2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)

Automated Resume Classification System Using Ensemble Learning

Spoorthi M
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
rajoslinc@gmail.com

Meghana Kuppala
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
kuppalameghana3102@yahoo.com

Indu Priya B
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
indu.balappagari@gmail.com

Vaishnavi Sunilkumar Karpe
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
vaishnavi.sunilkarpe@gmail.com

Divya Dharavath
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
divya773181@gmail.com

***Abstract*—One of job recruiters' biggest challenges is selecting a suitable resume from the pool of resumes. For a single role, thousands of candidates send their resumes. Manually selecting the resume from a large number of applicants and assigning them suitable positions is time taking and not feasible. An automated system can make this process easy and efficient. This system takes candidates' resumes in word, pdf, or any format and classifies them according to the skill set mentioned in the resume. We propose an ensemble deep-learning model to classify the resume.**

Keywords—resume; classification; job; deep learning; ensemble learning; skill set.

I. Introduction

Recruitments in the Information Technology field have been increasing exponentially. Recruiters must properly screen resumes to hire suitable candidates. The process of checking if a candidate is suitable for a particular role according to the information from their CV/Resume is called resume screening. Recruiters have to screen through a large amount of resume data fast and reliably.

The most important and basic tool in any selection process is the candidate's resume. Interviewing has become a time-consuming affair. The number of applications is in the millions, making it time-consuming to sort through them. Here we need a machine learning algorithm that can give a better way of screening and full fill the requirements in the industry.

The world of Artificial Intelligence and Machine Learning has grown significantly. In the discipline of machine learning, a dataset is used to train a model to predict the intended outcome from incoming data. The large amounts of data available have contributed to significant growth in the performance of ML models recently. We can take advantage of this growth in ML for automation and increase productivity in the areas which require manual human labor. We propose a mechanism that allows the recruiters to recruit based on the skill set and information mentioned in the resume of the candidates, by using an ensemble deep-learning model, which classifies the given resume into various categories. This reduces the time to screen a resume manually. Ease of Use.

II. Literature Survey

The research paper "A machine learning approach for Automation of resume recommendation system" describes an algorithm that analyses the characteristics extracted from the resume and categorizes those characteristics according to the job description. The categorized resume is mapped and recommends the candidate who is more suitable for the position. They have built two models. A Classification model was built using several algorithms like the random forest, Multinomial Naïve Bayes, Logistic Regression, and Linear Support Vector Machine Classifiers. Among these models, the

SVM model’s performance was better. The Recommendation model was built on content-based recommendation and K-Nearest Neighbours [1].

The research paper “automated tool for resume classification using semantic analysis” presents the development of a resume classification application. It uses a voting classifier which is based on ensemble learning. It categorizes a candidate’s profile into an appropriate domain in accordance with the work experience and other details given by the applicant in the profile [2].

The research Paper “Resume Classification using various Machine Learning Algorithms” describes a model using Nave Bayes, Random Forest, and SVM, which extracts skills and shows diverse capabilities under appropriate job profile classes. Random Forest gave the best accuracy among the three of them [3].

The research paper “Differential Hiring using a Combination of NER and Word Embedding” describes a methodology using the NLP, Word2Vec which is a pre-trained word embedding layer. Word embedding is a method of expressing words as real-valued vectors that convey their meaning in such a way that words that are adjacent to one another in the vector space are assumed to have identical meanings. This will help to get the most accurate resume according to the skillset provided [4].

The research paper “Resume Screening Using Machine Learning and NLP: A Proposed System” proposes a machine learning model which takes a student's resume and according to skills and other details mentioned in the resume, the model shows suitable job roles and the resume's relevance to the job description [5].

“Resume Ranking based on Job Description using SpaCy NER model” devised a method that lowers hiring costs and speeds up the process of selecting the best candidate for the job role [6].

“Domain Adaptation for Resume Classification Using Convolutional Neural Networks” employs a classifier to categorize resume data after training it on a large number of openly accessible job description excerpts. Despite just having a tiny amount of labeled resume data at their disposal, they empirically confirmed a respectable classification performance of the approach [7].

