Resume Parser with Natural Language Processing

Hamna Qaseem [1] Manahil Ahmad [2] Fiza Asghar [2] Samreen Jamil [2]

[1]Department of Data Science [1]The Islamia University of Bahawalpur

Abstract

Our project utilizes Natural Language Processing (NLP) techniques to parse information from PDF-formatted resumes. Through data collection, pre-processing, and exploratory analysis, we categorize resumes based on topics. Experimenting with various models, we evaluate their performance for classification. The results demonstrate the effectiveness of deep learning algorithms in improving resume parsing accuracy compared to traditional machine learning. This project showcases the successful application of NLP techniques for extracting meaningful information from resumes.

Literature review

Resume parsing techniques has gained significant attention in recent years. Different research papers have explored various approaches and methodology to extract relevant information from resumes. Following are few of research papers that we reviewed and gain insight from their work and study:

- 'Resume Parser with Natural Language Processing', Parse information from a resume using natural language processing, find the keywords, cluster them onto sectors based on their keywords and lastly show the most relevant resume to the employer based on keyword matching.[3].
- This paper aims to solve these issues by automatically suggesting the most appropriate candidates according to the given job description. Our system uses Natural Language Processing to extract relevant information like skills, education, experience, etc. from the unstructured resumes and hence creates a summarised form of each application. [4].

Project objectives

The present study investigates the following objectives:

- Objective 1: The objective of this project is to develop a Resume Parser using Python programming language.The specific objectives include data acquisition, exploratory data analysis, text pre-processing, topic modeling, classification, and future enhancements such as database integration and the creation of a streamlit app.
- Objective 2: The project aims to provide an efficient and effective solution for automating the recruitment process.

Study methodology

The present study adopted the following step-by-step methodology to achieve the project objectives.

