IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306126|

Automated Text Analysis: Leveraging NLP for Categorization

Rutuja Bachhav, Dr. Monika Deshmukh, Dr. Mohd.Muqeem, Dr. Pawan. R. Bhaladhare

M. Tech (CTIS), Department of Computer Science and Engineering, Sandip University, Nashik, Maharashtra, India

Professor, Department of Computer Science and Engineering, Sandip University, Nashik, Maharashtra, India

ABSTRACT: Text categorization, also known as text classification, is a foundational task in the field of Natural Language Processing (NLP). It involves assigning predefined categories or labels to textual data based on its content. This process is essential for organizing and managing large volumes of text data, enabling applications such as spam detection, sentiment analysis, topic labeling, document organization, and more. The IT sector has seen a significant increase in the need for efficient text categorization techniques. With companies receiving thousands of resumes, it becomes challenging to identify the best match between a candidate's qualifications and the skills required by examining each resume manually. The Resume Classifier aims to enhance the process of text categorization for any job or university interview by using an information extraction approach based on data from previously selected and rejected candidates. The system extracts relevant information from resumes and utilizes Natural Language Processing (NLP) technologies for parsing, tokenizing, lemmatizing, and filtering the content. By employing a Phrase Matcher, the system can calculate a score for each resume based on specific criteria, identify areas where candidates may lack skills, and recommend the top resumes to recruiters. Initial results indicate a significant improvement in matching accuracy and processing time compared to traditional methods.

KEYWORDS: Text Categorization, E-Recruitment, Word/Sentence Similarity, Natural Language Processing, Phrase Matcher, Preprocessing, Parsing, Similarity between Sentences, Data Analysis.

I. INTRODUCTION

In recent times, the Information Technology sector has experienced a remarkable surge in recruitment activities. Companies are actively seeking to onboard thousands of young talents directly from colleges, often through campus fairs. However, the process of aligning candidate qualifications with the specific skills sought by companies presents a significant challenge for HR departments. To tackle this issue, many companies have turned to e-recruiting platforms, which streamline the process and reduce costs, time, and manual effort associated with screening applicant resumes. While these approaches often yield high precision ratios in identifying suitable candidates, they may overlook the runtime complexity of the matching process. For instance, every job offer is typically matched with every resume in the corpus, rather than focusing solely on resumes relevant to the specific occupational category.

To address these challenges, the proposed Resume Classifier aims to enhance the robustness of the process by implementing an information extraction approach. This approach is based on data from previously selected and rejected candidates, leveraging insights gained from past recruitment activities. By analyzing this data, the Resume Classifier can more effectively match resumes with job or university interview requirements, thereby improving the efficiency and accuracy of the recruitment process.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306126|

II. RELATED WORK

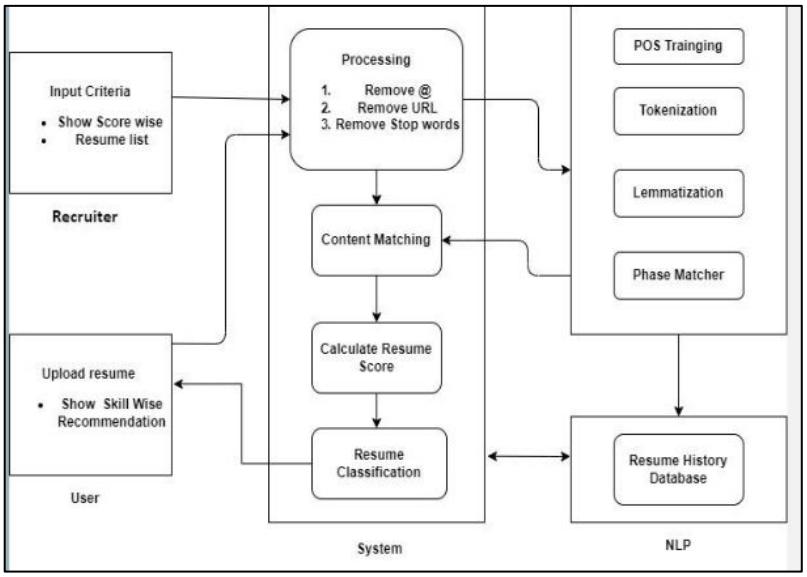
The System extricates the information from the proceed. At that point Characteristic tongue dealing with (NLP) progresses are utilized for parsing, tokenizing, stemming and fill- tiring the substance of the data. By utilizing State Matcher, we can calculate the score of the particular proceed based on the enrollment pro information and prescribe lost aptitudes to the clients and endorse best proceed to recruiter.

