|  |
| --- |
| GUS India NER modeling for User Resume summarization  Solution Design Approach |

**High Level Modelling Process**

1. Data Gathering & Storing to AWS S3
2. Primary Data Exploration, we had unstructured data (Resume pdf files) with us, we follow below methods to extract key entities:
   * Text extraction from PDF using textract
   * Custom NER modelling using Comprehend
     + Data set is split into training and testing in 9:1 ratio.
     + Annotated listing is prepared using sagemaker ground truth algorithm.
3. Training the Model
4. Test Model / Model Accuracy
5. Optimized the model using Hyper Parameter Optimization
6. Deployed Final Model using Flask API

**Learnings from multiple Iteration:**

* Must scrape improve the custom model training process by adding new set of data monthly.
* Labelling of data can be done through sagemaker instead of making the annotated entity lists.
* Maximize the data set ensuring resumes ranging from varied fields of work so that custom model is trained on large corpus of data.

**Test Cases**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test Case No.** | **Description** | **Test Data** | **Expected Result** | **Actual Result** | **Status (pass/fail)** |
| #1 | Verify response when a valid Job Description is passed to model API | Resume PDF files | Named Entity from Custom Entity Recognition | Named Entity Identified | Pass |
| #2 | Evaluated Model on Evaluation Data Set of CVs | Evaluation Test Data Set | Precision,  Recall,  F1\_score ,  Accuracy | 'precision': 0.80,  'recall': 0.81,  'f1\_score': 0.80 , ‘accuracy’: ‘0.80’ | Pass |

Solution Summary - Training & Handover Notes

**About Challenge(s)**

GUS Education has a Job Portal, used by recruiters and potential candidates / professionals. Recruiters use the job portal to upload job notifications with descriptions.

Candidate use the job portal to create their profile with their latest skills.

The customer was facing a challenge in profile creation. A large set of parameters like name, education, skill, certification etc was being manually filled by the user.

The customer requirement was that from the resume shared by the user, the key entities should be extracted automatically without the need for the user to fill in all the details manually.

This would help in mapping the right entities to the model for finding out the skills they need to learn by extracting them from their resumes to the right jobs, which enables savings in both candidate’s and recruiter efforts leading to a better user experience. This also reduces the cycle time of profile creation and result generation. This automation was a difficult problem to solve and NER modelling solution was recommended for this.

**Proposed Solution**

MIND discussed the problem with the customer and after analyzing the available data, it was determined that an NLP based ML solution would fit the business problem. Solution flow proposed consisted of the following steps

* Pre-processing of user resume files - i.e. resumes in PDF format are taken from user profile portal and stored in S3. When resume file comes to this S3 bucket then a lambda is triggered which sends this file to textract service for extracting text from PDF.
* Lambda then sends this UTF-8 text file to comprehend service which runs the pre-trained job for extracting the key entities. The set of key entities is divided into two- built-in entities like name, address etc and custom entites like education, certification etc.
* Labelling of resume data to extract custom entities is done using sagemaker Ground Truth labelling algorithm- This provides the labelled data for custom entities for initial training until the model was created.
* Pre-Processing for model training
* Input UTF-8 files to S3 and dividing the data into 90% training and 10% testing data.
* Create a labelling job using custom workflow.
* Set different labelers groups.
* Labelers complete tasks using the labeling UI with assistive labeling features.
* Accurate training data set is ready for use.
* Model Building: Custom NER model is trained using output manifest file from sagemaker. Hyper parameters were tuned iteratively to optimize the model.
* Model Deployment: Deployed the ML model using Flask API, the API takes job description as an input and outputs Technical Skills.
* Models are re-trained on bi-weekly basis.

The overall model development and deployment was done on Amazon Sagemaker. Cloudwatch was used for monitoring the training jobs and resulting metrics. The end point deployed on Sagemaker was called through API Gateway and Lambda function to return a set of labels for given resume. This API was integrated into the customer Profile Creation Portal.