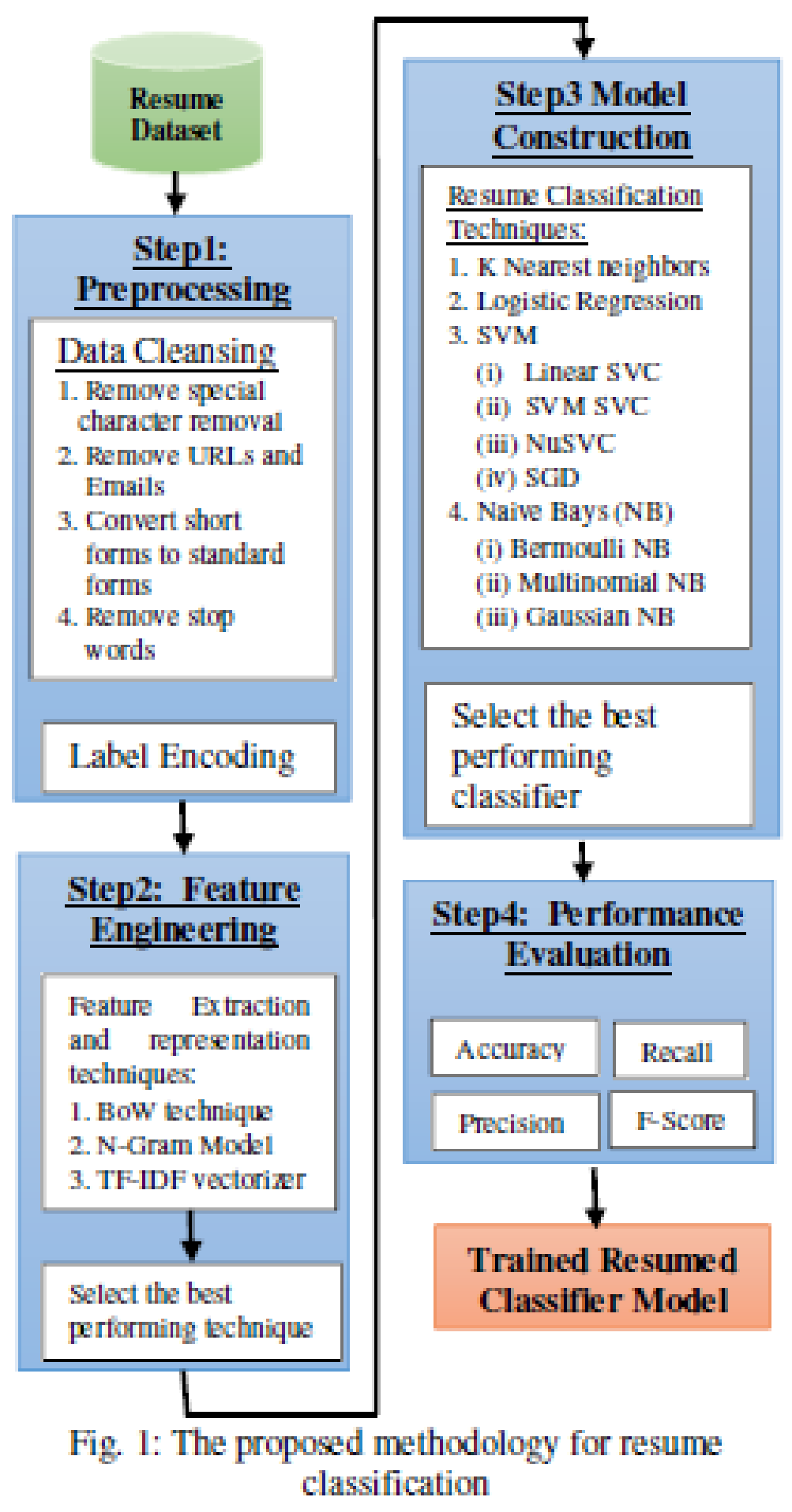

Resume Dataset
Step1: Preprocessing
Data Cleansing
1. Remove special character removal
2. Remove URLs and Emails
3. Convert short forms to standard forms
4. Remove stop words
Label Encoding
Step2: Feature Engineering
Feature Extraction and representation techniques:
1. BoW technique
2. N-Gram Model
3. TF-IDF vectorizer
Select the best performing technique
Step3 Model Construction
Resume Classification Techniques:
1. K Nearest neighbors
2. Logistic Regression
3. SVM
(i) Linear SVC
(ii) SVM SVC
(iii) NuSVC
(iv) SGD
4. Naive Bays (NB)
(i) Bermoulli NB
(ii) Multinomial NB
(iii) Gaussian NB
Select the best performing classifier
Step4: Performance Evaluation
Accuracy
Recall
Precision
F-Score
Trained Resumed Classifier Model
Fig. 1: The proposed methodology for resume classification

Figure 1. Methodology

Site selection and data collection

The resume datasets were collected from different sources such as kaggle[2] and Github[1] . The dataset is in Pdfs file format. The number of resume instance for each class job category. After gaining access to the dataset, the resume PDF files are downloaded from the available sources. The number of files and the size of the dataset may vary depending on the sources. It is important to ensure that the downloaded files are representative of the target population and cover a diverse range of industries, job roles, and experience levels.

Most Frequent Words

- The resume data is cleaned by removing any unnecessary elements such as punctuation, numbers, and common stop words (e.g., "the," "and," "is") that do not contribute much to the overall meaning.
- The frequency of each word in the cleaned text is calculated, capturing how often each word appears.
- The word cloud is generated by representing each word as a "cloud" element, where the size of the word corresponds to its frequency or importance.
- Identifying key topics or themes in our resume dataset.