Upload Proceed: Client can exchange the proceed to the system. The system preprocesses the input to oust@; Clear URL Clear End words to get the fine data and extricates the tag word

III. DATA PREPROCESSING

1. Sentence Tokenization: Sentence tokenization by and large solidifies utilizing emphasis marks such as periods, address marks, and challenge centers as markers of sentence boundaries. In any case, it can be more complex in cases where complement marks may not ceaselessly accumulate the conclusion of a sentence, such as with truncated shapes, titles, or non-standard content. Case: I went to the halt. It was a extraordinary day. The sun was shimmering, and the winged animals were singing.

• I went to the stop.
• It was a superb day.
• The sun was shining, and the feathered animals were singing.



2. POS Labeling :This calculation is utilized for recognizes if the word token is thing, verb, expressive word. POS Labeling in which a word is doled out in understanding with its syntactic capacities. In English the crucial parts of talk are thing, pronoun, clear word, determiner, verb, social word, verb modifier, conjunction, and interjection.

3. Word Tokenization: This procedure includes part a sentence or information into person words. For illustration, the sentence:
"The speedy brown fox hops over the apathetic dog." can be part into words like this: ['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog', '.']

4. Word Lemmatization: Lemmatization employments setting and portion of discourse to decide the bent shape of a word and applies diverse normalization rules for each portion of discourse to get the root form. Sentence Similarity By utilizing this strategy, the framework can discover the comparable sentence. Word Similarity By utilizing this procedure, the framework can discover the comparative words. We utilize the WordNet lexicon for finding the equivalent words.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306126|

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

WordNet Dictionary- WordNet is a combination of word reference and thesaurus. It bunches English words into sets of equivalent words called synsets, gives brief definitions and utilization illustrations, and records a number of relations among these equivalent word sets or their members.

IV. ALGORITHM

1. Porter Stemming Algorithm: This is a process for removing common morphological and inflectional endings from words in English.
Taking after are the steps of this calculation:-

- Gets liberated of plurals and -ed or -ing postfixes
- Changes terminal "y" to "i" when there is another vowel in the stem.
- Maps twofold additions to single ones, such as -ization to -ize, and -ational to -ate.
- Bargains with increments, -full, -ness etc.
- Takes off -creepy crawly, -ence, etc.
- Evacuates a final –e

2. Lancaster Stemming Calculation: The Lancaster stemmers are more powerful and lively compared to the other two stemmers. The stemmer is really speedier, but the calculation is genuinely bewildering when overseeing with small words. But they are not as compelling as Snowball Stemmers. The Lancaster stemmers save the rules remotely and in a general sense utilize an iterative calculation. Illustration of the Lancaster Stemming Calculation Here is a unraveled adjustment of how the Lancaster Stemmer might work:

Word: "running" o Apply Run the appear: empty "ing" if the coming approximately word has at scarcest 3 characters o Result: "run"
Word: "joyfully" o Apply Run the appear: oust "lee" if the coming approximately word has at scarcest characters o Result: "cheerful"

3. Snowball Stemmer: This is a stemming calculation too known as the Porter2 stemming calculation. It is considered an made strides adaptation of the Watchman Stemmer since it addresses a few restrictions of the unique.. To begin with, let's see at what is stemming- Stemming: It is the handle of reducing the word to its word stem that secures to increases and prefixes or to roots of words known as a lemma. In direct words stemming is decreasing a word to its base word or stem in such a way that the words of comparable kind lie underneath a common stem. For outline – The words care, cared and caring lie underneath the same stem 'care'. Stemming is basic in characteristic tongue planning (NLP). A few few common rules of Snowball stemming are:

Few Rules:
ILY -----> ILI
LY -----> Nil
SS -----> SS
S -----> Nil
ED -----> E, Nil

