# A Review on Text Categorization for E-Recruitment Based on Semantic Analysis

**Prajakta Bhapkar[1], Nikita Phulkar[2], Razkeen Shaikh[3], Sana Kauser Shaikh[4]**
**[1, 2, 3, 4](Students, Department of Computer Engineering, SAE Kondhwa, Pune, India)**

**Prof. Dr. Harsha Bhute**
**(Professor, Department of Computer Engineering, SAE Kondhwa, Pune, India)**

***ABSTRACT:- As we know Indian I.T sector is second largest candidate recruiting area of our country, it gives about 7.5% to our Gross Domestic Product . The recruitment in the Information Technology sector has seen an exponential increase in recent times. Companies recruit thousands of young talent, right from the college every year through campus fairs it is difficult to identify the good match between the qualifications of the candidate and match the skill that a company seeks by examining each resume for HR department in any organization. To address this challenge, many companies have shifted for using e-recruiting platforms. These platforms decrease the cost, time and effort required for physically processing and screening candidate resumes.***

***Keywords: - Resume Classification, Recruitment, Software companies, Natural language processing (NLP), Resume Examining, Text Categorization, and Semantic Analysis.***

## I. INTRODUCTION

Resume screening is the process of sorting resumes to disqualify candidates through successive rounds of scrutinizing using various filtering mechanisms. These mechanisms typically aim at reducing the number of resumes to be processed in the subsequent rounds of hiring. The proposed method presents a simple solution that serves as an initial screening of candidates and reduces human workload.

These systems employ different approaches to talk the challenges associated with screening, matching, and classifying candidate resumes. The different that made to make the process easy Although these approaches produce high precision ratios in finding candidates to fill a vacancy , they give less attention to the run time complexity of the corresponding process i.e. every job offer will be matched with every resume in the corpus instead of matching resumes that are only related to their occupational category.

The proposed resume Classifier tries to find the resumes for any job/university interview more robust by doing information extraction approach based on the data of previously selected and rejected candidates. The System extracts the information from the resume. Then Natural language processing (NLP) [1] technologies are used for parsing, tokenizing, stemming and filtering the content of the data. By using TF-IDF we can calculate the score of the particular resume based on the recruiter information and suggest lacking skills to the users and recommend top resume to recruiter.

Using NLP(Natural Language Processing) and ML(Machine Learning) to rank and classify the resumes according to the given constraint, this intelligent system ranks the resume of any format according to the given constraints or the following requirement provided by the client company.

The remaining part of this paper is organized as follows: Section I contains the introduction of "Text Categorization for E-Recruitment Based on Semantic Analysis" Section II elaborate the background of the Hiring system. Section III gives the brief overview of the approaches to Natural Language Processing (NLP). It also discusses about some goals and key challenges that are faced by the NLP. Section IV reviews some of the existing works related to Resume Classification using different techniques. Section V presents the detailed description for the proposed system with its architecture and the flowchart as well as the algorithms used in the proposed system. Section VI presents the literature survey of different papers. Finally, this paper is concluded to be carried out is stated in Section VI.

## II. BACKGROUND

- **First Generation Hiring Systems:**

In this System the Hiring team would publish their vacancies and invite applicants. Medium of publishing were newspaper, television and mouth. The interested candidates would then apply by sending their resumes.

These resumes were then received and sorted by the hiring team and shortlisted candidates were called for further rounds of interviews.

**Drawback: -** The whole process would take lot of time and human efforts to find right candidate suitable for their job roles.

- **Consultancy units based Hiring Systems→ Second Generation Hiring Systems:**

As the industries have grown, there hiring needs has rapidly grown. To serve these hiring needs certain consultancy units have come into existence. They offered a solution in which the candidate has to upload their information in a particular format and submit it to the agency. Then these agencies would search the candidates based on certain keywords. These agencies were middle level organizations between the candidate and company.