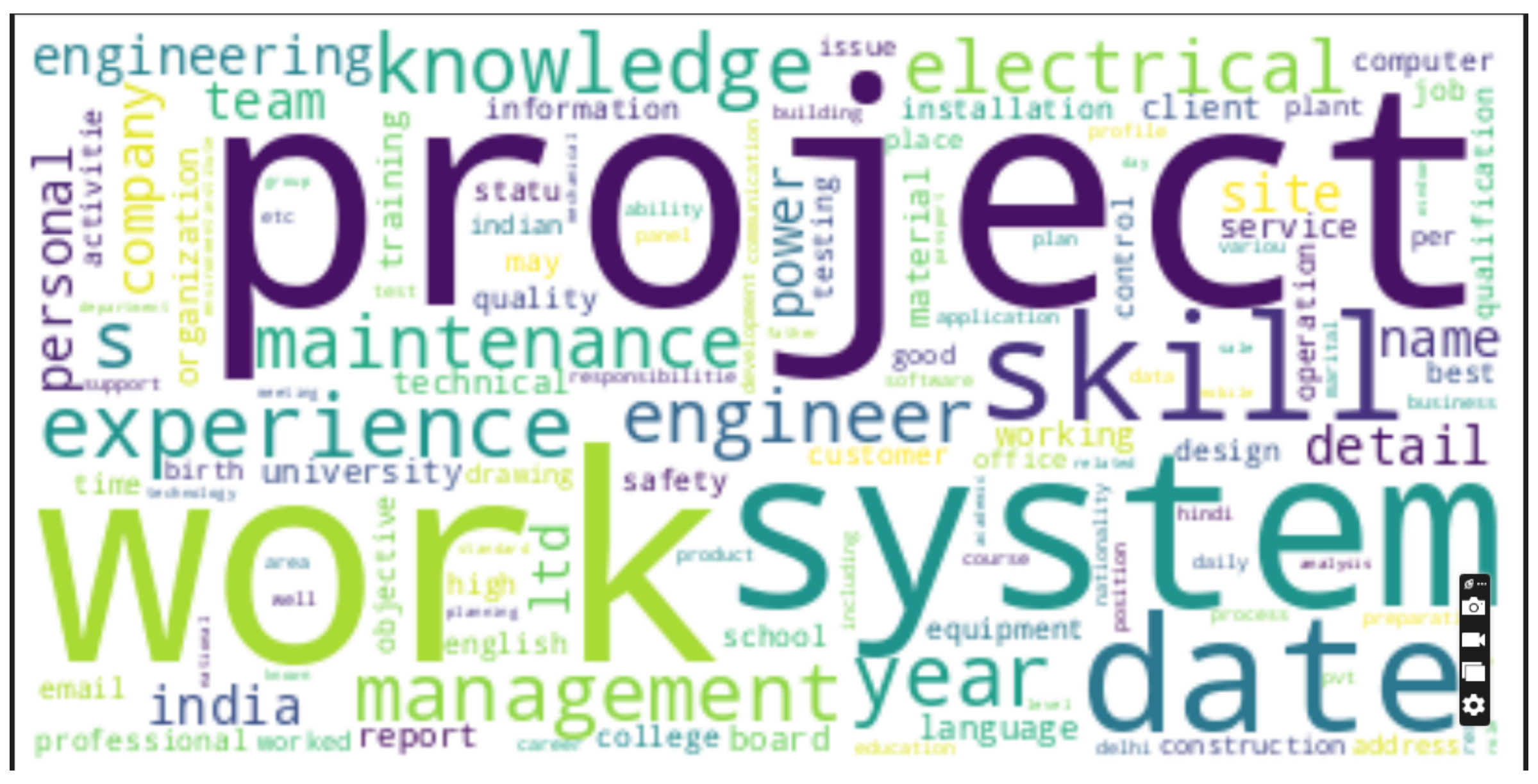


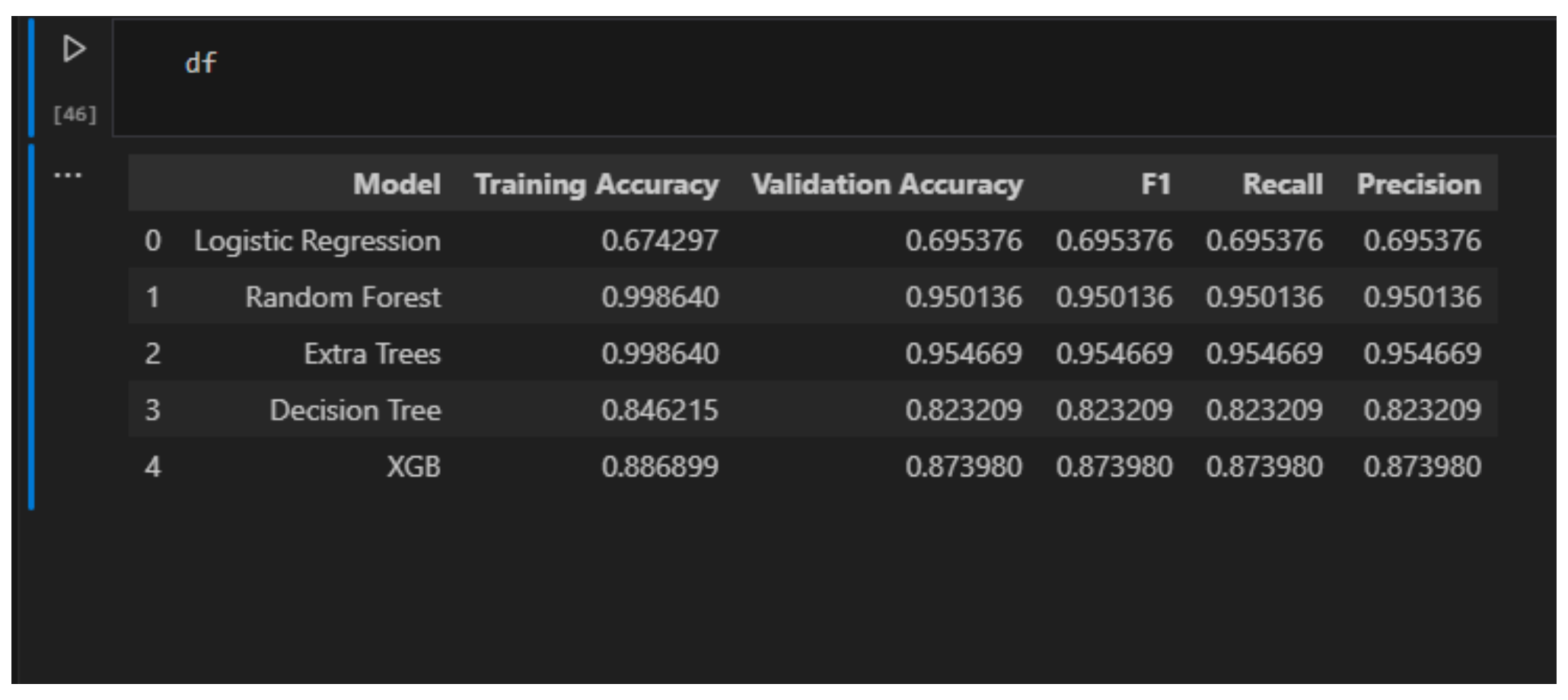
Figure 2. Word cloud

Results and discussion

Model estimates

In the Resume Parser project, several machine learning models were implemented and evaluated for their performance. So in our project we applied several machine learning algorithms. Following are our implemented models discussed with accuracy rate:

Overall, the Random Forest and Extra Trees models demonstrated the highest performance with high accuracies, F1 scores, recall, and precision. These models are suitable for the Resume Parser project as they can effectively classify the resume data and make accurate predictions.



```
df
[46]
                 Model  Training Accuracy  Validation Accuracy        F1    Recall  Precision
0  Logistic Regression           0.674297             0.695376  0.695376  0.695376   0.695376
1        Random Forest           0.998640             0.950136  0.950136  0.950136   0.950136
2          Extra Trees           0.998640             0.954669  0.954669  0.954669   0.954669
3        Decision Tree           0.846215             0.823209  0.823209  0.823209   0.823209
4                  XGB           0.886899             0.873980  0.873980  0.873980   0.873980
```

Figure 3. Machine learning Model performance

So after training and testing The reported test accuracy of 96-percent suggests that the BERT model achieved a high level of accuracy in classifying the samples in dataset. The test accuracy is calculated by dividing the total number of correctly predicted samples by the total number of samples in the test set.

Table 1. Deep learning Model Performance

Col 1	BERT Model Performance
Metric	Value
Training Loss	0.07218202252145114
Test Accuracy	0.9646579066606252

Conclusions

- When comparing the ML models' results to the BERT model's performance, it is evident that the BERT model outperformed the traditional ML models in terms of test accuracy. The BERT model achieved a significantly higher accuracy of 96-percent, while the ML models ranged from 69.54 to 95.47.Overall, the results and comparison indicate that BERT, with its advanced deep learning capabilities, offers significant advantages over traditional ML models for resume parsing and classification. Further analysis, including evaluation metrics such as precision, recall, and F1 score, can provide deeper insights into the models' performance across different classes and help identify any potential trade-offs between precision and recall.

