STM JOURNALS

International Journal of
Electronics Automation

Volume 1, Issue 2, 2023
July–December
DOI (Journal): 10.37591/IJEA

https://journals.stmjournals.com/ijea

Review IJEA

Machine Learning Approaches Towards Resume Classification

Subhajit Das[1,*], Saswati Naskar[2], Swapna Halder[3]

Abstract

Finding the right person for an open position can be an unnerving task, especially when there are many applicants, and if the recruiter or the Human Resource department must sort and further categorize all those resumes then it will be a labor intensive, time-consuming and tiresome task. Additionally, human assessment of resumes may be biased and prone to mistakes. Manually screening the proper candidate's resume from the pool is not practicable; instead, an automated method using natural language processing may assist in selecting the appropriate candidate's resume. The proposed web application is built so that job applicants and recruiters can use it easily for applying for job opportunities and screening, respectively. This study aims to revolutionize resume classification through the lens of machine learning. Unlike previous approaches, we provide a unique framework that combines advanced machine learning algorithms with complex feature engineering techniques. So, this can hamper the decision-making process and lead to delays in the hiring process. We examine the model architectures, assessment measures, feature extraction strategies, and underlying processes that went into creating and developing these systems. We also look at case studies and real-world applications to show the difficulties and benefits of using machine learning for resume screening. Consequently, the general expansion of the business will be impeded. An efficient resume classification process will ease the hiring process. Machine learning algorithms like Regular Expression, Multinomial Naïve Bayes, Logistic Regression, Random Forest, SVM etc. have been used for Resume Classification.

Keywords: Artificial Intelligence, Machine Learning, Logistic Regression, SVM, Naïve Bayes.

INTRODUCTION

Nowadays, job recruitment over the internet is more common and helpful to both businesses and employees. Traditional resume screening approaches, such as keyword searches or manual review, are typically subjective, inefficient, and biased. Having the option to submit a formal application without attending an interview saves time for both recruiters and applicants. But these days, there are a lot of resume templates online. Hiring brilliant minds for an organization is a difficult task. And as the Indian economy is booming, young and energetic workforce will be required in different organization to support the economic activities [1]. The actual problem arises here, finding a qualified candidate amongst thousands of people is a tedious task. And the person who is responsible for this task must have a sound knowledge of various domains to find appropriate candidate and this is not imaginable in real world scenarios as numerous job roles exists in today's world, so it will be more and more difficult for a human being to manually screen all those resumes [2]. Machine learning algorithms empower us to solve such kind of problems [3, 4]. It is

***Author for Correspondence**
Subhajit Das
E-mail: subhajitdas2512@gmail.com

[1]Student, Department of Computer Science & Engineering, Greater Kolkata College of Engineering and Management, India
[2]Student, Department of Computer Science & Engineering, Greater Kolkata College of Engineering and Management, India
[3]Student, Department of Computer Science & Engineering, Greater Kolkata College of Engineering and Management, India

Received Date: March 30, 2024
Accepted Date: April 15, 2024
Published Date: April 25, 2024

Citation: Subhajit Das, Saswati Naskar, Swapna Halder. Machine Learning Approaches Towards Resume Classification. International Journal of Electronics Automation. 2023; 1(2): 1–7p.

possible to train the model with the help of huge number of datasets, and the model can learn from previous results and can predict reliable outcomes [5].

Most proposed approaches to text categorization aim to enhance classifier accuracy. Text categorization involves categorizing materials based on their wording. One approach is to classify documents using machine learning, treating the absence of words as a logical property, like an early statistical model of language. The Multivariate Bernoulli Naive Bayes (BNB) model is completed. This research used data from job candidates' resumes, primarily sourced from LinkedIn. This study used 250 resumes and experiments as its primary research methodology. We classified resumes into three categories: employable (ER), waiting (WR), and not employable (NER).

In this study, we provide a complete review of the current resume classification landscape, emphasizing the limitations of existing methods. Advanced HR strategies like hiring, screening, training, and performance reviews are used by organizations. Online recruitment improves applicant information collecting, provides uniform data, and saves time.

We then provide our suggested framework, describing its main components such as deep learning architectures, NLP approaches, and ensemble learning algorithms. We hope to transform how recruiters analyze applications by using the power of artificial intelligence, resulting in better informed hiring decisions, and driving organizational success in the competitive global marketplace. Machine learning techniques can be used to solve real-world problems, including online job applications, job recommendation systems, sentiment analysis, and spam filtering [7–9].