**Drawback: -** These systems were not flexible as the candidate has to upload there resume in a particular formats, and these formats changed from system to system.

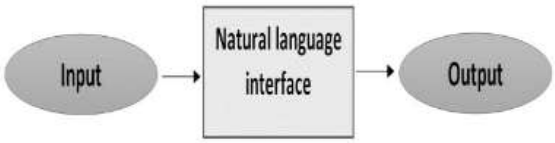
- **Proposed system→Third Generation Hiring Systems:**

This is our proposed system, which allows the candidates to upload their resumes **inflexible format.** These resumes are then analyzed by our system, indexed and stored in a specific format. This makes our search process easy. The analyzing system works on the algorithm that uses **Natural Language Processing.** It reads the resumes and understands the natural language/format created by the candidate and transforms it into a specific format.

## III. APPROACHES TO NATURAL LANGUAGE PROCESSING-NLP

- **NLP SYSTEM INCLUDES:**

1) User input
2) It goes to the natural language interface
3) Output obtained in a language that is understood by the application program




- **GOALS OF NLP**

The most natural means of communication between humans is Natural Language, spoken, written or typed. The dominance of natural language as a means of communication among humans suggests that it would be an agreeable medium in human-computer interaction. Thus, the major goal of NLP would be the ability to use natural language as effectively as humans do.

There is no tool that can provide an expert human quality word-sense disambiguation.

Also, the goal of NLP is to enable computers to engage themselves in communication using natural human speech and language, so that non-programmers can interact with the computers easily and effectively.

- **CHALLENGES IN NLP**

**A) Machine Translation:** Machine Translation is the task of automatically converting one natural language into another natural language, preserving the meaning of the input text and producing fluent text in the output language. Correct translation requires not only the ability to analyze and generate sentences in human languages but also human like understanding of world knowledge and context, despite the ambiguities of languages i.e., computers should be able to understand input in more than one language, provide output in more than one language and translate between languages.

**B)** Today, most NLP resources and systems are available only for high resource languages (HRLs), such as English, French and German. Whereas many low resource languages (LRLs) such as Indonesian, Swahili-spoken and written by millions of people have no such resources or systems available. So a future major challenge for the NLP community is to develop resources and tools for hundreds and thousands of languages, not just a few.

**C) Reading and Writing Text:** Text reading and writing is one of the major challenges in NLP. Machine reading is the idea that machines could become intelligent, integrate and summarize information for humans, by reading and understanding the text available i.e., computers should be able to understand and process the data.

**Semantic Analysis:** Semantic Analysis is related to create the representations presentations for meaning of linguistics inputs. It deals with how to determine the meaning of the sentence from the meaning of its parts. So, it generates a logical query which is the input of Database Query Generator. It is another form of representation for user tokens and user input symbols in the form of semantic word.

## OBJECTIVES OF THE PROPOSED SYSTEM

1. To reduce human workload for resume classification.
2. To provide quick process of sorting resumes.
3. To improve the performance of resume selection process.
4. To suggest lacking skills to the users and recommend top resume to recruiter.

## IV. LITERATURE SURVEY

**"AUTOMATED PROFILE EXTRACTION AND CLASSIFICATION WITH STANFORD ALGORITHM"** Renuka S. Anami et.al introduced the system that provides an effective approach for the extraction of information from the resumes. Here the cascaded Hybrid Model of Stanford Algorithm In the cascaded hybrid model. This system will make the task of both candidate and HR Manager easier and faster. This system avoids the complexity in form filling procedure of the candidates by directly asking the user to upload only the resume.

**Algorithm Used: -** Cascaded Hybrid Model of Stanford Algorithm

**Advantage: -** Classifier specific method is applied and dependent on the combination of the different feature selection method is done.

