RESUME SCREENING USING KNN, RANDOM FOREST CLASSIFIER AND DISTILBERT

BY

TAIFUL HAQUE ANAN
ID: 172-15-10111

MD. SHAON MIA
ID: 172-15-9815
And

TANVIR SHIHAB TUKU
ID: 172-15-9723

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Sharmin Akter
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

5th JANUARY 2022

APPROVAL

This Project/internship titled **“Resume Screening Using KNN, Random Forest Classifier and DistilBERT”,** submitted by **“Taiful Haque Anan”**; **“Md. Shaon Mia”** and **“Tanvir Shihab Tuku”**, ID No: **172-15-10111, 172-15-9815 and 172-15-9723** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **05-01-2022**.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan (DTB) **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Riazur Rahman (RR) **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Ohidujjaman Tuhin (MOT) **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Imran

Shah Md. Imran **External Examiner**
Industry Promotion Expert
LICT Project, ICT Division, Bangladesh

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ms. Sharmin Akter , Lecturer , Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Ms. Sharmin Akter
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Anan

Taiful Haque Anan
ID: 172-15-10111
Department of CSE
Daffodil International University

Shaon

Md. Shaon Mia
ID: 172-15-9815
Department of CSE
Daffodil International University

TANVIR SHIHAB TUKU

Tanvir Shihab Tuku
ID: 172-15-9723
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Supervisor Ms. Sharmin Akter , Lecturer , Department of CSE** Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine Learning and Data Mining” to carry out this project. Her endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Prof. Dr. Touhid Bhuiyan** and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

One of the most significant and critical tasks for every firm is to find the right person for the position. As online recruitment becomes more prominent, conventional hiring practices are becoming inefficient. Conventional approaches typically consume more time due to manually reviewing all applicants, assessing their resumes, and then creating a list of candidates who should've been interviewed. Many company hires other firms to screen their candidates resume and find out the suitable person for the position. In this information age, job searching has become both smarter and easier. Companies get a lot of resumes/CVs, and many of them aren't well-structured. Finding suitable candidate for any position takes a significant amount of time and effort. In this study, we have come up with an easy and effective solution for this tedious work. We build three models KNN, Random Forest Classifier and DistilBERT on same dataset for resume classification process. KNN and Random Forest Classifier model have achieved highest accuracy 98% among all the models.

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

CHAPTER

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

To hire the appropriate person at the right time, recruiters have to evaluate resumes effectively. Resume screening is the process of determining if a candidate is qualified for a job based on their qualifications, schooling, work experience, and other information from their CV [1]. The need of effective and useful resume screening is the crucial part in every recruitment process. The motive of resume screening process is to find the most qualified person for a certain vacancy. To evaluate the appropriate and suitable candidate for the role, successful resume screening necessitates domain knowledge. It is difficult for human resource department to make a short-list of all candidates, huge number of various job roles exist currently,in addition hr department receives many applicants. Nowadays industry is facing two major issues while hiring candidates:

- Choosing the best candidates from the crowd - Many people are searching for job , screening applicants suitable for the role is quite difficult. It is becoming time consuming, waste of money and tedious for companies to choose the right person.
- Meaningful resume of the candidate - The fact that CVs available in the domain are not per industry level creating big issue. Applicants use different formats for resume according to their preference. Reviewing all the resumes manually is the only option for HR to find the best candidate for the role. The entire manual screening process is time consuming, mostly inaccurate and also it creates biasness in recruitement process. Sometimes suitable preson for the position get ignored due to inaccuracy.

In our paper, we are proposing a machine learning approach using natural language process to resolve the resume screening process. Our model categorize candidates resume by taking features as an input, mapping the classified resume and matching them with the approriate job description given by the recruiter and recommend the suitable person to HR.

1. We created a system that recommends resumes automatically.
2. To select the most relevant resume, machine learning-based categorization approaches with similarity functions are applied.
3. KNN and Random Forest Classifier performed best compare to other ML classifiers.

1.2 Motivation

There are many automated tracking system for screening process to find any job. Still many resume screening systems are insufficient to find and recommend the best applicant. To solve the problem, we have used Natural Language Processing (NLP) and Machine Learning (ML) to make an efficient system for the entire process. Nowadays, there is a lot of research on both Natural Language Processing and Machine Learning [8]. Most importantly, these two topics we are using almost every day in our life while using mail, online shopping etc. Many recruiters encounter difficulty while filtering suitable applicants from huge number of resume they receive, especially when recruiters receives thousands resumes on average from which 75 percent to 88 percent of them are unqualified [2]. Their team does not have time to study resumes and choose the best CV based on their needs because they are working on a lot of significant projects with big firms. To address this issue, the corporation always hires a third party whose job it is to create a resume that meets the requirements. Hiring Service Organization is the name given to these businesses. It's all about the resume information screen. Resume screening is the process of picking the best candidates for jobs, tasks, and online coding competitions, among other things. Due to a lack of time, large corporations do not have enough time to open resumes, forcing them to enlist the assistance of another firm. For which they must pay a fee. This is a significant issue. To address this issue, the company intends to use machine learning models to process all the applicants resumes matching with their job description .

