IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

INFORMATION EXTRACTION FROM DIGITAL RESUME –CONVENTIONAL AND DEEP LEARNING APPROACH

[1]N.Gayathriy ,[2]Jayadevkadri Ananda,[3]Vijaya Vittal B N
[1]Research Scholar, [2]Team Lead,[3]Credit Risk Compliance Manager
[1]Department of Computer Science and Engineering
[1]Coimbatore Institute of Technology, Coimbatore, India

Abstract: Applicant Tracking System has become a popular toolin recent times for recruitment and talent acquisition process among organizations.It's "Easy apply option "and automated resume summarization are the most important features that allowcandidates to just upload the resume without additional information's unless required and provides the recruiters with brief report. However, its conventional rule-based method may lead to false positive or false negative predictions which compromises the quality of automated resume screening process.Beginning with the traditional approach of manual resume screening process, in this research studywe propose a generalized deeplearning-based information extraction model to locate & classify entities across digital resumes. Finally, a recommendation model has been built and deployed using Flask application thereby aiming to provide an end-to-end solution for automatic hiring process**.**

***Index Terms* – Information Retrieval, IR, resume parsing, Regex, NER, Spacy, Conditional Random Field, CRF, Bi-LSTM and ensemble model.**

I. INTRODUCTION

Recruitment is one of the most important processes for anyorganization. It marks the potential for growth and developmentof a team or an organization in general.It is estimated that on an average for every Job posting, the hiring organization receivesmore than 250+ resumes. After receiving the application, theHR or recruitment team is tasked to go over everyresume to short list deserving candidates to the next interviewprocess. However, asresumesareunstructured and containunique formats it makes it hard to read all resumes with the same level of consistency.To solve the issue of quality, organizations have resorted to use Application Tracking Systems (ATS) which decreases turn over rates, eliminates costly screening calls, makes more confident hiring decisionsand integrates with in-house talent management systems. Butextracting information from each resume with the traditional string match using Regex, dictionaries & Rule based approaches works well but has its limitations.The major drawback is in identifying a non-critical entity as an importantentity (falsepositives) or identifying & classifying important entity as an on-important entity (falsenegatives).Hence a smart screening model is required to extract the rightinformation from theresumeand thereby addvaluetotheentirerecruitment process.

II. PROBLEM DEFINITION

The problem of accurately reading/parsing are sume to extract the right information and classify into the right entity(Name,Email,Contact,Skills,Education/Qualification,Designation,Company,Institutions&Experience) to aid in decision making.The potential solution should minimize the false positives & false negatives in identifying the right class.The final recommendations should create value additions that enables:

- Resume–JobDescriptionfitmentscore.
- ResumeRanking.

- Resume Classification.

III. MODELLING METHODOLOGY AND VALIDATION

A. *Methodology*

The Extraction process was broadly classified into four majorstages:

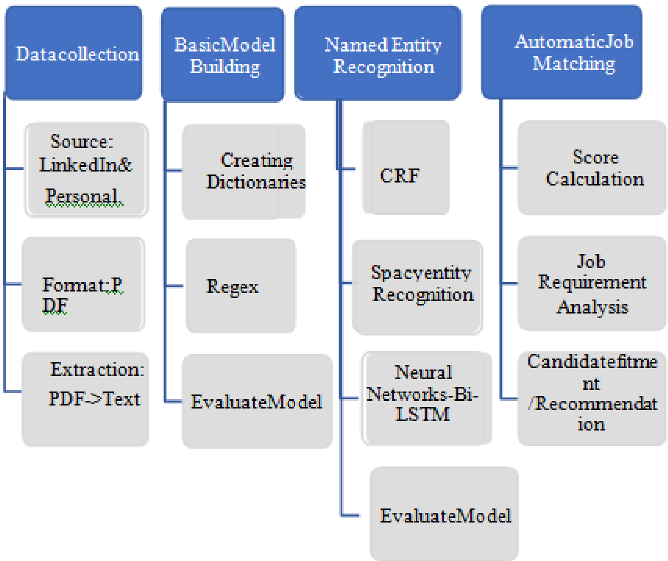


FIG 1: EXTRACTION PROCESS FROM DIGITAL RESUME

A. *Evaluation Metrics*

Precision, Recall & F1-Score were used to evaluate the model.Consider the confusion matrix below, where Cij represents thenumber of data instances which are known to be in the grouping i(truelabel) and predicted to be in groupj(predictedlabel)