RELATED WORKS

1. Machine learning techniques have recently been widely used in many different fields. Many studies are focusing on utilizing machine learning approaches to categorize job posts in today's technology-driven world.
2. The resume classifier is a text-based categorization system. Different approaches to representing text based on the syntactic and semantic relationships between words are offered by Sam Scott and Stan Matwin. Various machine learning models, including SVM and GBDT, describe feature creation techniques.
3. Since the introduction of deep learning, there has been a shift toward using neural network architectures for resume classification problems. It has been demonstrated that Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can identify sequential and hierarchical patterns in textual input. Zhang et al. (2019) suggested a CNN-based strategy for resume categorization, which outperformed existing methods. Similarly, Liu et al. (2020) investigated the use of bidirectional Long Short-Term Memory (BiLSTM) networks for resume categorization and found gains in classification accuracy.
4. Human resource professionals encounter numerous obstacles during the recruitment process, making artificial intelligence a viable solution. There hasn't been much research done in this field, though. Data for the study came from a variety of sources, including Statista, Tractica, MIT Sloan Management Review, LinkedIn, Boston Consulting Group, and CV Library [6].
5. With the advent of fifth-generation sophisticated technology, spam messages have become a major issue, and a variety of email programs have grown in number. Different spam filtering techniques exist. Historically, the probability-based Bayesian classification system has been used to efficiently filter spam messages and prevent major issues [10–11]. Bayesian email spam filtering currently has a 90% accuracy rate.
6. There are huge applications used in medicine for predicting diseases. The Naïve Bayes classification approach can predict illnesses based on hemoglobin protein sequences. Using data mining in protein analysis allows for efficient identification of protein properties, leading to improved medication design. In the investigation, Naïve Bayes achieved an accuracy of 85% [8].

7. These methods typically examine training data (i.e., pair data with predetermined input-output) to generate an inferred function that may be used to map other examples. Unsupervised categorization, on the other hand, groups documents based on their similarity rather than using a specified criterion. Document classification techniques include Naïve Bayes, TF-IDF, SVM, KNN, and Decision Tree.

METHODOLOGY AND PRE-PROCESSING

This section provides data pre-processing experimental setups and various methods and technics have been used to classify and categorize resumes accordingly. The model works as described below.

Gathering Data from Various Resources

This is the first and foremost step to build this Machine Learning (ML) model as we cannot build and train the model without relevant data. So, datasets have been downloaded from kaggle.com. And another thing is that these datasets are unstructured. So, the datasets need to be cleaned before we can use that the dataset has the following category wise and percentage distribution is shown in Figures 1,2. Distribution of categories throughout a range of fields, including sales, consulting, chief, finance, healthcare fitness, business development, advocacy, design, information technology, and human resources, among others.