**"RESUME CLASSIFICATION USING ANALYTIC HIERARCHY PROCESS AND KEYWORD EXTRACTION"** Vishnunarayanan.R et.al presents a classification technique using analytic hierarchy process (AHP) and keyword extraction has been proposed in a hiring scenario. The proposed approach acts only as a support tool that can be used to obtain a subset of resumes that might be suitable to consider for subsequent rounds of the hiring process.

**Algorithm Used: -** Analytic Hierarchy Process (AHP)

**Advantages: -** 1) AHP is broadly spread in the scholastic group and connected in distinctive fields like Engineering, medicine and other sciences.
2) The qualities incorporate (i) Its usability (ii) It's an effortlessly reasonable system
**Disadvantages:-** There is inconsistency in positioning when including or erasing options utilized as a part of the information set.

**"AUTOMATED CV CLASSIFICATION USING CLUSTERING TECHNIQUE"** Prof. Sagar More et al. focus on the calculating a score for resume by using clustering techniques which makes it easy for HR department to get to the eligible candidate. The system provides an option to HR team to customize each and every job before uploading as per requirement and skill set required.
**Algorithm Used: -** K-Means Clustering
**Advantage: -** If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.
**Disadvantages: -** 1) Different initial partitions can result in different final clusters.
2) It does not work well with clusters (in the original data) of Different size and Different density

**"DOMAIN ADAPTATION FOR RESUME CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS"** Luiza Sayfullina et. al The system classifies the resume data of job applicants into 27 different job categories using convolutional neural networks. Here classifiers on a large number of freely available job description snippets and then use it to classify resume data.
**Algorithm Used: -** Convolutional Neural Network
**Advantage: -** High Accuracy
**Disadvantages: -** 1) A Convolutional neural network is significantly slower due to an operation such as maxpool.
2) If the CNN has several layers then the training process takes a lot of time if the computer doesn't consist of a good GPU.

**"RESUME ANALYZER AN AUTOMATED SOLUTION TO RECRUITMENT PROCESS"** Ankita Vaidya & Pooja Sawant introduced proposed an analyzer system that uses text mining that extracts the details of the job seeker from its resume. Here, the author has used text mining as it is used for extracting the text from the unstructured document. Also, the extracted text is converted into data for analysis.
**Algorithm Used: -** Text Mining

**"SEMANTIC MATCHMAKING FOR JOB RECRUITMENT: AN ONTOLOGY-BASED HYBRID APPROACH"** Maryam Fazel-Zarandi et. al describes an ontology-based hybrid approach that matches job seekers and job advertisements by utilizing a similarity-based approach to rank applicants. The proposed system exploits semantic technologies in order to improve the matching process. However, the main drawback of this approach is the huge cost (run time complexity) of the matching process
**Algorithm Used: -**Ontology-Based Hybrid Approach
**Advantages: -** Context reasoning is to check the consistency of contexts as well as deducing high-level implicit context information from low-level explicit contexts.

The author introduced the method based on recommendation system based where Job Description provided by the employer is matched with the content of resumes in the space and the top *n* (n being configurable) matching resumes are recommended to the recruiter. The model takes the cleansed resume data and job description and combines the two into a single data set, and then computes the cosine similarity between the job description and CVs.
**Algorithm Used: -** Content-Based Recommender

## V. PROPOSED SYSTEM

The process of allotment of projects to the new recruits is a manual affair, usually carried out by the Human Resources department of the organization. Here an attempt is made to automate the process of resume classification. Figure 1 gives the system architecture of the proposed system.

1) **Upload Resume:** User can upload the resume to the system. The system preprocess the input to Remove@, Remove URL Remove Stop words to get the fine data and extracts the tag word from the data. The resume should have the skill mentioned.