PREDICTED LABEL(J)

Negative Positive

True Label(i)			
	Negative	C_{00} True NegativeTN	C_{01} FalsePositive FP
	Positive	C_{10} False Negative FN	C_{11} TruePositive TP

FIGURE 2: CONFUSION MATRIX

Accuracy represents the number of correctly classified entity labels (i.e.,thepredicted orextracted labelsare compared with the annotated data to build the classification matrix).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

However, Accuracy may not be a good measure if the data labels are not balanced and given for our scenario the "Others" non-important entity that consists of almost 87% of the overall corpus will result in high accuracy% but the overall model will be poor in its predictions. To eliminate this, the project focuses on precision, recall & f1-score. As the objective of the model is to minimize both False positives & False

Negatives, equal importanceis given. Hence, f1-Score isused to decide the best sequence to sequence classifier.

Precision,it calculate show precise/accurate based on the predicted positive,how many are actually positive.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall calculates how many of the actual positives the models are able to identify as positives.

(3)

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Finally, f1-score, calculates the harmonic mean between precision & Recall.

IV. DATAPREPARATION

A. Data Collection

Digital resumes are the primary data sources.Due to intellectual property behind each resume carefully moderation was done in the collection stage to select the required resume for training & analysis. Different file formats like doc, docx,txt, pdf etc. were obtained. For this study, we considered the dataset of around 260 resumes in PDFformat.

B. Data Preprocessing

Due to the nature of the project, the input data in the resumes had to be extracted first. The Apache Tika toolkit was used asthe parser. This was decided based on scalability in mind, asthe Apache toolkit has the ability extract metadata and textfrom over thousands of file formats including (PPT, doc, Pdf etc.). Apache Tikaisa content detection and analysis framework that is written in Java and stewarded at Apaches of software foundation. Some of the major advantages of the toolkit are Unified parser interface, Low memory usage, Fastprocessing, Flexible metadata., Language detection.

Post, the parsing stage extracted data was passed through pre-processing phase.This stage will clean and sort out any issuesin the data related to training the models. The following pre-processing steps were used.

- Removal of punctuations, symbols, hyperlinks, Nextline etc.
- Removal of stopwords(NLTK library has a corpus of frequently used stop words)
- Lower case conversion.
- Finally, the cleaned data was stored as a pandas dataframe.The data was also exported to aid in manual annotation.

C. *DataAnnotation*

Data annotation is the process of labelling texts, videos, images and other content.The process is mainly needed in deeplearning models to train and help the model to understandtheinput and label these input or predict outcome. The process helps the machine to understand and memorize the input patterns.Since the objective of the project is to identify important information from the rest, the below entities were selected for training and annotated.

- Name
- E-mailID’s
- Contactinformation.
- Skills
- Education/qualification
- Institutions
- Designation
- Company
- Yearsofexperience

Two of the main annotation for mats being Offset & BILOU method are considered:

Offset method is the basic form of Spacy input, that represents the entities with their starting and ending index numbers.Theformatisshownbelow:

[(Text,{'entities': [(176, 187, 'Name'), (605, 639, 'Institutions'), (669, 682, 'Institutions'), (199, 212, 'Company'), (237, 250, 'Company'), (271, 284, 'Company'), (364, 377, 'Company'), (488, 506, 'Company'), (79, 133, 'Skill'), (706, 715, 'Skill'), (645, 652, 'Skill'), (293, 317, 'Date'), (396, 414, 'Date'), (533, 561, 'Date'), (656, 667, 'Date'), (719, 730, 'Date'), (640, 643, 'Qualifications'), (683, 691, 'Qualifications'), (188, 195, 'Designation'), (252, 259, 'Designation'), (285, 292, 'Designation'), (388, 395, 'Designation'), (507, 532, 'Designation')]})]

FIGURE 3: BASIC FORM OF SPACY INPUT

BILOU method stage very single token with respective entitytags:

The first task is to create manually annotated training data to train the model. For this purpose, a manual Annotation Toolhas been built as a part of the study. Here, the 260 resumes collected were manually labelled and classified in Excel. After the labelling task wascompleted, python query toread andconvert the binned entities to the spacy offset format shown above was created. Spacy's Gold Parse function was used to convert the offset annotations into BILOUform.

TABLE1: BILOU METHOD

TAG	DESCRIPTION
B EGIN	The first token of a multi-token entity.
I N	An inner token of a multi-token entity.
L AST	The final token of a multi-token entity.
U NIT	A single-token entity.
O UT	A non-entity token.