The research work “A Hybrid Approach to Conceptual Classification and Ranking of Resumes and Their Corresponding Job Posts" presents a hybrid approach using a conceptual-based classification of resumes and a ranking system that ranks the candidates according to the corresponding job offers. They collected 2000 resumes from online sites and 10,000 different job postings for the experiment. They used job titles and skill sets in the classification process. They got higher precision results [8].

In “Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping EXPERT”, EXPERT mapping-based candidate screening, an intelligent ontology tool, was utilized to construct an automated system for the intelligent screening of prospects for recruitment, improving the precision with which candidates are matched to the job criteria [9].

III. Methodology

A. Data Collection and Visualization

The data was collected from Kaggle. The data is in Comma Separated Value(CSV) format, with two columns Category, and Resume. The category column is the resume’s sector or field, and the resume column is the content of the resume. There are 962 resumes in 25 different categories.

B. Data Preprocessing

Data Preprocessing involves converting raw data in the dataset suitable to our task. The information supplied by the Curriculum vitae in this procedure would be cleaned. Unnecessary data would be removed. Then the data would be converted into vectors. The following steps were performed in the data pre-processing of resume data.

1) Data Cleaning:

In the cleaning process, numbers, special characters, and words with single letters are removed. Then we get the cleaned resumes .

2) Tokenization

Tokenization was performed on the resume data using the tokenizer class of TensorFlow.

3) Removal of stop words

Words such as ‘is’, ‘are’, ‘was’ etc are called stop words. They appear in most of the text. Such stop words that do not provide any important information for the classification task are removed from the generated tokens.

English stop words were imported from the NLTK corpus and used for the stop word removal process.

4) Label encoding

Label encoding should be done to assign a numerical label to all categories. The sklearn Label Encoder was used. Fig. 2. shows how after label encoding, the categories are given unique numeric values.

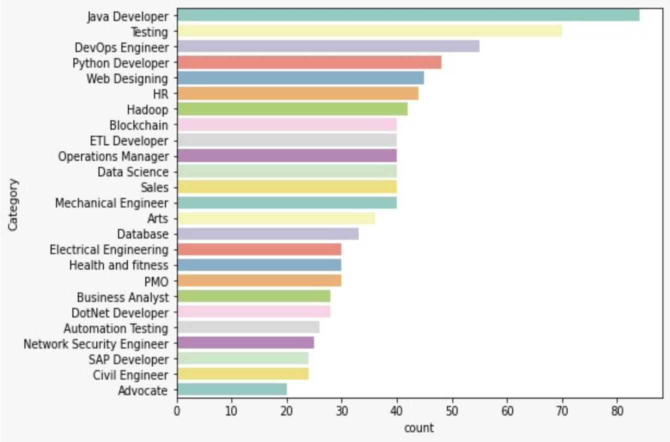


Fig. 1. Number of instances of different domains in the dataset.

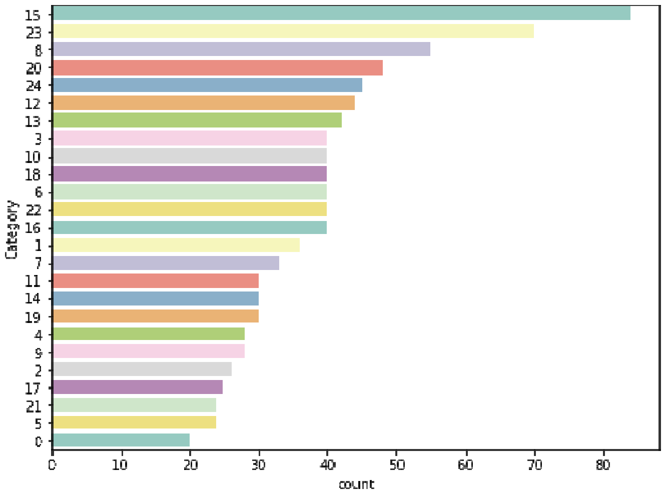


Fig. 2. After label encoding

C. Model Architecture

We created an ensemble model using 1D Convolutional Neural Network (CNN) and Bi-directional Gated Recurrent Unit (GRU), as in Fig. 3. They both act as two channels in our model. Each and every text message is mapped to a reality we used the pre-trained word embeddings trained with a skip-gram model using the 3-billion-word Google News corpus [10]. The input sequences are fed to the embedding layer.

