RESUME SCREENING USING LSTM

Divya Mule[1], Samiksha Doke [2], Sakshi Navale[3], Prof. S.K.Said[4]

[1,2,3]BE Student, Jaihind College of Engineering, Kuran, Junnar, Pune,
[4]Assistant Professor, Jaihind College of Engineering, Kuran, Junnar, Pune,

---***---

Abstract - Résumé screening is the process of identifying whether or not a candidate is qualified for a position based on their education, experience, and other information on their resume. It's a pattern matching and classification technique. Long-term memory networks (LSTM) are neural networks that can learn long-term dependencies. The PDF file will be used as input for Resume Screening with LSTM. The file's data will be extracted. The extracted information will be saved in a CSV file. After cleaning the data from the csv file, the LSTM model will be utilised to produce predictions. The LSTM will be trained and stored for usage in the web app. The GitHub profile of the applicant will also be retrieved. To find information about the GitHub profile, web scraping will be employed. The result will be influenced by the study of the GitHub profile. The database will keep track of the user's information. After that, the data set will be cleaned and vectorized. After that, the vectorized dataset will be used for training. Weights will be retained for future usage and the bidirectional LSTM model will be trained.

Key Words**: Machine Learning, Supervised learning, Deep Learning, LSTM.**

1. INTRODUCTION

The practise of analysing a resume to evaluate if the individual is qualified for the post is known as resume screening. Education, experience, abilities, and other relevant information on the resume will be used to determine whether or not the individual is qualified. Resume screening is still the most time-consuming aspect of the hiring process. For a single hire, resumes are projected to take up to 23 hours. When a job posting generates 250 resumes on average, and 75 percent to 88 percent of them are unqualified, it's no surprise that filtering the right people from such a large applicant pool is the most difficult element of the job.

To make matters worse, according to a recent survey of talent acquisition leaders, 56 percent plan to increase their hiring volume in the coming year, but 66 percent of recruiting teams will either remain the same size or shrink. In this work, we propose a system that will benefit recruiters by ensuring that the best candidate is found in the shortest amount of time.

1.1 Objectives

To create a computerised system for the aim of recruitment.

To create a system that industries can utilise to recruit.

Find the best applicant in the shortest amount of time.

2. PROPOSED METHODOLOGY

The PDF file will be used as input for Resume Screening with LSTM. The file's data will be extracted. The extracted information will be saved in a CSV file. After cleaning the data from the csv file, the LSTM model will be utilised to produce predictions. The LSTM will be trained and stored for usage in the web app. The GitHub profile of the applicant will also be retrieved. To find information about the GitHub profile, web scraping will be employed. The result will be influenced by the study of the GitHub profile. The database will keep track of the user's information. After that, the data set will be cleaned and vectorized. After that, the vectorized dataset will be used for training.

Weights will be retained for future usage and the bidirectional LSTM model will be trained. VScode and the Django framework will be used to build the system. The Python language will be utilised to carry out the work. Figure 1 depicts the proposed system's design. There are two sorts of users in the proposed system.