**2) Preprocessing:** The system can Understand Each Word from all the resumes using Natural Language Processing (NLP). It can apply different techniques for understanding the sentence and word. They can analyze the words using two different ways like Sentence & Word Understanding

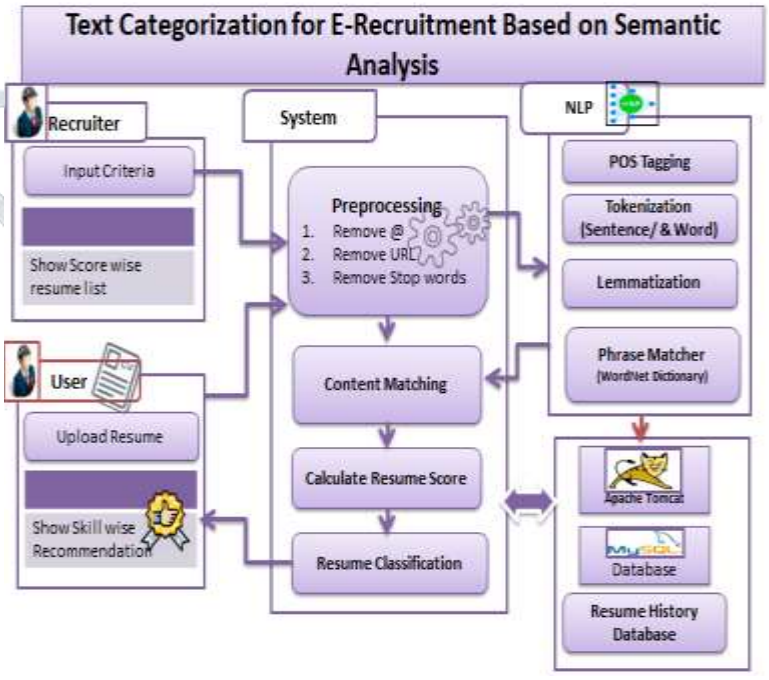



**Figure: - System Architecture.**

- **Sentence Tokenization:** In this techniques system can divides the sentence into several tokens. It split the large raw text into several sentence to get more meaningful information out. For eg. "All work and no play make jack a dull boy, all work and no play". The above sentence is divide into sentence like,{All work and no play make jack dull boy}All work and no play makes jack a dull boy.

- **POS Tagging:** This algorithm is used for detects if the word token is noun, verb, adjective POS Tagging in which a word is assigned in accordance with its syntactic functions. In English the main parts of speech are noun, pronoun, adjective, determiner, verb, preposition, adverb, conjunction, and interjection.

- **Word Tokenization:** This technique the sentence or data can split into several words. For eg. "All work and no play makes jack a dull boy, all work and no play" This sentence split into word like,[All, work, and, no].
- **Word Lemmatization:** Lemmatization is a more methodical way of converting all the grammatical/inflected forms of the root of the word. Lemmatization uses context and part of speech to determine the inflected form of the word and applies different normalization rules for each part of speech to get the root word (lemma).

| Rule | | | Example | | |
|---|---|---|---|---|---|
| SSES | → | SS | caresses | → | caress |
| IES | → | I | ponies | → | poni |
| SS | → | SS | caress | → | caress |
| S | → | | cats | → | cat |

- **Word Similarity:** By using this technique the system can find the similar words. We use the Word Net dictionary for finding the synonyms.

- **Sentence Similarity:** By using this technique the system can find the similar sentence.

**i) Phrase Similarity**:- By using this technique the system can find the similar words. We use the WordNet dictionary for finding the synonyms. WordNet Dictionary-WordNet is a combination of dictionary and thesaurus. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. For e.g. "Last night" →"yesterday"

**3) Recommend Resumes:** The score is calculated for each of the resume and recruiters show all the shortlisted resumes score-wise.

**4) Recommend Skills:** Based on the rejected resumes, which skill the particular user needs to improve according to the market condition.