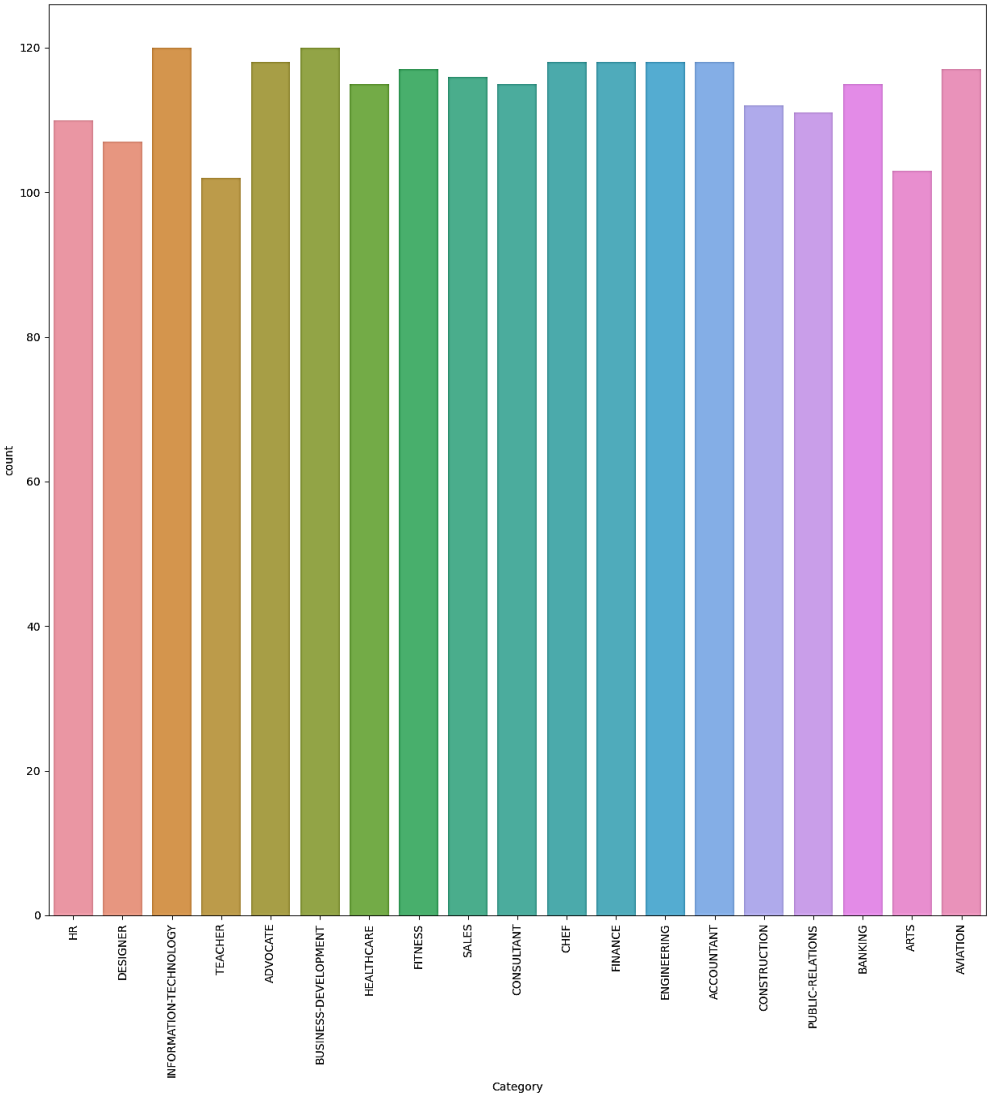


Figure 1. Category Distribution in Various Domain.

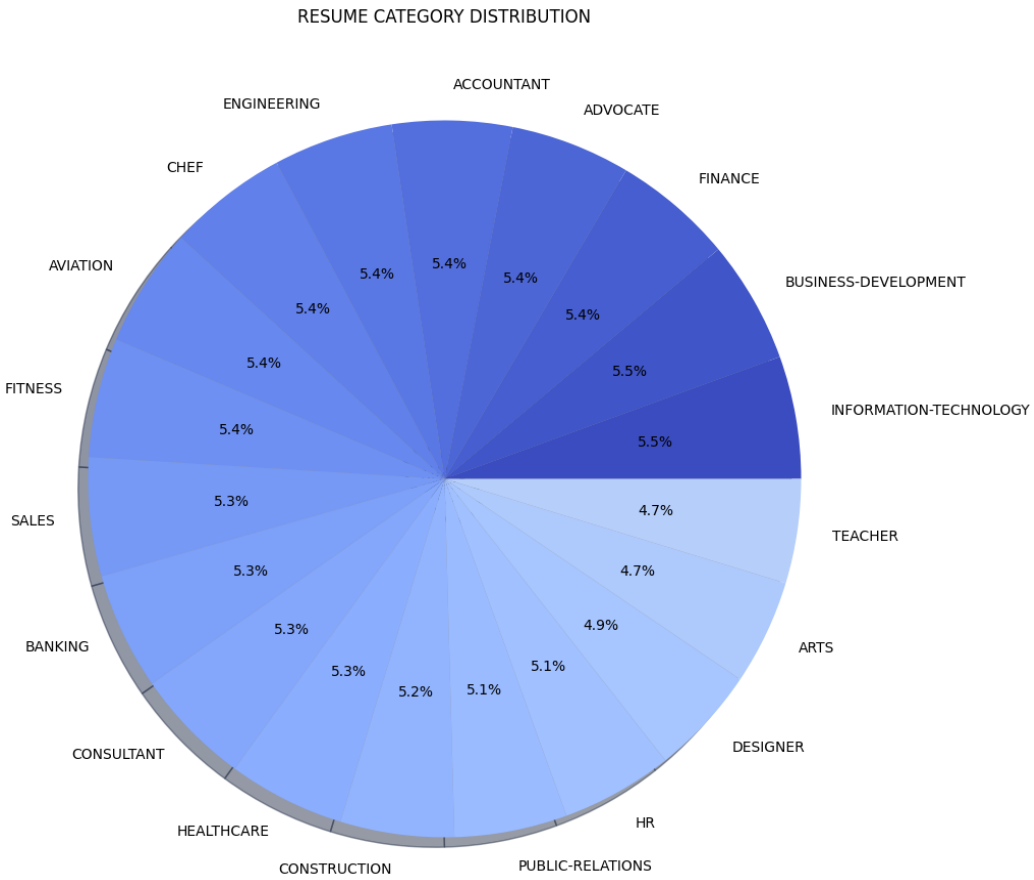
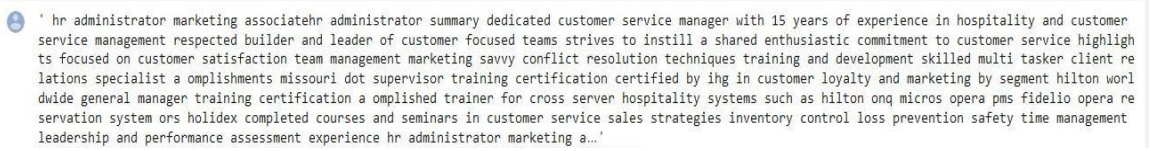