In channel 1, the embedding layer's output is fed into a 1D convolutional layer. The number of filters is 100. The kernel size is 3. The rectified linear unit (ReLU) is used as the activation function. ReLU helps in preventing the exponential computation growth in neural networks. The input feature space is convolved as a result of this. The convolved input feature space is then down-sampled. The Max pooling layer is used to downsample the feature space. The pool size is 2. Each of the dimensions of the output can be considered an 'extracted feature'. Then we used a flattened layer. Then this is fed to a drop-out layer. The drop-out rate is 0.5. It is used to regularize learning and prevent overfitting.

At last, a softmax layer is added.

In channel 2, the output from the embedding layer feeds into a GRU layer. Then the output is fed to a drop-out layer. The drop-out rate is 0.5. It is used to regularize learning and prevent overfitting. At last, a softmax layer is added.

The output from both channels is combined to get the final output.

Fig. 3. Shows the model architecture.

IV. Experiment

We used Keras with the Tensorflow backend. For each dataset, we split it into 80:20. Accuracy and other metrics are shown in tables Table. I, Table. II.

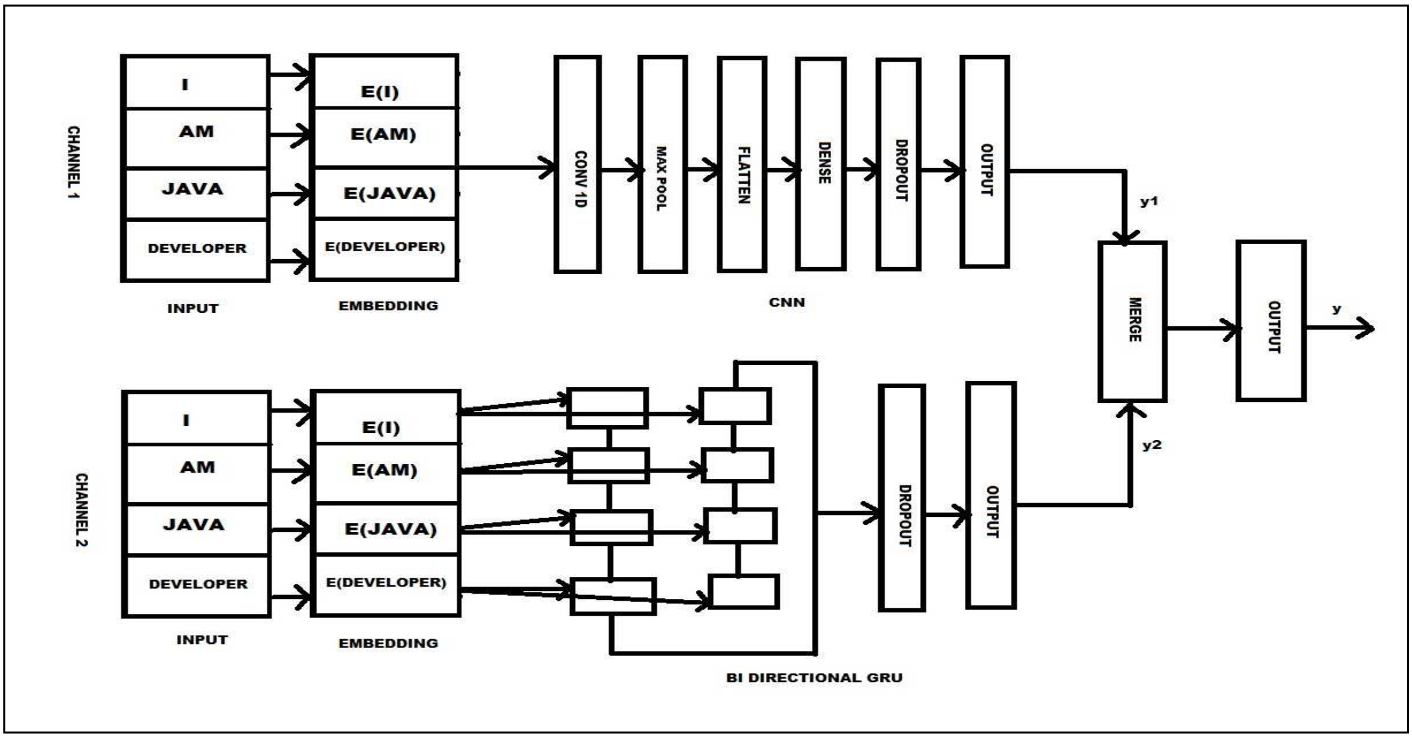


Fig. 3. Model Architecture

V. Result

Table I. gives precision, recall, and F-1 score values.