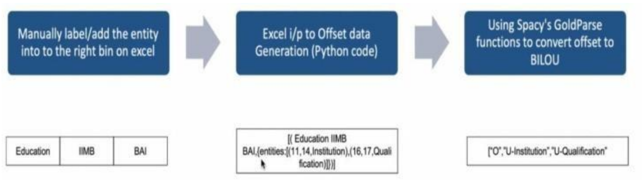


FIGURE 4: EXTRACTION OF DATA

D. Feature Extraction and DescriptiveAnalytics

The feature variables used for model building are Skill, Qualification and experience. They are extracted and analyzed to get better understanding and knowledge of the entities.

WordCount:

After parsing through all the resumes,a basic descriptive analysis was conducted to see theword distribution. As shown below, the majority of the words are stop words and needs to be removed.This also highlighted few additional non-important words like link sand other characters not needed for the analysis.Stop words such as and ,of ,in have higher representation and hence need to be excluded.

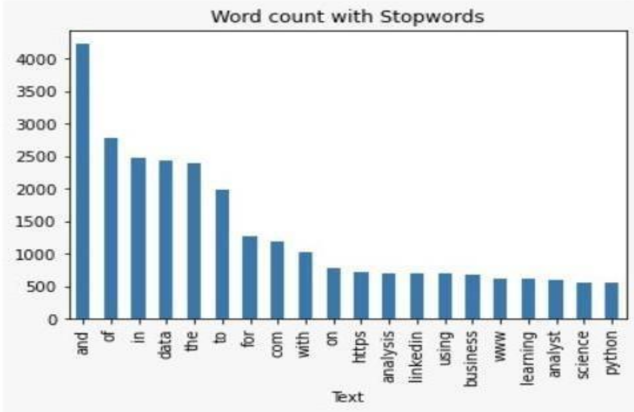
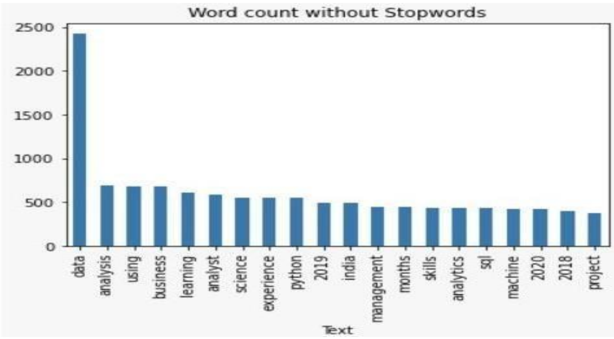


FIGURE5: FILTERATION OF WORD COUNT USING NLP

Stop words are words which are filtered out before or afterprocessing of natural language data. Most common being the,is,at,which,on,a,about,all,etc.For the purpose of analyzing text data, these stop words might not add much value to the meaning of the documentor stop words are excluded from the given text so that more focus can be givento those words which define the meaning of the text. After there moval of stopwords we can see that actual critical Key words being better represented in the distribution chart below.



Annotation Distribution

To train a deep learning named entity model it is also important to understand the distribution of the important Tags.

From the below distribution we can see that almost 87%of thewords in the corpus are non-important entities and are taggedas “Others”. Only theremaining 13% need to be located andthen classified by the information extraction model & NERmodels.

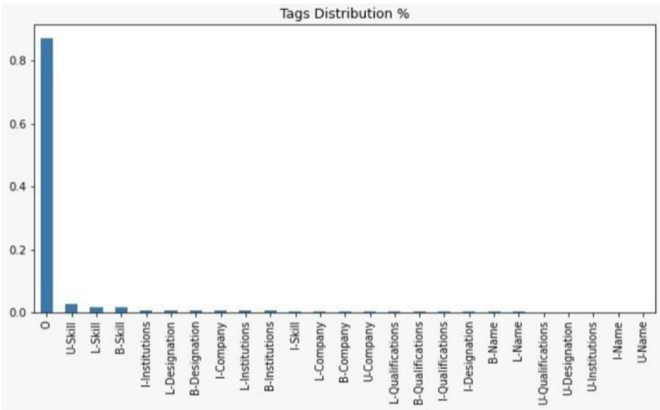


FIGURE6:ANNOTATIONDISTRIBUTION

Designation ofcandidates:

Based on the candidate's resumes the data of designations ofthe candidates are presented below in a barchart, where in each candidate might have multiple designations.

