**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**II SEMESTER 2019-2020**

DSE CL ZG628T **DISSERTATION**

**Dissertation Outline**

**BITS ID No:** 2018AH04003  **Name of Student:** Shravan Mateti

**Name of Supervisor:** Dr. V. Vijay Kumar

**Designation of Supervisor**: Dean Research and Development; Anurag University

Former Professor SIT, JNT University

**Qualification and Experience:** M.S. Engg (USSR), Ph.D and 30 years’ experience

**E- mail ID of Supervisor:** drvvk\_dean@cvsr.ac.in

**Topic of Dissertation**: Duplicate candidate profiles identification during application screening of a hiring process

**Name of First Examiner:**

**Designation of First Examiner**:

**Qualification and Experience:**

**E- mail ID of First Examiner:**

**Name of Second Examiner:**

**Designation of Second Examiner:**

**Qualification and Experience:**

**E- mail ID of Second Examiner:**

(Signature of Student) (Signature of Supervisor)

Date: 10-Sep-2020 Date: 10-Sep-2020

**Dissertation Title**: Duplicate candidate profiles identification during application screening of a hiring process

**Date**: 10-Sepetember-2020

**Student details**:

**Name**: Shravan Mateti

**Roll #**: 2018AH04003

**Email ID**: 2018ah04003@wilp.bits-pilani.ac.in

**Mobile #**: 9177435353

**Supervisor details**:

**Name**: Dr. V. Vijay Kumar [M.S. Engg (USSR), Ph.D]

**Role in organization**: Dean Research and Development; Anurag University

**Email ID**: drvvk\_dean@cvsr.ac.in

**Mobile#**: 9849452287

**Problem statement**:

A candidate can apply for open roles in a company through multiple channels. The channels could be, candidate directly applying over company’s job portal, through employee referral, through a recruiting agency or sourced by recruiter from social network or campus hiring events.

Due to GDPR (General Data Protection Regulation) regulations the hiring company can’t collect candidates personal identification data like Govt. ID proofs, Social Security Number (SSN)…etc. during the application process and due to this restriction the only information available for the recruiters are candidate name, email ID, phone number and details in the resume. Also there are company policies like if a candidate’s profile is dispositioned for not meeting the job role criteria at any point during the interview process then they cannot apply for any other role for the next six months. But candidates go ahead and apply for other roles through other channels with different email IDs or phone numbers or different resume formats. This way multiple duplicate profiles of same candidate get created in hiring system.

The challenge for recruiters is to identify duplicate records of the candidate as he/ she might have applied for multiple roles through multiple channels at different time intervals. Once the duplicate profiles are identified; the profile which is active in hiring process is kept as survivor profile and the rest of the duplicate profiles are merged. This will make sure the hiring system is clean without duplicates and all the hiring decisions including the interview feedbacks are updated on the survivor profile and same details are updated during offer generation and onboarding process.

The challenge with current manual search process for duplicates with the existing Boolean search query system is; search results are mainly dependent on the keywords or combination of keyworks chosen by the recruiter to search a given profile in hiring system at the same time accuracy of search results using Boolean search is low as search results are many and profiles are not ranked by similarity; leading to manual effort for recruiter to open each profile and confirm duplicate. On an average there are 20% duplicate profiles observed in the hiring system and recruiters spend almost 50% of their effort in identifying duplicate profiles and merging them rather than focusing on sourcing candidates which is the core job role of the recruiter.

**Scope of work**:

The scope of the work is to develop an efficient support system to recruiters to identify potential duplicates which are ranked based on highest similarity to lowest similarity for the profile queried by recruiter. Also the scope of work is confined to candidate profiles in English only and other languages are out of scope.