Figure 2. Percentage Wise Distribution.



```
' hr administrator marketing associatehr administrator summary dedicated customer service manager with 15 years of experience in hospitality and customer
service management respected builder and leader of customer focused teams strives to instill a shared enthusiastic commitment to customer service highligh
ts focused on customer satisfaction team management marketing savvy conflict resolution techniques training and development skilled multi tasker client re
lations specialist a omplishments missouri dot supervisor training certification certified by ihg in customer loyalty and marketing by segment hilton worl
dwide general manager training certification a omplished trainer for cross server hospitality systems such as hilton onq micros opera pms fidelio opera re
servation system ors holidex completed courses and seminars in customer service sales strategies inventory control loss prevention safety time management
leadership and performance assessment experience hr administrator marketing a...'
```

Figure 3. datasets.

Data Cleaning

These kinds of dataset are usually filled with huge number of irrelevant data. So, to clean the dataset, blank spaces, special characters, tags, links etc. must be removed from the dataset is shown in Figure 3.

Also, all the characters have been converted to lowercase. And the cleaned data stored in a separate column.

Lemmatization

Lemmatization is the process of putting together various inflected versions of a single word. Its applications include computational linguistics, natural language processing (NLP), and chatbots. Lemmatization combines related meaning terms into a single word, making chatbots and search engine inquiries more effective and accurate. The purpose of lemmatization is to reduce a word to its root form, often known as a lemma.

Lemmatization transforms words into the root word and the meaning remains intact. For example, “playing” will be converted into “play” after lemmatization. Words with maximum frequency can be visualized and understood with the help of word cloud is shown in Figure 4.

Figure 4. Word Cloud.

Label Encoding and Feature Extraction Using TF-IDF

The cleaned data stored into a separate column after preprocessing. Then, texts from category have been converted into numerical values and stored into a different column named "Category Labeled", which can be treated as target variable. TF is used to check the frequency of a word in a text. In IDF, the words which are more common are less significant. TF-IDF is used to test relevancy of a word in a body to a text. So, the word with the higher TF_IDF score is most significant. The purpose is to find the most relevant words related to a job role.

The TF-IDF method efficiently and simply matches query words to relevant documents. TF-IDF finds documents that are very relevant to a query, according to data analysis. When a user searches for a particular subject, TF-IDF can find relevant documents.

TF-IDF is simple to encode and can serve as the foundation for complex algorithms and retrieval systems [3].

RESULT ANALYSIS

The various Machine learning estimators are used to train the model such as Multinomial Naïve bayes, Logistic Regression, Support Vector Machine.

Multinomial Naïve Bayes (MNB)

Multinomial NB is a "probabilistic classifier" that works on probabilistic output, and it is appropriate for discrete features. Then the dataset is fragmented into train and test set and the training set which are fitted into the MNB classifier. And training set is used to estimate the enactment of the model and the accuracy score, precision, recall and f1 score has been attached. Figure 5. Shows the accuracy of MNB.

Logistic Regression (LR)

LR is a statistical method, and it is used to train ML models where the dependent variables are binary. This model gave an accuracy of around 68% which is better than MNB in this case. Figure 6. Shows the Accuracy of LR.

Logistic regression is a useful analysis tool for classification problems, which includes determining whether a new sample fits well into a category. Because components of cyber security involve classification problems, such as attack detection, logistic regression is an effective analytic tool.