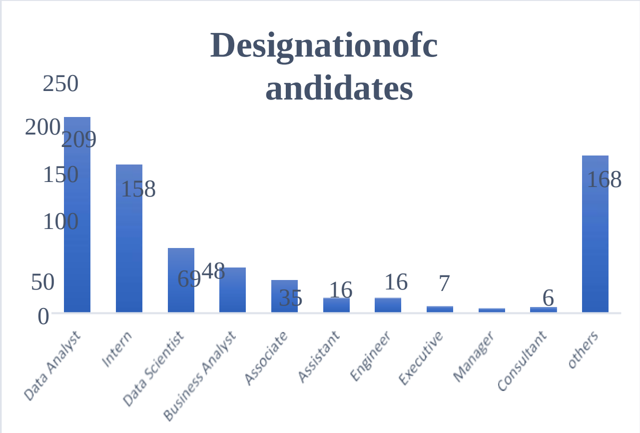


FIGURE7: DESIGNATION OF CANDIDATES

Qualification background of candidates

Based on the candidates resumes the data of qualification background of the candidates are presented below in a barchart

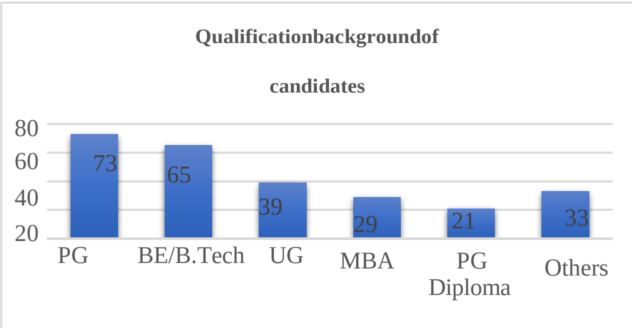


Figure 8: Qualification of background candidates

V. MODELS

A. *Information Extraction using Regex, Dictionaries & Rule-Based.*

Rule-based NLP approaches, such as the one used in the study, are based on an expert system of rules hand-coded byhumans. Even though creating a rule-based system is a time-consuming process and requires domain knowledge but are reliable and useful to automated data processing.A relative study on sentiment analysis done by Dwivedi et al. [2] was based on rule-based model (RBM) and it was found to give better result compared to other sentiment lexicons mentioned.Below is the methodology involved in building base model for information extraction using Regex & Dictionary.

Regular expressions (Regex) are special string for describing a search pattern, they are similar to wild cards in functionality.Regex are used by string search algorithms to find or find & replace operations on strings. Regex are known to be used insearch engines, search and replace dialogs of word processors and text editors. Python library re was used to perform the entity extraction as shown below. Regular expressions were used in finding Email-id's,contact information,dates,Company & Institutional in formation based on suffixes.

Dictionaries are a large set or list of identical entities that together with Regex can be used to lookup for a given wordand extract the information. Dictionary for Skills, Languages, Qualifications & designation was created froma sampleof100 resumes and each string on the specific dictionary wassearchedandextractediffound.

To extract the name several approaches were used, howeverthebest approach that resulted in 70% accuracy was usingrule based i.e., extracting strings with the biggest fonts. Asfrom the initial observations it was found that the name of thecandidate in the resume was the largest font. Libraries such asApache &PyPDF have the ability to extract metadata such assize, font&color.Using the semeta data,Word with the highest size was tagged a sname.