1.3 Rationale of the Study

Although there was some study done to automate the process in some other way and there was some research to make the process less boring and easier at the same time, but there is still some room for improvement. Many of the natural language processing techniques and machine learning techniques came from analysing the brain interprets real-life data. For example, Artificial Neural Networks (ANN) is a computer program that came from the concept of the biological neural network in the animal brain [11]. Therefore, the primary objective of our paper is to examine how human brain works in case of analyzing a piece of CV/ Resume. Employers take few minutes to check all resumes. This implies that in most of the time employers focus on the important parts of the resume which are relevant with the description and ignore rest of the parts. Segmentation method of the specific resumes makes it easier and faster to inspect and summarize appropriate informations. Therefore, the primary goal is to classify all resumes picking keywords and then match them with the given job descriptions for recommendation process [3].

1.4 Research Questions

1. Can it classify all types of resumes using ML classifiers?
2. Can it eliminate the biasness in resume screening process?
3. Can this model match resumes effectively according to the job requirements given by the recruiters?
4. Can this model does the resume screening process faster?
5.How does all the models perform?

1.5 Expected Outcome

The main objective of our study is to classify resumes efficiently and fast

1. Classifying resumes accurately according to their respective domains.
2. Recommending suitable candidate matching with the job description.
3. Increase accuracy in resume screening.
4. Making light weight model to deploy in any platform effectively.

The aim of this research is to make light weighted model for resume screening and increase the accuracy of the classification so that it can be used for further deployment.

1.6 Report Layout

We covered the introduction of the sufficiency to effective resume screening, motivation, the rationale for the study, and thesis outome. The layout of our report is then followed.

Chapter 1, we have covered the motivation of our work, rationale study, research questions and expected outcome

Chapter 2, we have covered the background of our research topic.

Chapter 3, we have covered the related works on our study, comparative analysis of our study, challenges we have faced.

Chapter 4, we have covered the acquired results, analysis and discussion.

Chapter 5, we have covered the future work we will do for further advancement.

CHAPTER 2
BACKGROUND

2.1 Introduction

Companies are receiving thousands resumes per job recruitement. In every job post they receive different styles of resumes from different domains including their work experience [5]. As a result, not only is it necessary to extract specific data from such resumes in order to support fully automated match according to the job description, but it is essential to match the resume with their suitable job to reduce time and effort to manage but it is also necessary to efficiently match them to their appropriate job categories in order to reduce the time and effort required to manage them. For that reason, rather than scanning across all resumes and postings, resumes that match with their specific category will automatically match to their corresponding job posting [6]. We introduce a technique in this paper that uses a conceptual-based categorization of resumes accoridng to job descriptions.

2.2 Related Works

Senthil Kumaran et al. [1], in their work they have used EXPERT mapping-based screening for the recruitement process. For the precision enhancement they have matched the job requirement with the applicants resume.

Jagan Mohan Reddy D et al. [2] they have discussed about a process where they can make a process so that it will be time consuming and cost-effective. In their study they have suggested not to use specific characteristics such as age, salary and so on directly for the problem of value variations.

Frank Färber et al. [3] discussed about an automated recommendation approach for indiviual profiling and selecting the candidates. Lack of proper data can be a reason for not obtaining the potential result which was expected.

Chirag Daryania et. al [4] used natural language processing for their screening system.

To match candidates resume with the job requirements and for recommend they have used Vector Space Model. They have used both cosine similarity and vectorisation model for ranking resumes and search the suitable person for the role.

Momin Adnan et. al [5] they have proposed a screening system model used by Linear SVM classifier. In their model they used KNN, cosine similarity and content-based recommendation to match the resume with the job requirements. For summarisation they have used “genim” package. Due to compression crucial informations get lost.

2.3 Comparative Analysis and Summary

2.3.1 CV / Resume Analyzing Process

Recruiters used to manually review and judge CVs/Resumes provided by job seekers in the past. In modern times, this method is still used. However, because large firms frequently deal with hundreds of CVs/Resumes each day, handling such a large number of CVs/Resumes one by one is becoming extremely difficult and time-consuming. As a result, many businesses have begun to offer certain templates or forms that job searchers must fill out with needed information before the CV/Resume is reviewed by a machine using simple pattern recognition and keyword searches [14]. While this strategy lowered the workload for companies, it dramatically raised the workload for applicants, who must maintain distinct forms of resume according to the job offered by companies. In addition, it may limit candidate’s flexibility to use distinct format for their resume according to their skills.