Table II. gives accuracy.

Fig. 4 is the confusion matrix.

Creating a web app: After training the model we created a web application using streamlit, where the user can upload resumes in .pdf, .docx, and .txt formats. The resume will be classified into a suitable category by clicking the submit button.

TABLE I. PERFORMANCE OF THE MODEL

	Precision	Recall	F1- Score
Macro	0.84	0.85	0.83
Weighted	0.86	0.88	0.86

TABLE II. TRAINING AND TESTING ACCURACY

Accuracy	
Training	90.377
Testing	88.083

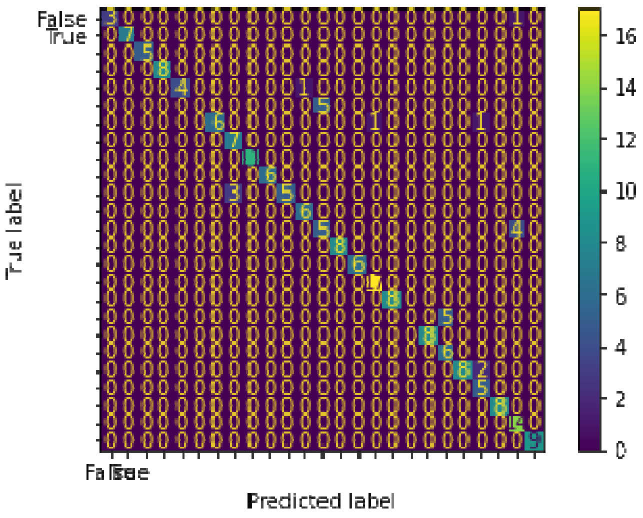
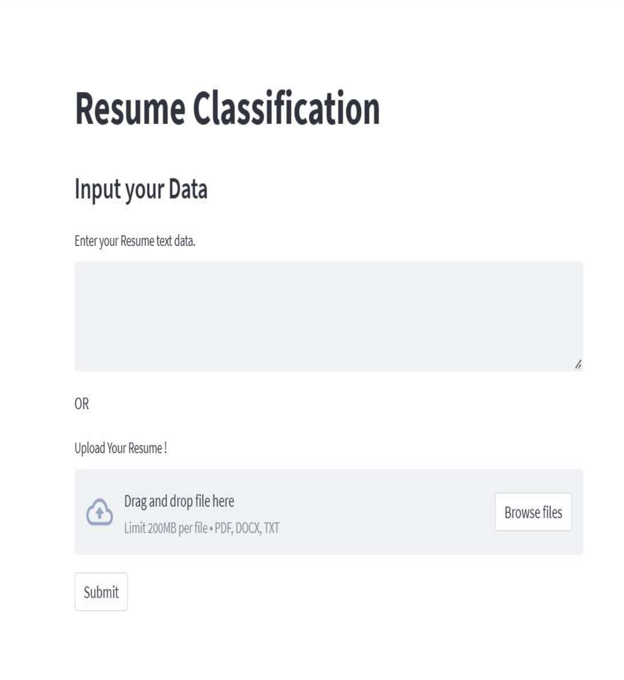


Fig. 4. Confusion Matrix



Resume Classification

Input your Data

Enter your Resume text data.

OR

Upload Your Resume !