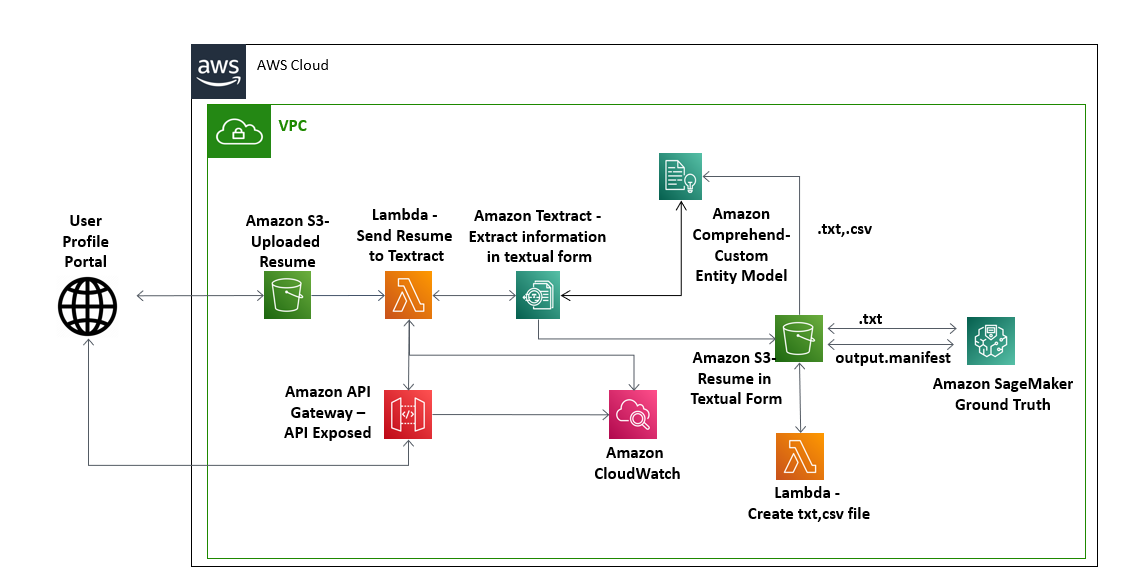
**AWS Services used**

* Amazon SageMaker
* AWS S3
* AWS Lambda
* AWS Comprehend
* Amazon Textract
* Amazon API Gateway

**Solution Outcome**

* With all the customizations and hyper parameters tuning we trained a model with which we were able to extract the key entities from user resume with up to 80% accuracy.
* Saved up to 60% of human effort which was used to manually enter all details.
* Saved up to 90% of training cost by using spot instances for training.

**Architecture Diagram**



**How AWS services helped in building the ML Model for Technical skill Mapping**

**Amazon SageMaker**

Amazon SageMaker is used to create and manage Jupyter notebooks that were used to prepare and process data and to train and deploy the machine learning models. Its built-in Ground Truth Algorithm is used for training of NER Model and is optimized using Amazon SageMaker Hyper Parameter Optimization Service. Final Model is deployed on Amazon SageMaker using Flask API.

**AWS Lambda to handle the backend API calls**

It helped to initialize and validate the input and acted as the backend of the whole task. AWS Lambda lets us run code without provisioning or managing servers. It helped to connect with API’s of Textract and Comprehend.

**AWS API Gateway**

It helped in exposing the model endpoint. This API will be used in backend of the job Portal for enhancing user experience.

**Amazon S3 to store CSV raw documents**

It is an object storage service that offers industry-leading scalability, data availability, security, and perfomance.

**Amazon Textract for extracting text from PDF**

It automatically extracts text, handwriting, and other data from scanned documents, such as PDF's, tables and form.

**Amazon Comprehend for Custom NER modelling**

It is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. It processed resumes in UTF-8 format and developed insights by recognizing the entities, key phrases, language, sentiments, and other common elements in a document. It automatically tests for the best and most accurate combination of algorithms and parameters to use for model.