Support Vector Machine (SVM)

SVM distinctly classifies data points by a hyperplane. Figure 6 and Figure 7. shows the accuracy of

Linear SVM.

Classification report for classifier MultinomialNB():

	precision	recall	f1-score	support
0	0.53	0.94	0.68	17
1	0.52	0.48	0.50	27
2	0.50	0.21	0.30	19
3	0.88	0.58	0.70	24
4	0.79	0.44	0.57	34
5	0.57	0.74	0.64	23
6	0.73	0.80	0.76	20
7	0.83	0.83	0.83	24
8	0.11	0.05	0.07	21
9	0.83	0.50	0.62	20
10	0.56	0.72	0.63	25
11	0.62	0.42	0.50	24
12	0.76	0.62	0.68	21
13	0.36	0.42	0.39	19
14	0.63	0.81	0.71	21
15	0.57	0.79	0.67	29
16	0.67	0.70	0.68	23
17	0.44	0.79	0.57	19
18	0.63	0.71	0.67	24
accuracy			0.61	434
macro avg	0.61	0.61	0.59	434
weighted avg	0.62	0.61	0.59	434

Figure 5. Accuracy of MNB.

```
0.684331797235023
```

Figure 6. Accuracy of LR.

```
0.6658986175115207
```

Figure 7. Accuracy of Linear SVM.

Table 1. Accuracy score in different classifier.

Classifier	Accuracy
Multinomial Naïve Bayes	0.62
Logistic Regression	0.684331797235023
Linear SVM	0.6658986175115207

Support vector machines (SVMs) are training algorithms that teach regression and classification rules from data. For instance, they can learn multilayer perceptron classifiers, polynomial, and radial basis functions. Vapnik first proposed SVMs for classification in the 1960s, and since then, there has been a lot of research on the topic thanks to advances in theory and practice as well as applications to regression and density estimation.

The Accuracy scores of Multinomial Naïve Bayes, Logistic Regression and Linear SVM are provided below in Table 1.

CONCLUSION

Resume screening for job opportunities is influenced by a variety of elements during the process. This procedure selects student as per their qualifications and talents. The analysis of Three models

showed that the individual Multinomial Naïve bayes gives probabilistic output, Logistic regression gave accuracy around 60% which is better than MNB. The accuracy of Linear SVM model is better than MNB and LR. The resume classification process based on machine learning approaches shows promising results in task automation. Further advancements can be made by integrating more data sources and refining algorithms to improve accuracy and efficiency. In summary, this research work demonstrates the machine learning approaches to transform the recruitment process and make it more effective and efficient.

REFERENCES

1. Balci B, Saadati D, Shiferaw D Handwritten text recognition using deep learning. 1–8p.
2. Kim SW, Gil J-M Research paper classification systems based on TF-IDF and LDA schemes. Human-centric Computing and Information Sciences 2019; 9(30): 1–21p.
3. Berger, A & Lafferty, J. (1999). Information Retrieval as Statistical Translation. In Proceedings of the 22nd ACM Conference on Research and Development in Information Retrieval (SIGIRí99), 222–229.
4. Friedman C, Hripcsak G Natural language processing and its future in medicine. Academic medicine: Journal of the association of the American Medical Colleges 1999; 74(8): 890–5p.
5. Ramos J Using TF-IDF to determine word relevance in document queries. 1999.
6. Yi X, Allan J, Croft, WB Matching resumes, and jobs based on relevance models. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2007.
7. H. Rodney, K. Valaskova and P. Durana, "The artificial intelligence recruitment process: how technological advancements have reshaped job application and selection practices," Psychosociological Issues in Human Resource Management, vol. 7, no. 1, pp. 2–47, 2019.
8. Mwaro PN, Ogada DK, Cheruiyot W, SCIT J. Applicability of naïve Bayes model for automatic resume classification. International Journal of Computer Applications Technology and Research. 2020;9(9):257–64.
9. Roy PK, Singh SK, Das TK, Tripathy AK. Automated Resume Classification Using Machine Learning. InAdvances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2022 2022 Jul 28 (pp. 307-316). Singapore: Springer Nature Singapore.
10. T. Joachims. Making large-scale support vector machine learning practical. In Advances in Kernel Methods: Support Vector Learning. B. Schölkopf, C.J.C. Burges, and A.J. Smola (Eds.), MIT Press, 1998.
11. F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review, 65:386-408, 1959.