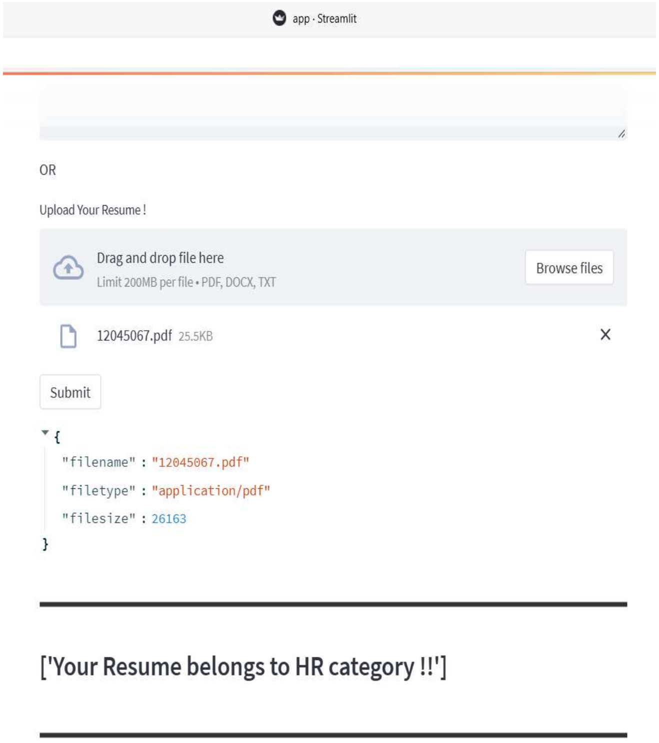
Drag and drop file here

Limit 200MB per file • PDF, DOCX, TXT

Browse files

Submit

Fig. 5. Web app



app · Streamlit

OR

Upload Your Resume !

Drag and drop file here

Limit 200MB per file • PDF, DOCX, TXT

Browse files

12045067.pdf 25.5KB ×

Submit

```
{
  "filename" : "12045067.pdf"
  "filetype" : "application/pdf"
  "filesize" : 26163
}
```

['Your Resume belongs to HR category !!']

Fig. 6. Output in web app

VI. Conclusion AND Future Work

We have studied the performance of a CNN + GRU model for the resume classification task.

The resume classification model will improve the effectiveness of the recruitment process. This strategy will help organizations to streamline the hiring process and save time. We created an Automated Resume Classification model which classifies the resume with an accuracy of 88%

We will explore further different aspects like other structures of neural networks; different types of word embedding layers etc.

Future work includes the ranking to the resume classification with the Ensemble model

We classified resumes only based on skill set but then we can classify them by adding more criteria.

References

[1] Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020). A Machine Learning approach for automation of Resume Recommendation system. *International Conference on Computational Intelligence and Data Science, 167*(Elsevier B.V), 2318–2327.

[2] Gopalakrishna, S. T., & Varadharajan, V. (2019). AUTOMATED TOOL FOR RESUME CLASSIFICATION. *International Journal of Artificial Intelligence and Applications*, *10.* Bengaluru.

[3] Pal, R., Shaikh, S., Bhagwat, S., & Satpute, S. (2022). Resume Classification using various Machine Learning Algorithms. *International Conference on Automation, Computing and Communication*, *44.* Navi Mumbai.

[4] (2020). Differential Hiring using a Combination of NER and Word Embedding. *International Journal of Recent Technology and Engineering.*

[5] Kinge, B., Mandhare, S., Chavan, P., & Chaware, S. M. (2022). Resume Screening Using Machine Learning and NLP : A Proposed System. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8*(2), 253-258.

[6] Dr.K.Satheesh, A.Jahnavi, L.Iswarya, K.Ayesha, G.Bhanusekhar, & K.Hanisha. (2020). Resume Ranking based on Job Description using SpaCy NER model. *International Research Journal of Engineering and Technology, 07*(05), 74-77.

[7] Sayfullina, L., Malmi, E., Liao, Y., & Jung, A. (2017). Domain Adaptation for Resume Classification Using Convolutional Neural Networks. *Springer, Cham.*

[8] Zaroor, Abeer & Maree, Mohammed & Sabha, Muath. (2017). A Hybrid Approach to Conceptual Classification and Ranking of Resumes and Their Corresponding Job Posts. 10.1007/978-3-319-59421-7_10.

[9] V, Senthil kumaran & Annamalai, Sankar. (2013). Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping EXPERT. International Journal of Metadata, Semantics and Ontologies. 8. 56-64. 10.1504/IJMSO.2013.054184.

[10] Zhang, Z., Robinson, D., & Tepper, J. (2018). *Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network* (Vol. 48). Europe: Springer, Cham, June 2018. doi:10.1007/978-3-319-93417-4_48