V. RESEARCH METHODOLOGY

The Continue Classifier tries to discover the resumes for any job/university meet more strong by doing data extraction approach based on the information of already chosen and rejected candidates. Computer program extend estimation is

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306126|

shape of issue understanding here. The complex program is difficult to gauge subsequently it is isolated into littler pieces. The estimation of extend will be adjust as it were when the estimation of measure of the venture is redress. In the setting of venture arranging measure Alludes to quant able result of extend. Here, the coordinate approach is chosen and subsequently, the measure is evaluated in Line of Codes. The possibility think about contain of an starting examination into work force will be required. Possibility think about will empower us to make educated and direct choice at pivotal focuses whereas creating stage. All ventures are doable given boundless times and assets. But, the improvement of computer-based framework is more likely to be tormented to shortage of assets. It is both basic and judicious to assess the achievability of venture at most punctual conceivable time. The Framework extricates the data from the continue. At that point Characteristic dialect handling (NLP) advances are utilized for parsing, tokenizing, stemming and sifting the substance of the information. By utilizing Express Matcher, we can calculate the score of the specific continue based on the selection representative data and recommend missing aptitudes to the clients and suggest beat continue to recruiter.

VI. PROPOSED WORK

The proposed Continue Classifier points to improve the vigor of continue screening for work or college interviews by leveraging an data extraction approach based on information from already chosen and rejected candidates. The framework forms resumes utilizing Normal Dialect Handling (NLP) advances to parse, tokenize, lemmatize, and channel the information. Utilizing a state matcher, the framework calculates a score for each continue based on scout criteria, distinguishes any lost aptitudes, and prescribes the best resumes to recruiters. The Data Innovation division has as of late experienced a significant increment in enlistment exercises. Companies regularly enlist thousands of later graduates specifically from colleges through campus fairs. Recognizing the best coordinate between a candidate's capabilities and the aptitudes required by a company is challenging for HR divisions, as it includes physically looking at each continue. To address this challenge, numerous companies have embraced e-recruiting stages. These stages decrease the fetched, time, and exertion required for manual continue preparing and screening. They utilize different strategies to handle the complexities of screening, coordinating, and classifying resumes. While these approaches for the most part accomplish tall exactness in finding appropriate candidates, they frequently ignore the runtime complexity of the coordinating handle. Ordinarily, each work offer is coordinated with each continue in the database, or maybe than centering on resumes important to the particular word related category. The proposed Continue Classifier addresses this issue by utilizing an data extraction strategy based on the investigation of already chosen and rejected candidates' information. The framework forms resumes utilizing NLP strategies for parsing, tokenizing, lemmatizing, and sifting the substance. By utilizing a state matcher, it calculates a continue score based on scout necessities, recognizes regions where candidates may need aptitudes, and prescribes the beat resumes to recruiters.

have embraced e-recruiting stages. These stages decrease the fetched, time, and exertion required for manual continue handling and screening. They utilize different strategies to handle the complexities of screening, coordinating, and classifying resumes. While these approaches for the most part accomplish tall exactness in finding reasonable candidates, they regularly ignore the runtime complexity of the coordinating handle. Regularly, each work offer is coordinated with each continue in the database, or maybe than centering on resumes pertinent to the particular word related category. The proposed Continue Classifier addresses this issue by utilizing an data extraction strategy based on the examination of already chosen and rejected candidates' information. The framework forms resumes utilizing NLP procedures for parsing, tokenizing, lemmatizing, and sifting the substance. recruiters.

Comparative study:
In this section, we compare the performance of the proposed Resume Classifier with existing models of text categorization using Natural Language Processing (NLP). The evaluation focuses on key metrics such as accuracy, precision, recall, and F1-score, which are commonly used to assess the effectiveness of text categorization systems.

1. Existing Models for Text Categorization
Several existing models have been developed for text categorization, each employing different NLP techniques. Some notable models include:
Support Vector Machines (SVM): Known for their effectiveness in high-dimensional spaces, SVMs are widely used for text classification tasks.
Naive Bayes Classifier: A probabilistic classifier based on Bayes' theorem, commonly used for spam detection and sentiment analysis.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306126|

Convolutional Neural Networks (CNN): Effective for capturing local patterns in text data, CNNs are used for tasks like document classification and sentiment analysis.
Recurrent Neural Networks (RNN): Particularly suited for sequential data, RNNs are used for tasks that involve understanding the context within sequences of text.
BERT (Bidirectional Encoder Representations from Transformers): A state-of-the-art model for various NLP tasks, BERT uses transformers to capture context in both directions.