**ALGORITHMS USED:-**

**Word Order Similarity between Sentences**

- Let's consider a particular case to illustrate the importance of word order. For example, for two sentences:

**T 1:** A quick brown dog jumps over the lazy fox.

**T 2:** A quick brown fox jumps over the lazy dog.

- These two sentences contain exactly the same words and most words appear in the same order. The only difference is that dog appears before fox in T 1 and dog appears after fox in T2.
- Since these two sentences contain the same words, any methods based on "bag of word" give a decision that T 1 and T 2 are exactly the same. However it is clear for a human interpreter that T 1 and T 2 are only similar to some extent. The dissimilarity between T 1 and T 2 is the result of the difference in word order.
- Therefore any efficient computational method for sentence similarity must take into account the impact of word order. Sentences containing the same words but in different orders may result in very different meanings. It is easy for humans to process word order information.
- However the incorporation of order information in to computational methods for understanding natural language is a difficult challenge. This may be the reason why most existing methods do not tackle this type of information.
- In this section we introduce a method that takes word order information into account when computing sentence similarity. Assume that for a pair of sentences, the joint word set is T . Recall the above two example sentences, their joint word set is: T = {A quick brown dog jumps over the lazy fox} For each word in T 1 and T 2 , a unique index number has been assigned respectively.
- The index number is simply the order number that the word appears in the sentence. For example, the index number is 4 for dog and 6 for over in T 1. In computing word order similarity, a word order vector r is formed for T 1 and T 2 respectively based on the joint word set T. For each word w i in T , we try to find the same or a similar word in T 1 as follows:

1. If T 1 contains an occurrence of the same word, we fill the entry for this word in r 1 with the corresponding index number in T 1. Otherwise we try to find the most similar word i w ~ in T 1.

2. If the similarity between w i and i w ~ is greater than a pre-set threshold, the entry of w i in r 1 is filled with the index number of i w ~ in T 1

3. If the above two searches fail, the entry of w i in r 1 is null. Having applied the above procedure for T 1 and T 2, the word order vectors for are r 1 and r 2 respectively. For the example

sentence pair, we have: r 1 = {1 2 3 4 5 6 7 8 9}, r 2 = {1 2 3 9 5 6 7 8 4}

Thus a word order vector is the basic structural information carried by a sentence. The task of dealing with word order is then to measure how similar the word order in two sentences is. We propose a measure for measuring word order similarity of two sentences as:

$$S_r = 1 - \frac{\|\mathbf{r}_1 - \mathbf{r}_2\|}{\|\mathbf{r}_1 + \mathbf{r}_2\|}$$

**FLOWCHART:-**

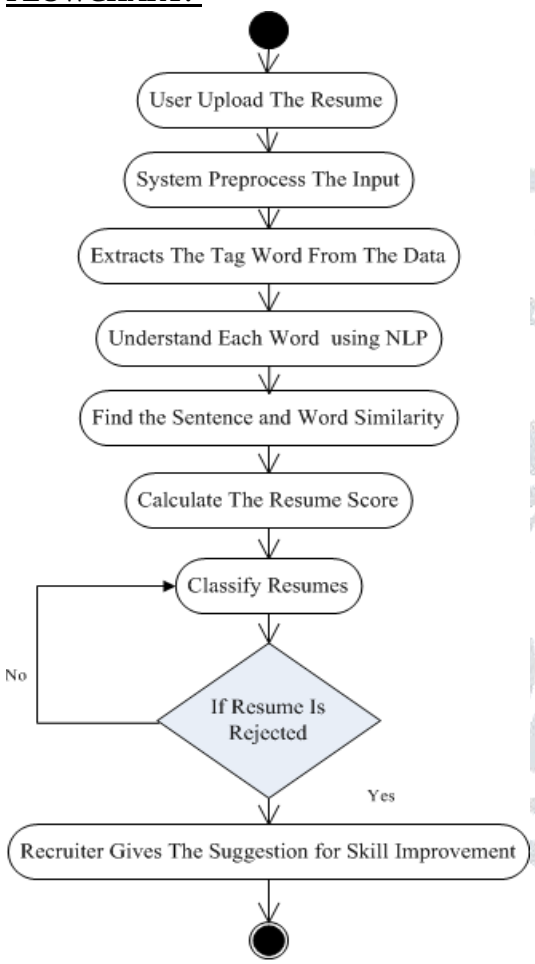



Figure: - Flow chart of the proposed system

## VI. ACKNOWLEDGMENT

We express our sincere thanks to Project guide, **Harsha Bhute,** for his continuous support. We also thankful to our Head of Department of Computer **Prof. B.B. Gite** For support