2.3.2 Natural Language Processing Approach

Considering all the advantages and disadvantages, there is a high demand for a process which is automated so that it’s easy to find suitable candidates, employers are able to select suitable applicants quickly and can demonstrate their creativity while using a single application format to apply to multiple organizations. In this scenario, advancements in both Natural Language Processing and Machine Learning have proved quite beneficial. Understanding unorganized written language and extracting vital information to train any model is required for analyizing any text documents, such as resume papers, in the same way that humans do.

2.3.3 Machine Learning Approach

For precise and accurate result many researhcers prefer to use machine learning combining with the natural language processing [12]. Machine learning approach can be used in any type of problem solving by training models. The availability of differenct techniques in machine learning makes it popular in research community. Logistic regression, naive Bayes classifier, Decision trees are the most common aprroaches in machine learning to to solve any kind of problems related to the domain. They have also been used in the past to determine various diseases, like cancer. Thus it has becoming popular for making precise decisions we can use this approach to solve our resume screening problem.

2.3.4 CV / Resume Processing

Apart form resume screening system in many field both machine learning method and natural language processing were applied for example Grading system, Question and Answer format, Summarization etc. For better accuracy and result different techniques were used to make this systems easy and fast. Thereby, we are using these procedure for screeing our resumes to get positive values for our model. For precision, recall, and better accuracy of our models we have used four algorithms. High accuracy and precision will make a better system for our model. First we used stop words to eliminate unusual text from our resume. Then we labeled them in our dataset to train our models. Precision is crucial for any model for better result comapre to real world. Recall is from all truth how many it got right.

2.4 Challenges

Different format of resumes are the biggest challenge to extract informations. Applicants use different formats to write their resume. So, their is no one format for resume. Training model different formats of resume is difficult. Exctracting appropriate informations from unstructured resumes are the most challenging task to do. In case of model training it is difficult to train model every keywords involved. Eliminating biasness is another big challenge in this entire training process.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In research methodology, we discussed about concepts that facilitate the construction of the proposed Automated Resume Screening System. The system works in two phases. First it will classify all the resumes in their respective categories and then in second phase it will match the resume according to it's suitable job requirements. Figure 3.1, represents the entire working procedure our study.

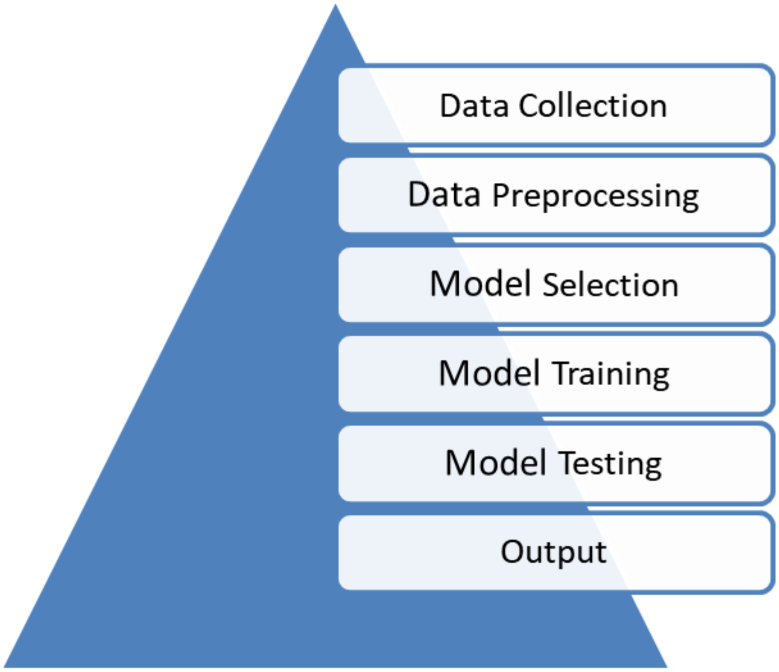


Figure 3.1: Methodology flow chart

3.2 Research Subject and Instrumentation

Getting approriate data for research is the most crucial thing. The first step is to ask valid questions and find out those answers to solve the problem. And by using those organized data and applying them on different models will help us to obtain best results. A better research depends on valid questions which we followed to solve our problem.