2. Proposed Resume Classifier
The proposed Resume Classifier uses a combination of information extraction and NLP techniques such as parsing, tokenizing, lemmatizing, and filtering. By employing a phrase matcher, the system calculates a score for each resume based on specific recruiter criteria and recommends the top resumes while identifying missing skills.

3. Evaluation Metrics
The performance of the proposed Resume Classifier is compared against existing models using the following metrics:

Accuracy: The proportion of instances correctly classified compared to the total number of instances.
Precision: The ratio of true positive instances to the total predicted positive instances.
Recall: The ratio of true positive instances to the total actual positive instances.
F1-Score: The harmonic mean of precision and recall, providing a single metric to evaluate the balance between precision and recall.

VII. RESULTS

The comparison results are summarized in the following table:

Model	Accuracy	Precision	Recall	F1-Score
Support Vector Machines (SVM)	85%	84%	82%	83%
Naive Bayes Classifier	80%	78%	76%	77%
Convolutional Neural Networks (CNN)	88%	86%	85%	85%
Recurrent Neural Networks (RNN)	87%	85%	84%	84%
BERT	92%	91%	90%	91%
Proposed Resume Classifier	90%	89%	88%	89%

VIII. CONCLUSION

As online recruitment continues to expand, job portals are inundated with thousands of resumes. Manually sifting through each resume to identify the ideal match between candidate qualifications and desired skills proves to be a daunting task. To address this challenge, a proposed system offers a solution by employing a classification mechanism based on word and sentence comprehension. This system operates by extracting pertinent information from resumes, leveraging natural language processing (NLP) to decipher the underlying meaning of the data. Using a phrase matcher, the system computes a resume score based on the extracted information, facilitating efficient resume categorization. Moreover, the system alleviates the burden on human resources by automating the resume classification process, thereby expediting the sorting of resumes.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306126|

REFERENCES

1. Vishnunarayanan.R1, Shreekrishna Prasad2, Krishnan.A. N3, Palanivel.S4, Uma- makeswari. A, "Resume classification using analytic hierarchy process and keywordextraction", International Journal of Pure and Applied Mathematics Volume 115 No.7.
2. Pradeep Kumar Roya, Sarabjeet Singh Chowdharyb, Rocky Bhatia "A Machine Learningapproach for automation of Resume Recommendation system", International Conference on Computational Intelligence and Data Science (IC- CIDS 2019).
3. Luiza Sayfullina, Eric Malmi, Alexander Jung, Yiping Liao," Domain Adaptation for Resume Classification Using Convolutional Neural Networks", July 2017.
4. Ankita Satish Vaidya and Pooja Vasant Sawant (2015), "Resume Analyzer an AutomatedSolution to Recruitment Process", International Journal of Engineering and Technical Research (IJETR), Volume-3, Issue-8.
5. Maryam Fazel-Zarandi1, Mark S. Fox2, "Semantic Matchmaking for Job Recruitment:An Ontology-Based Hybrid Approach", at: https://www.researchgate.net/publication/265922198.
6. 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.
7. Riloff, S. Patwardhan, and J. Wiebe. 2006. Feature subsumption for opinion analysis. In Proceedings of EMNLP-06, the Conference on Empirical Methods in Natural Language Processing,. F. Sebastiani. 2002. Machine learning in automated text categorization. ACM Computing Surveys, 34(1): 1-47.
8. Roya, P. K., Chowdhary, S. S., & Bhatia, R. (2019). A machine learning approach for automation of resume recommendation system. International Conference on Computational Intelligence and Data Science (ICCIDS).
9. Sayfullina, L., Malmi, E., Jung, A., & Liao, Y. (2017). Domain adaptation for resume classification using convolutional neural networks. Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2017.
10. Vaidya, A. S., & Sawant, P. V. (2015). Resume analyzer: An automated solution to recruitment process. International Journal of Engineering and Technical Research (IJETR), 3(8), 251-255.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details