## VII. CONCLUSION

Due to the constant growth in online recruitment, job portals are starting to receive thousands of resumes. It is difficult to identify the good match between the qualifications of the candidate and match the skill that a company seeks by examining each resume manually.

As resumes contain unformatted text or semi-formatted text, extending the concept of special features for the development of an approach to process resumes is a complex task. Sorting the resumes to disqualify candidates through successive rounds of scrutinizing using various filtering mechanisms is quite completed and time consuming task.

The proposed system helps in classifying resumes by word/ sentences understanding. The system extracts information from resume; by using natural language processing system understands the meaning of the data. Resume score is calculated by using phrase matcher based on resume information and classify the resume. The system also help to reduce human workload for resume classification providing quick process of sorting resumes.

## REFERENCES

[1] Divyanshu Chandola1, Aditya Garg2, Ankit Maurya3, Amit Kushwaha ,\ONLINE RESUME PARSINGSYSTEM USING TEXT ANALYTICS",Volume: 09Issue: 01, July-2015, Available www.jmdet.com

[2] Renuka S. Anami, Gauri R. Raoo,"Automated Profile Extraction and Classification with Stanford Algorithm",International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-4 Issue-7, December 2014.

[3] Vishnunarayanan.R1, Shreekrishna Prasad2, Krishnan.A.N3, Palanivel.S4, Umamakeswari.A,"RESUME CLASSIFICATION USING ANALYTIC HIERARCHY PROCESS AND KEYWORD EXTRACTION",nternational Journal of Pure and Applied MathematicsVolume 115 No. 7.

[4] Prof. Sagar More1, Bhamare Priyanka2, Mali Puja3, Kachave Kalyani4,"Automated CV Classification using Clustering Technique" , International Research Journal of Engineering and Technology(IRJET)e-ISSN: 2395-0056Volume: 06 Issue: 06| June2019

[5] Luiza Sayfullina,Eric Malmi, Alexander Jung,Yiping Liao, "Domain Adaptation for Resume Classification Using Convolutional Neural Networks",July 2017 .

[6] Ankita Satish Vaidya and Pooja Vasant Sawant (2015), “Resume Analyzer an Automated Solution to Recruitment Process”, International Journal of Engineering and Technical Research (IJETR), Volume-3, Issue-8.

[7] Maryam Fazel-Zarandi1, Mark S. Fox2 ,“Semantic Matchmaking for Job Recruitment: An Ontology-Based Hybrid Approach”, at: https://www.researchgate.net/publication/265922198

[8] Abeer Zaroor,Mohammed Maree,Muath Sabha,\A Hybrid Approach to Conceptual Classification and Ranking of Resumes and Their Corresponding Job Posts",DOI: 10.1007/978-3-319-59421-7 10 Conference: Intelligent Decision Technologies 2017.

[9] Suhas Tangadle Gopalakrishna1 and Vijayaraghavan Varadharajan,\AUTOMATED TOOL FOR RESUME CLASSIFICATION USING SEMENTIC ANALYSIS", International Journal of Arti_cial Intelligence and Applications (IJAIA), Vol.10, No.1, January 2019.

[10] Kun Yu,Gang Guan, Ming Zhou,\Resume Information Extraction with Cascaded Hybrid Model.",DOI: 10.3115/1219840.1219902 · Source: DBLP Conference: ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA Cite this publication