- How to collect data, from where and in what manner?
- Should we have to modify our data?
- Where should we store our data?
- What will be the labeling procedure of our data?
- What will be the size of our dataset?
- How should we clean our data?

3.3 Data Collection Procedure

Our dataset consist of 1000 resumes from which we have collected 600 cleaned resume from kaggle. We have manually made rest of the resumes using different format styles. We categorised our resume data in four formats such as Chronological, Functional, Combined and Focused. Then we proceeded our cleaning process by removing digits, unique characters and single letter words. For tokenization we have used NLTK library. For processing the dataset we have included lemmatization, vectorization and stop word removal.

3.3.1 Removing Stop Words

There are some words such as was, and , the which doesn’t help for text processing are eliminated. We have used following steps to filter our dataset:

- Words taken as an input were tokenized and stored in an array.
- Every tokenized words correlate with the NLTK library.
- If any word listed in the library appears it will be removed from the sentence.

3.3.2 Lemmatization

In lemmatization the root of each word will be extracted. The only difference between lemmatization and stemming is lemmatization extract the root word and remain it meaningful. The following are the regular stages of lemmatization:

- Text corpus should be converted and make a word list.
- Extracting the root of each word and list them.
- Linking lemmas from the extracted words.

3.3.3 Information extraction

The first phase of our proposed system involves information extraction using Natural Language Processing. The information in the resumes is not present in a structured format. There are noises, inconsistencies and irrelevant bits of data which is of no use to the recruiters. The objective is to derive relevant keywords from the unstructured textual data in the resume without any need of human crawling efforts. Using techniques like Tokenization, Stemming, POS Tagging, Named Entity Recognition, etc., our system obtains important job-related content (skills, experience, education, etc.) from the uploaded candidate resumes. The result is a summarised version of each resume in a CSV format which can be easily used for further processing tasks in the next phase of this resume screening system.

3.3.4 Tokenization

After converting the various resume formats (.docx, .pdf, .jpg, .rtf, etc.) into text, we begin the tokenization process to identify terms or words that form up a character sequence. This is important as through these words, we will be able to derive meaning from the original text sequence. Tokenization involves dividing big chunks of text into smaller parts called tokens. This is done by removing or isolating characters like whitespaces and punctuation characters. Tokens are sentences initially (when tokenized out of paragraphs) and then are further split into individual words. By performing Tokenization, we can derive information like the number of words in a text, frequency of a particular word in the text and much more. The tokenization can be performed in multiple ways such as using Natural Language

Toolkit [NLTK], the spaCy library, etc. Tokenization is a mandatory step for further text processing such as removal of stop words, stemming and lemmatization.

3.3.5 Parts of speech (POS) tagging

In pos tagging assigning grammatical information in word with the context and relationship with the other words in a sentence. The part-of-speech tag specifies whether the word is a noun, pronoun, verb, adjective, etc. according to its usage in the sentence. It is important to assign these tags so as to understand the correct meaning of a sentence and for building knowledge graphs for named entity recognition. This process is not as simple as mapping a word to their corresponding part of speech tags. This is so as a particular word may have a different part of speech based on different contexts in which it is used. For example: In the sentence "I am building a software", building is a Verb, but in the sentence "I work in the tallest building of that street", building is a Noun. Also called grammatical tagging or word-category disambiguation, it is a type of supervised learning technique to analyses the extracted features such as preceding word, following word, capitalizing first letter or not etc. to label the words after tokenization. Rule-Based POS tagging, Stochastic POS tagging, and Transformation based tagging are mostly used.

3.3.6 Chunking

Chunking is a process that aims to add more structure to sentences by grouping short phrases with parts of speech tags. Because parts of speech tags alone cannot give information about the structure of the sentence or the actual meaning of the text, chunking combines parts of speech tags with regular expressions to give a result as a including a chunk consists of verb phrase, noun pharse etc. Also called Shallow Parsing, it involves the construction of a parse tree that can have a maximum one level of information from roots to leaves. This ensures there is more information than just part of speech of the word without needing to create a full parse tree. Chunking segments and labels multi-token sequences, mostly making groups of "noun phrases" that are used for finding named entities.

3.3.7 Named entity recognition

Named Entity Recognition is an information extraction technique which extracts relevant information by classifying chunks of unorganized text into predefined categories like names of persons, companies, contact info, educational credentials, and skills. After classifying the unstructured resume data into such different sets of categories, our aim is to use a similarity model to determine the similarity between the categorized resume data and the requirements provided by the recruiters. There are many approaches to implement the Named Entity Recognition in order to derive relevant categories from unstructured data. These include the Rule-Based approach in which we define our own algorithms according to the required domain. We can also use regular expressions, which finds patterns in a string to detect the named entities. Another approach is using Bidirectional-LSTM with the an algoritm named conditional random field as a sequence labelling problem. We have used the spaCy module which consists of various pre-trained models that can recognize a number of default entities from the content of the documents. These models use language information to detect these entities. We also trained the model on a large annotated set of resume samples for better accuracy in the entity recognition. We could detect entities like name, phone number, email, educational institute, organisation etc.