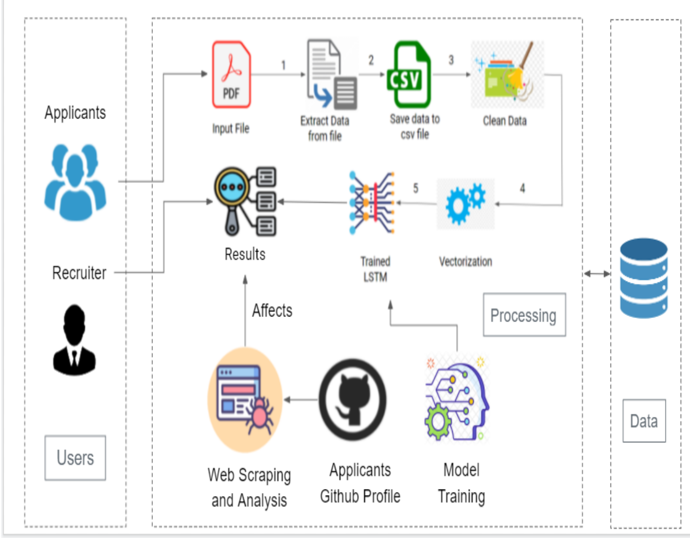

Fig -1: Proposed System Architecture

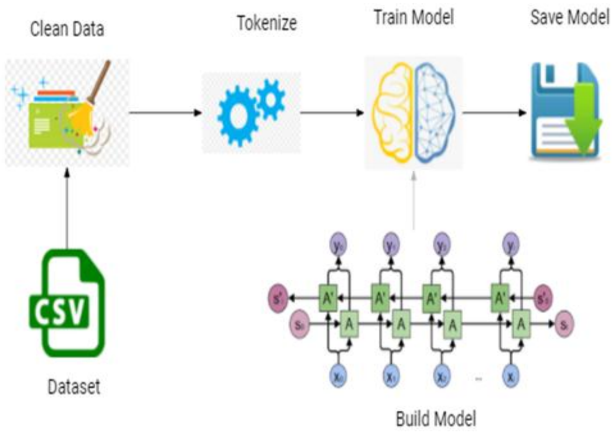

Fig -2: Model Training

3. MACHINE LEARNING TECHNIQUES

Long short-term memory (LSTM) is a deep learning architecture that employs an artificial recurrent neural network (RNN) architecture (DL). LSTM has feedback connections, unlike traditional feedforward neural networks. It can analyse whole data sequences as well as single data points (such as photos) (such as speech or video). LSTM can be used for tasks like unsegmented, linked handwriting recognition,[2] speech recognition,[3][4] and anomaly detection in network traffic or IDSs, for example (intrusion detection systems).

A cell, an input gate, an output gate, and a forget gate are the components of a typical LSTM unit. The three gates control the flow of information in and out of the cell, and the cell remembers values across arbitrary time intervals.

Because there might be lags of undetermined duration between critical occurrences in a time series, LSTM networks are well-suited to categorising, processing, and making predictions based on time series data. LSTMs were created to solve the problem of vanishing gradients that can occur when training traditional RNNs. In many cases, LSTM has an advantage over RNNs, hidden Markov models, and other sequence learning approaches due to its relative insensitivity to gap length.

OCR: Human eyes are capable of recognising a wide range of patterns, fonts, and styles. It is difficult task for computers. A graphics file, or a pattern of pixels, is created when a document is scanned. Characters on an image are localised, detected, and recognised by a computer, which then converts the image of paper documents into a text file. Then, and only then, is it possible to extract useful information. Machine-readable texts can then be used for a variety of reasons. They may be scanned for trends and important data, utilised to create reports and charts, and distributed into spreadsheets, among other things. Optical character recognition (OCR) methods allow computers to automatically analyse printed or handwritten documents and convert text data into editable formats so that computers can process them more effectively. It's yet another method for extracting and utilising business-critical data. Businesses that use data can gain a competitive advantage and generate $430 billion in productivity gains by 2020, according to the International Institute of Analytics.

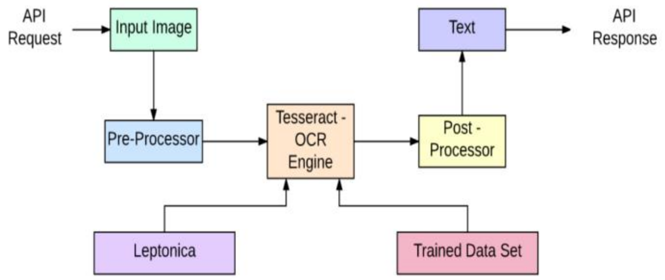

Fig -3: Optical Character Recognition

Support Vector Machine is a supervised machine learning technique for solving classification and regression issues. It is, however, mostly used to tackle categorisation problems. SVM builds a hyperplane in a high or infinite dimensions space that can be used for classification, regression, or other tasks. The purpose of the SVM algorithm is to discover the best line or decision boundary for categorising n-dimensional space into classes so that subsequent data points can be easily placed in the relevant category. The ideal choice boundary is known as a hyperplane. A hyperplane can be numerous lines or decision borders to segregate the classes in n-dimensional space, however we must choose the best decision boundary to categorise the data points.

Support Vector Machine is a supervised machine learning technique for solving classification and regression issues. It is, however, mostly used to tackle categorisation problems. SVM builds a hyperplane in a high or infinite dimensions space that can be used for classification, regression, or other tasks. The purpose of the SVM algorithm is to discover the best line or decision boundary for categorising n-dimensional space into classes so that subsequent data points can be easily placed in the relevant category. The ideal choice boundary is known as a hyperplane. A hyperplane can be numerous lines or decision borders to segregate the classes in n-dimensional space, however we must choose the best decision boundary to categorise the data points.

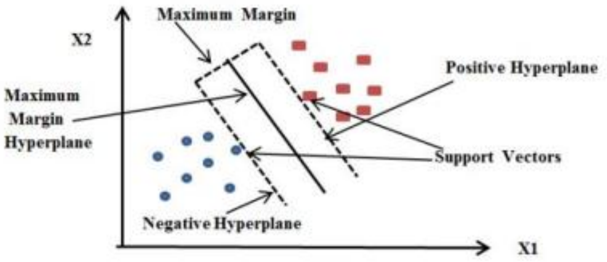

Fig -4: Support Vector Machine

KNN stands for K Nearest Neighbor and is one of the most basic machine learning algorithms. Machine learning models anticipate output values based on a set of input values. KNN is a type of machine learning algorithm that is primarily used for categorization. It classifies the data point based on the classification of its neighbours.

