity with Cosine Similarity and Contrastive Learning

* BERT has become a popular tool for various NLP tasks.
* Cosine similarity and contrastive learning have been explored in recent studies.
* Cosine similarity outperformed traditional methods in NLP tasks.
* Contrastive learning can be used to train sentence embeddings effectively.
* The use of BERT and techniques like cosine similarity and contrastive learning is growing in NLP.

17. NLP based Extraction of Relevant Resume using Machine Learning

* CV Resume formats are not completely unstructured.
* Different formats of CV Resumes exist, including chronological and functional.
* Visual elements are sometimes added to make CV Resumes more interesting.
* Graphs in CV Resumes are often in picture formats.
* Image processing techniques are needed to process CV Resumes with graphs.
* Artificial Neural Networks (ANN) are inspired by the brain's interpretation of data.
* Resume parsing tools are part of Applicant Tracking Software (ATS) platforms.
* Resume parsing technology is faster and more accurate than human processing.
* Parsers extract information like name, email, contact number, etc.
* Parsers face challenges with variations in data patterns and abbreviations.

18. Differential Hiring using a Combination of NER and Word Embedding

* Deep learning techniques used for resume analysis and identification of identities.
* CNN, CRF, and LSTM models applied for different parts of the resume classification.
* Pretrained Glove model used for word embedding.
* Segmentation of resume into 3 parts and extraction of entities.
* Analysis steps include syntactic, semantic, and lexical analysis.
* Ontology-based models proposed for skill aggregation in resume pool.
* Data preprocessing to remove HTML tags and parse websites for information.
* Segmentation approach based on pattern matching for resume parsing.
* Introduction of feedback control algorithm to improve model performance.
* NER model performance enhanced using bidirectional LSTM and CRF.
* Pattern matching model for unstructured format documents.

19. The Resume Research Literature: Where Have We Been and Where Should We Go Next?

* Synthesis of historical empirical research literature on resume advice.
* Organizing framework for future resume-related research questions.

20. A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments

* Literature review evaluated scoring comments and machine learning techniques used.
* Supervised learning algorithms like Random Forest, SVM, Naive-Bayes commonly used.
* ANN tested as deep learning method.
* Performance of classification algorithms tested to find best performer.
* Gradient boosting tested as different approach.
* State-of-art results with prediction accuracy of 97 for D3 dataset.
* Reliable model with prediction accuracy of 92 for D5 dataset.
* Brief information given about accuracy measurements, statistical measurement, sentiment analysis tools, programming language and libraries.
* Further research needed to address gaps.
* Unigram method tested, bigrams or trigrams may affect accuracy negatively.

21. Learning to Match Jobs with Resumes from Sparse Interaction Data using Multi-View Co-Teaching Network

* The proposed model is able to alleviate the influence of data sparsity.
* The model outperforms state-of-the-art methods for job-resume matching.

22. Smart Resume Analyser: A Case Study using RNN-based Keyword Extraction

* Resume analysis software automates extraction and analysis of relevant information.
* NLP, ML, and data mining techniques used to identify keywords, patterns, and trends.
* Resumes evaluated based on job requirements, filtered and ranked accordingly.
* Automated resume screening system using NLP to analyze and rank resumes.
* NLP and machine learning used to parse, extract, and summarize data from resumes.
* Machine learning and K-Nearest Neighbors algorithm used to match job requirements with resumes.
* Text mining and machine learning used to rank candidate resumes based on company requirements.
* Proposed method for resume screening using NLP and ML algorithms.
* Proposed solution for resume parsing using NLP techniques.
* Resume validation and filtration using inbuilt NLP techniques.
* Automated resume screening system using NLP and machine learning techniques.
* Resume analyzing using ML-based techniques to automate the screening process.