3.3.8 Vectorization

In vectorization process it converts texts in vectors. It assigns a number to individual word and match it with the recommended words. It is an algebraic model for representing text information for Information Retrieval, Natural Language Processing and Text Mining. Representing documents in a vector space model is called vectorisation. It is the process of turning a document into a numerical vector [8]. An important reason behind performing vectorisation is that most machine learning models require the input to be numerical vectors rather than strings. A common way of vectorising text is mapping words in individual integer. Every word will be represented in an array in where every word fits in an array. We can determine the value by the number of occurance of the word. Generally, if array size lesser than the corpus we should thus have a vectorisation strategy to account for this.

3.3.9 TF-IDF

TF-IDF used for determining frequency of any given words in text documents. Using TF-IDF for text mining gives better result. It was invented for extracting informations and document search. The given value is a numerical measure for determining the importance of that term with respect to the corpus. If any word appears mostly then it will be rank as an important in the document. So, terms we use frequently used in most documents, for example whom, and, this, what, is, if, the, etc. all this words will be given low importance doesn’t matter how many times it occurs in the document.

3.3.11 Transfer Learning

In transfer learning a model built and trained for one purpose on different dataset then used in an another domain to decrease model render time. Considering vast time consuming and compute processing required for implementing neural network models on these problems, and the performance and precision it provides to solve other problems, a popular strategy in deep learning where pre-trained models are being used as the initial point on machine learning and natural language processing tasks [10]. Transfer learning is commonly used for natural language processing tasks in which it takes text as input or output. To reslove this issues word embedding works better, in which it map words to a wide representation of vector having identical vector representations for distinct words with similar meanings.

3.3.12 Data Organizing

We have separated our dataset for tesing and training and stored them in two different folders. In each folders we have stored pdf format of each resumes. Different categories of data were also stored within two folders using sub folders.

3.3.13 Data Storing

As we are using google colab platform for our model tarining storing all data in google drive makes it easier to work with the dataset. By uploading our dataset in google drive we can use these dataset which is in csv file to colab directly. We can import this dataset in colab anytime and use it for training and testing.

3.3.14 Machine Learning Algorithms

For better performance and accuracy we have used K-Nearest Neighbors, Random Forest Classifier and DistilBERT for buliding our model and imported the dataset.

3.4 Statistical Analysis

In our dataset, we have 1000 resumes from different catagories. We splited our dataset in 80% for training and 20% for testing data. Applying machine learning algorithms such as (K-Nearest Neighbors, Random Forest Classifier and DistilBERT) we tried to obtain best accuracy in our model.

3.5 Implementation Requirements

3.5.1 Python 3.8

We are using current versiono of python 3.8. It is a programming language with a high level of abstraction. It is used by the majority of researchers to conduct their studies. It is a highly recommended programming language for machine learning projects and is extremely popular among programming vommunity due to its ease of learning and comprehension .

3.5.2 Google CoLab

Google CoLab is an open source platform from Python distribution that is free to use. We may work here online using our browsers in the same way. Biggest advantage using colab is it gives the free access of the virtual gpu in online.

3.5.3 Hardware and Software Requirements

- Chrome browser
- Operating System (Windows 7)
- Ram(more than 4 GB)
- Hard Drive or SSD (minimum 120 GB)

CHAPTER 4

CLASSIFIACTION AND MODEL STUDY

4.1 Classification

Classification is one of two types of supervised machine learning models (the other being regression). For data analysis, classification is a far more straightforward technique of prediction. The exact value, which can be yes or no, is determined by classification. It can more precisely determine whether a prediction is correct or incorrect. It may be used to anticipate the exact result for the low and high ranges of any dataset.

4.2 Model Summary

In figure 4.1 shows the whole process of training data, extracting features, labeling and testing them for evaluation.