The K-Nearest Neighbour algorithm is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms.

The K-NN method assumes that the new case/data and existing cases are similar and places the new case in the category that is most similar to the existing categories.

The K-NN method stores all available data and classifies a new data point based on its similarity to the existing data. This means that new data can be quickly sorted into a well-defined category using the K-NN method.

The K-NN approach can be used for both regression and classification, but it is more commonly utilised for classification tasks.

The K-NN algorithm is a non-parametric algorithm, which means it makes no assumptions about the underlying data.

It's also known as a lazy learner algorithm since it doesn't learn from the training set right away; instead, it saves the dataset and performs an action on it when it comes time to classify it.

During the training phase, the KNN algorithm simply stores the dataset, and when it receives new data, it classifies it into a category that is quite similar to the new data.

The process of model training is shown in Fig-5 below. The data from dataset will first be cleaned. Then it will be tokenized to be used for training purpose. The the data will be used for training. A Bidirectional LSTM model will be trained. The weights of model will be saved. The saved weights will then be used for Prediction.

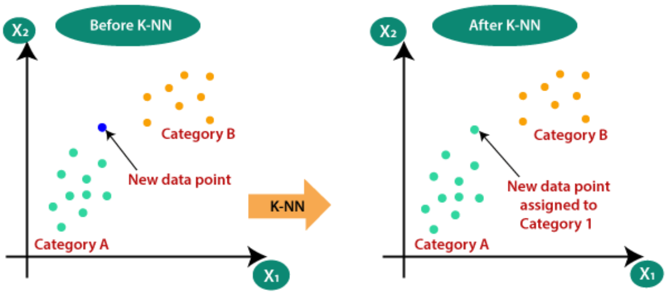

Fig -5: Support Vector Machine

3. EXPERIMENTAL RESULT AND ANALYSIS

The author of [1] describes a web-based resume screening application. The use of a natural language processing pipeline. Text extraction is carried out using a segmentation system based on parts. To train a machine learning model, use semi-supervised learning. There are various flaws in the system. On the recruiter's side, the web application displays results in the form of a rating. One of the features is that their resume is only compared to job positions for which they are qualified and have applied.

The author of [2] employs Principal Component Analysis' OCR (Optical Character Recognition) Feature Extraction. The categorization algorithm is based on a decision tree. However, the technique only works with Urdu text. Conventional SVM can be optimised for text categorization, according to paper [3]. The traditional SVM is optimised through entropy-based feature selection. Three categorization algorithms are compared by the author of [4]. Support vector classifier with the TFIDF feature outperform naïve bayes and KNN in terms of accuracy.

The author of [5] offers an online document classification system based on a fuzzy k-NN network, with TF/IDF used to choose document features during the classification process. The findings reveal that classification results is superior to that of k-NN and SVM, although classification time is slower than that of KNN.

4. CONCLUSIONS

Recruiters can benefit from the Online Recruitment application by finding the best applicant in the shortest amount of time. It will reduce the amount of time spent screening and provide the best profile. The proposed system might include a function that allows recruiters to check an applicant's previous work. Because it takes into account the GitHub profile, the application could be beneficial to the IT business.

REFERENCES

[1] Sujit Amin, Nikita Jayakar, Sonia Sunny, Pheba Babu, M. Kiruthika, Ambarish Gurjar, Web Application for Screening Resume,2019 International Conference on Nascent Technologies in Engineering (ICNTE 2019),978-1-5386-9166-3/19/$31.00 ©2019 IEEE.

[2] K. Khan,R. Ullah khan, Ali Alkhalifah, N. Ahmad, Urdu Text Classification using Decision Tree,978-1-4673-9268-6/15/$31.00 ©2015 IEEE.

[3] Zi-qiang wang, xia sun, de-xian zhang, xin li, an optimal svm-based text classification algorithm, Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006 1-4244-0060- 0/06/$20.00 ©2006 IEEE

[4] Fang Miao, Pu Zhang, Libiao Jin*, Hongda Wu, Chinese News Text Classification Based on Machine learning algorithm,2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetic.

[5] Huabe iNie,Yi Niu,Juan Zhang, Web Document Classification Based on Fuzzy KNN Algorithm,2009 International Conference on computational intelligence and Security 978-0-7695-3931-7/09 $26.00 © 2009 IEEE