Future Work to focus On?

- The integration of a database to store the parsed resume data. This will enable efficient data management, retrieval, and scalability asthe volume of resumes increases.
- The user-friendly Streamlit application can be developed to allow candidates to easily upload their resumes and receive instant feedback on the parsed information.
- Integration of a chatbot into the application The chatbot can serve as a virtual assistant, providing guidance to candidates, answering their queries, and offering personalized feedback based on the analysis performed by the resume parser.

What does this study add?

- Working on this project that involves Natural Language Processing enhances our understanding of this subfield of artificial intelligence. We gain hands-on experience in text preprocessing, feature extraction, and machine learning models specific to NLP tasks. This project enriches our knowledge beyond what is covered in our curriculum.
- Throughout the project, we acquire and refine various technical skills. These include data collection and preprocessing, implementing machine learning algorithms, working with NLP libraries and tools, and evaluating model performance. These skills are highly sought after in the field of data science and can strengthen our profile for future academic or professional opportunities.
- As we encounter challenges during the project, we developed problem-solving and critical thinking skills. We learn to analyze problems, devise strategies, and experiment with different approaches to overcome obstacles.

Practical implications

- Data Collection, preprocessing
- Parsing and Information extraction
- Machine learning and deep learning algorithms

References

[1] 2021.

[2] Pdfs resume dataset.
2021.

[3] CH Ayishathahira, C Sreejith, and C Raseek.
Combination of neural networks and conditional random fields for efficient resume parsing.
In *2018 International CET Conference on Control, Communication, and Computing (IC4)*, pages 388–393. IEEE, 2018.

[4] Nirali Bhaliya, Jay Gandhi, and Dheeraj Kumar Singh.
Nlp based extraction of relevant resume using machine learning.
2020.

Visit our[1] project here

Scan me!

Scan this QR code