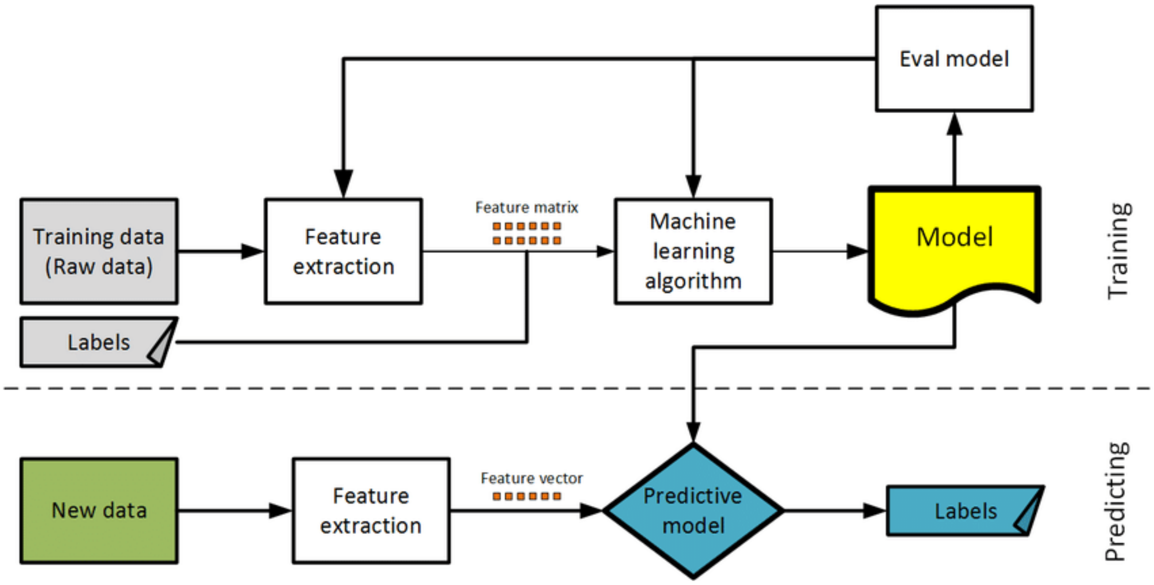


Figure 4.1: Model Training and Predicting Process

4.2.1 K-Nearest Neighbor

KNN model is a type of supervised model in which classification is used. It is a very popular and common algorithm in machine learning. Its operating technique is based on similarities between data that we previously inserted. It primarily calculates output from prior analyses. It calculates the distance between fresh and old data. Then it measures the distance and chooses the closest value before comparing the old and new data. As a result of comparing the distance between the old and new data using classification, the final decision value is obtained.

4.2. Random Forest Classifier

Random forest classifier is the most used algorithm in case of training due to it's simplicity and scalability. It usually use bagging method for training group of decision trees. It creates various decision trees and integrate them to get better accuracy and prediction. It can be used in different machine learning problems as it utilizes in both classification problems and regression problems. The advantage of using this algorithm is while measuring individual features value it is not difficult to extract it. It predicts by computing score for individual features automatically and then execute prediction.

4.2.3 DistilBERT

DistilBERT is the light weight version of BERT. Where BERT takes so much time in computing due to it's huge size distilbert is the compressed version of the bert. In this paper, we offer DistilBERT method which is pre trainned smaller general-purpose language representation model that can subsequently be used in differen tasks by fine tuning the parameters. Distlbert transfromer built in top of bert transformer by distilling and building task specific models. It reduces the bert model by 40% but shows the same performance like the bert model. We develop a loss that combines language modeling, distillation, and the loss of cosine distance.

Table 4.1: DistilBERT Training Configurations

Configuration	**Value**
Vocabulary Size	30522
Number of Hidden Layers	6
Number of Attention Heads	12
Intermediate Layer Size	3072
Hidden Activation Function	gelu
Hidden Dropout Probabilty	0.1
Dropout ratio for the attention probabilities.	0.1
Maximum Sequence Length	512
Initializer Range	0.02

4.3 Training the Model

For trainning Distilbert we have used 1000 resumes using 5 epochs , 64 training batch size, 32 validation batch size, 256 number of sequence with the learning rate of 5e-05.

4.4 Discussion

Our dataset and method have been changed. As a result of the update, we now know that this classifier can be used to predict whether or not a dataset is accurate over a wide variety of different datasets. We've succeeded in describing the precision of 98 percent of impact expectations. The model allows us to think about as well as obtain the suitable result.

CHAPTER 5

EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Results

Multiple indicators, including recall, precision and f1 score are used for measuring the performanc of model.