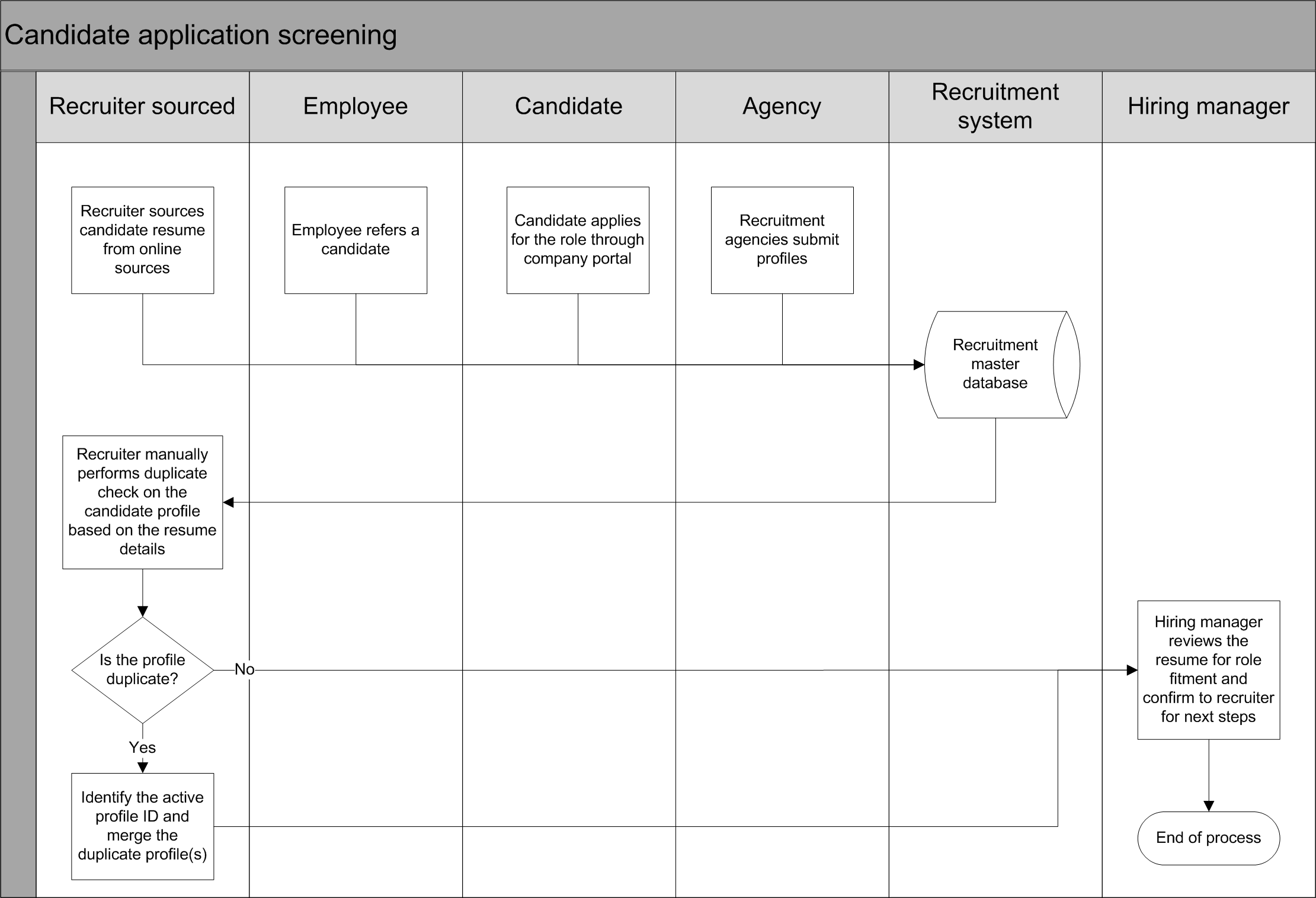
**Objective of the project**:

The objective of this project is to study pairs of candidate profiles manually annotated by human as duplicate/ not duplicate, identify various features from the candidate profiles, use machine learning (ML) to learn from human-labeled data, and predict whether a new pair of candidate profiles are duplicate or not. During the study explore natural language processing (NLP), ML and neural network approaches for classification and suggest a model which has the highest accuracy and AUC (Area Under Curve) in ROC (Receiver operating characteristic) curve.

**Benefit to the organization**:

Effort saved for recruiters which can be utilized in core sourcing activities.

**Business process flow**:



**Background of previous work done in the chosen area:**

There exists no solution apart from the Boolean search, which is manual and more labor intensive as the search results are many and recruiter need to manually open each profile and confirm if it’s a duplicate or not.

**Solution architecture, if any:**

Nil

**Uniqueness of the project**:

Identifying duplicates in a dataset (candidate profiles) where there are no unique IDs exist apart from email ID and phone number of the candidate but even then the challenge is a candidate can have more than one email ID and phone number. Also the other challenge is the data is unstructured (candidate’s resume) which makes this problem even more complex. Due to these challenges it makes this problem unique and the approach in this project is believed to be efficient support system to recruiters in identifying duplicate candidate profiles.

**Potential challenges & risks in doing the project:**

The solution cannot support candidate profiles in languages other than English.

**Resources needed for the project, including people, hardware, software, etc**.:

In this project Python will be used for feature extraction and model development.

**Detailed Plan of Work:**

| **#** | **Task** | **Expected date of completion** | **Names of Deliverables** |
| --- | --- | --- | --- |
| 1 | Data collection and preparation | 19-Sep-2020 | Data preparation |
| 2 | Creation of pairs of duplicate resumes and manual annotation if the pair is duplicate (positive) or not a duplicate (negative) | 26-Sep-2020 |
| 3 | Features extraction from the pairs of duplicate resumes | 10-Oct-2020 | Feature extraction |
| 4 | Testing with Machine learning models Logistic and xgboost | 24-Oct-2020 | Machine learning model development |
| 5 | Testing with Deep Learning models | 14-Nov-2020 | Deep learning model development |
| 6 | Summarization of observation and recommendation | 28-Nov-2020 | Summarization |

**Supervisor’s Rating of the Technical Quality of this Dissertation Outline**

EXCELLENT / GOOD / FAIR/ POOR (Please specify): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Supervisor’s suggestions and remarks about the outline (if applicable).**

Date: 10-Sep-2020 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (Signature of Supervisor)

**Name of the supervisor**: Dr. V. Vijay Kumar

**Email Id of Supervisor**: drvvk\_dean@cvsr.ac.in

**Mob # of supervisor**: 9849452287