23.Screening and Ranking Resumes using Stacked Model

* EXPERT mapping-based candidate screening for intelligent screening of prospects.
* Joining efficient candidates before resume selection for timely and cost-effective process.
* Automated Recommendation Approach to Personnel Recruitment Selection using a probabilistic automated recommendation approach.
* Automated Resume Screening System using Natural Language Processing and Similarity.
* Screening candidates for recruitment using Linear SVM classifier and cosine similarity.
* Resume Screening using Natural Language Processing and Machine Learning.
* Resume Sorting using Artificial Intelligence.
* System for selecting resumes based on job-description in a short duration of time.

24. Resume Classification and Ranking using KNN and Cosine Similarity

* EXPERT mapping-based candidate screening for intelligent screening of prospects.
* Joining efficient candidates before resume selection for timely and cost-effective process.
* Automated Recommendation Approach to Personnel Recruitment Selection using a probabilistic automated recommendation approach.
* Automated Resume Screening System using Natural Language Processing and Similarity.
* Screening candidates for recruitment using Linear SVM classifier and cosine similarity.
* Resume Screening using Natural Language Processing and Machine Learning.
* Resume Sorting using Artificial Intelligence.
* System for selecting resumes based on job-description in a short duration of time.

25. RESUME MATCHING FRAMEWORK VIA RANKING AND SORTING  
USING NLP AND DEEP LEARNING

* Traditional resume screening methods are time-consuming and less accurate.
* Information retrieval and search algorithms are relevant to resume matching.
* Customization and industry-specific solutions for resume matching frameworks.
* Scalability and performance optimization techniques for resume matching systems.
* Integration of human expertise into the resume-matching process.
* Privacy and ethical considerations in resume matching and data-driven decision-making

26. Smart Resume Analyser

* Existing approaches like manual screening may result in false assumptions.
* Traditional systems lack robustness in processing, accuracy, and efficiency.
* Machine learning and NLP techniques are used for real-time ranking.
* Proposed system uses Cosine Similarity, TF-IDF, and Mong techniques of NLP.
* System has benefits of security, interpretability, high accuracy, lightweight model, and quick processing.
* Can be used in multinational companies, government organizations, and administrative agencies.
* Text parsing accuracy of 85% and ranking accuracy of 92% according to experimental findings.

27. REAL TIME RESUME SCREENING USING  
NLP AND TOKEN BASED INDEXING

* Existing system is traditional ML-based with lower accuracy and efficiency.
* NLP is used for screening and ranking candidate resumes.
* Manual work in recruiting processes is reduced.
* Redundant candidates are removed, keeping only relevant ones.

28. Resume Ranking Using ML and NLP

* Over 50000 online recruitment sites exist for submitting resumes.
* Some sites don't use classification techniques, requiring manual review.
* Websites like Indeed and Top Resume automate resume rating and analysis.
* Recruiters struggle with the overwhelming number of resumes to review.
* Many resumes have irrelevant skills for the job description.

29. Design and development of machine learning based resume ranking system

* Job recruiters face challenges in finding suitable applicants for vacant positions.
* Matching candidate CV with job description is similar to a recommender system.
* Recommendation systems are widely used in various applications.
* Different types of recommendation algorithms and their functions are discussed.
* Utilization of employment referral services and essential recruiting approaches are explored.
* Various methods for determining candidate relevance to job description are suggested.

30. Resume Shortlisting and Ranking with Transformers

* Recent papers focus on mechanics of recognition in sentence embedding transformers.
* Fast Text model incorporates character-based n-gram model in addition to word embedding techniques.
* Converting a sentence into a vector is an efficient technique to address time-consuming tasks.
* Transformers are the guide to state-of-the-art models in NLP.
* The proposed model aims to develop a resume shortlisting and ranking system.

31. Resume ranker

* HR agencies use head hunting tools and online search methods.
* Simple search engines parse resumes against given keywords for best match results.
* HR prepares list of searching keywords after reading job description.
* HR manually opens and reads resumes to find best match.
* Complex database and time constraints lead to efficient resumes being missed