Precision is all the correct detection over all the detections the model has detected. It is true positive (TP) over the sum of true positive (TP) and false positive (FP) value.

$$Precision = \frac{TP}{TP + FP}$$

Recall on the other hand is all the correct detection over ground truth detection value.

$$Recall = \frac{TP}{TP + FN}$$

There is always a trade-off between Precision and Recall. If precision value gets higher, recall value will fall. So maintaining a good balance between these two indicators is crucial for a good model. There is another indicator called F1-score which combines both precision and recall.

$$F1\ score = 2 \times \frac{Precision\ \times Recall}{Precision + Recall}$$

Table 5.1: Accuracy Table

Algorithm Name	**Accuracy**	**Precision**	**Recall**	**F1 Score**
KNN	98%	0.98	0.96	0.94
Random Forest Classifier	98%	0.94	0.95	0.94
DistilBERT	96%	0.92	0.94	0.92

5.2 Discussion

Our models will classify all the resumes according to their categories and macthing with employers job requirement so that recruiter can extract and find the right person easily and fast. Recruiter can rank all the resumes suitable for the job and classify them for better selection. It will help employer to find best candidate with less effort and less time with higher precision rate. Even our system will eliminate human biasness in resume screening which is a big issue in recruitement process.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

In every job position, organization receives a high number of applicants. In today's world, finding best candidate for the given job role from thousands of resumes is becoming difficult and complex. Screening resume manually takes more time, effort, resource. We used different machine learning models and nlp for extracting resume informations and recommend best applicant for the job. Our system worked in two stages first it will classify all the resumes according to their categories and then it will match the resume with job description and recommend suitable candidate.

6.2 Future Work

More accurate and precise model can be constructed by getting feedbacks from domain experts such as HR professionals. We simply developed the model that displays the accuracy score, but we plan to develop a website in the future where machine will help to detect whether the resume informations are fake or not. It will match the information of a candidate from his social media profiles and verify the authenticness of the information. HR can easily import any type of resume file(pdf or doc) and can find the best candidate for the position.

REFERENCES

[1] Sankar, A. (2013). “Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (EXPERT)”. International Journal of Metadata, Semantics and Ontologies, 8(1), 56. https://doi.org/10.1504/ijmso.2013.054184

[2] Jagan Mohan Reddy D, Sirisha Regella., “Recruitment Prediction using Machine Learning”, IEEE Xplore, 2020.

[3] Färber,F., Weitzel, T.,Keim, T., 2003. “An automated recommendation approach to selection in personnel recruitment”. AMCIS 2003 proceedings , 302.

[4] Chirag Daryania, Gurneet Singh Chhabrab, Harsh Patel, Indrajeet Kaur Chhabrad, Ruchi Patel., “An Automated Resume Screening System using Natural Language Processing and Similarity”. (2020). Topics In Intelligent Computing And Industry Design.

[5] Momin Adnan, Gunduka Rakesh, Juneja Afza, Rakesh Narsayya Godavari, Gunduka and Zainul Abideen Mohd Sadiq Naseem., “Resume Ranking using NLP and Machine Learning”, (2016b). Institutional Repository of the Anjuman-I-Islam’s Kalsekar Technical Campus. https://core.ac.uk/display/55305289.

[6] Nimbekar, Rohini, et al. "Automated Resume Evaluation System using NLP." 2019 International Conference on Advances in Computing, Communication and Control (ICAC3). IEEE, 2019.

[7] Roy, Pradeep Kumar, Sarabjeet Singh Chowdhary, and Rocky Bhatia. "A Machine Learning approach for automation of Resume Recommendation system." Procedia Computer Science 167 (2020): 2318-2327.

[8] Sinha, Arvind Kumar, Md Amir Khusru Akhtar, and Ashwani Kumar. "Resume Screening Using Natural Language Processing and Machine Learning: A Systematic Review." Machine Learning and Information Processing: Proceedings of ICMLIP 2020 1311 (2021): 207.

[9] Chen, Jie, Chunxia Zhang, and Zhendong Niu. "A two-step resume information extraction algorithm." Mathematical Problems in Engineering 2018 (2018).

[10] Tejaswini, K., et al. "Design and Development of Machine Learning based Resume Ranking System." Global Transitions Proceedings (2021).

[11] Gopalakrishna, Suhas Tangadle, and Vijayaraghavan Vijayaraghavan. "Automated Tool for Resume Classification Using Sementic Analysis." International Journal of Artificial Intelligence and Applications (IJAIA) 10.1 (2019).

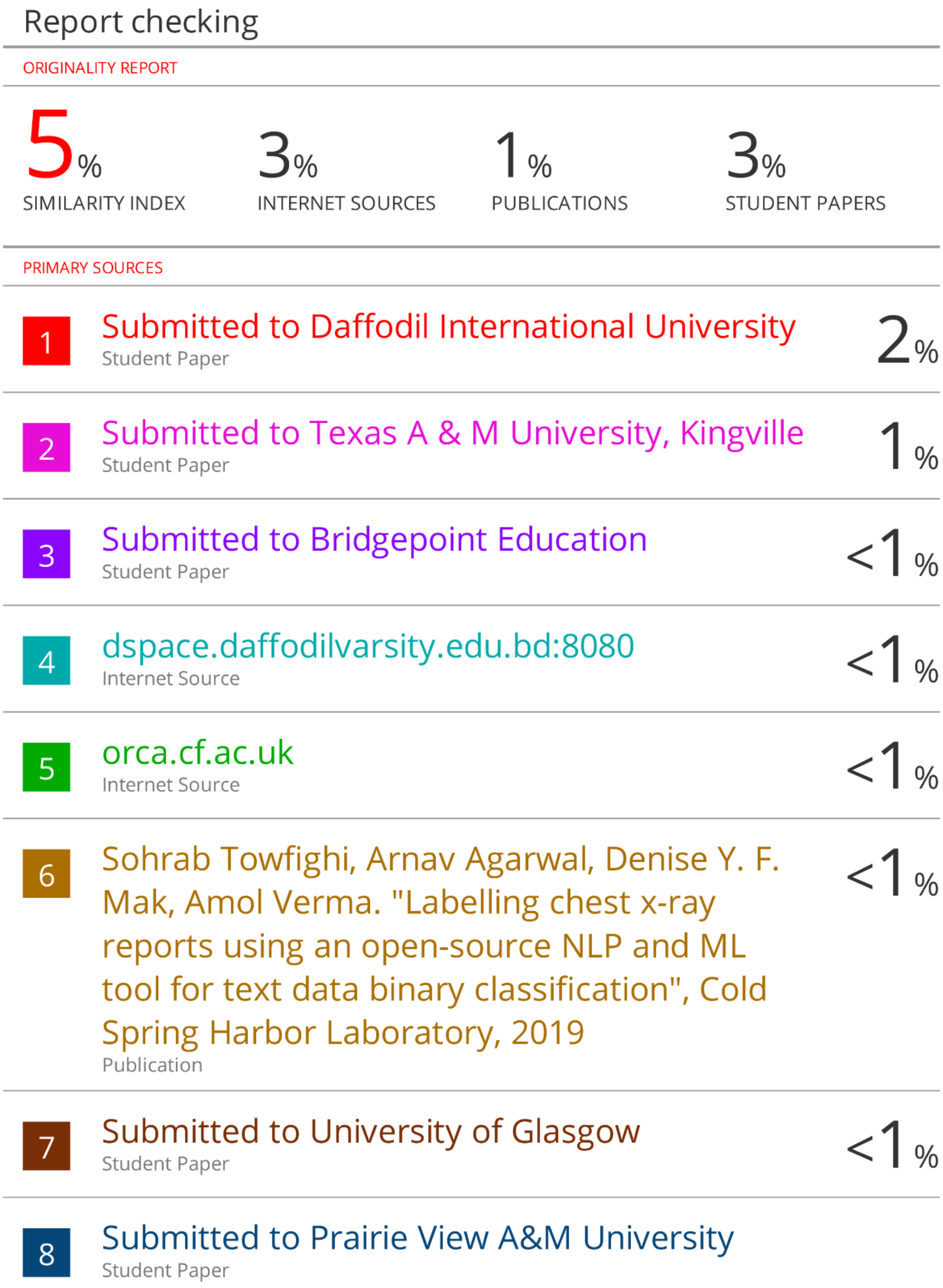
[12] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

[13] Clark, Kevin, et al. "Electra: Pre-training text encoders as discriminators rather than generators." arXiv preprint arXiv:2003.10555 (2020).

[14] Mohamed, Ashif, et al. "Smart Talents Recruiter-Resume Ranking and Recommendation System." 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS). IEEE, 2018.

[15] Goyal, Umang, et al. "Resume Data Extraction Using NLP." Innovations in Cyber Physical Systems. Springer, Singapore, 2021. 465-474.

[16] Satheesh, K., Jahnavi, A., Iswarya, L., Ayesha, K., Bhanusekhar, G. and Hanisha, K., 2020. Resume Ranking based on Job Description using SpaCy NER model.



Report checking

ORIGINALITY REPORT

5%	3%	1%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Submitted to Texas A & M University, Kingville Student Paper	1%
3	Submitted to Bridgepoint Education Student Paper	<1%
4	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
5	orca.cf.ac.uk Internet Source	<1%
6	Sohrab Towfighi, Arnav Agarwal, Denise Y. F. Mak, Amol Verma. "Labelling chest x-ray reports using an open-source NLP and ML tool for text data binary classification", Cold Spring Harbor Laboratory, 2019 Publication	<1%
7	Submitted to University of Glasgow Student Paper	<1%
8	Submitted to Prairie View A&M University Student Paper	