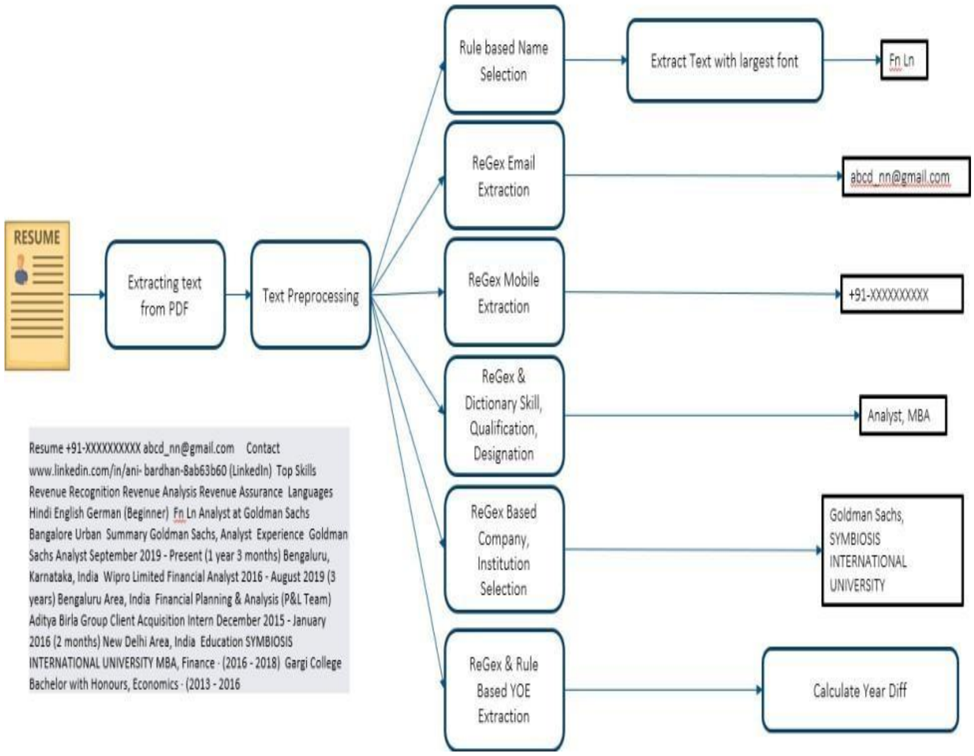


Figure 9: Meta data Extraction

The screenshot of entities extracted are given below. Hereregex & rule-based method for retrieving email,company details was successful with universal format for email and predefined suffixes as private Ltd,.com etc.for company.

Email	Institute	Qualification	Company	Overall_Experience	Relevent_Experience	Spoken Languages	Skills
thanisornsr@gmail.com	[CHALMERS UNIVERSITY OF TECHNOLOGY, Gothenburg...	Information Not Found	[Design Engineer Gerenga Service (Thailand) Co...	6	2	[Thai, English, Swedish, Portuguese]	[Project Management, Deep Learning, Machine Le...
Information Not Found	[SRM University]	Information Not Found	Information Not Found	4	6	Information Not Found	[Leadership, Communication, R]
Information Not Found	[Siddaganga Institute Technology, Tumakuru (Sl...	Information Not Found	[TATA Consultancy Services Ltd, Bangalore.]	7	8	[Kannada, English]	[Communication, Analytics, Deep Learning, Mach...
Information Not Found	[Data ByteDance • IIM Shillong, MBA IIM Shillo...	[Mba]	[Outlook Publishing (India) Pvt. Ltd., Outlook...	6	3	Information Not Found	[Teamwork, Data Analytics, Analytics, Marketin...

Figure10: Here regex & rule-based method for retrieving entities

The average Precision, Recall and F1 Score came out around 55% with individual percentile given below:

Table2:

Entity	Precision (%)	Recal l (%)	F1(%)
EMAILID	98.00	98.00	98.00
Mobilenumber	87.00	84.00	82.00
Institute	62.84	59.42	61.09
Company	25.42	14.19	18.21
Skills	40.43	32.02	35.74
Languages	78.93	79.33	79.13
Qualifications	8.00	5.00	6.25

B. Named Entity Recognition

Named entity recognition(NER) or Entity chunking,extraction or identification is a task of locating & classifying aword/token in a sentence into a set of pre-defined classes ortags. It is a sequence-to-sequence prediction, labeling, tagging model. An entity can beasingleor series of words that refer or related to a class.Consider the below example:

Sam studied at IIM Bangalore.
Name Institution Location

In the above example, Sam is the name of a person and is tagged as name entity.IIM is the name of the institution, therefore tagged as Institution and finally Bangalore isthelocation.

NER architecture is fairly simple, consisting of 2 majoractivities:

- Detectanamedentity.
- Classifytheentity.

According to a transitioned based approach borrowed fromshift – reduceparsers, every NLP problem can bebroken into4important sections:

Embed: This stage starts off by first converting text or stringsinto tokens (word or sentences). After tokenizing, the tokens are converted to numeric alid's followed by embedding (representing to kens as a vector of numbers).

Encode: This stage concentrates on learning the hidden features,language from the embedded inputs.The initial embedded id's are now converted to sequence or pattern matrix.

Attention: All important features from the sequence matrix are extracted.

Predict: Finally,a classifier,predicts the right class or classes depending on the problem

AsNER is a Sequence-to-Sequence model, every input token needs to be classified into one of the pre-defined classes. Hence, a multi categorical classifier is used in the prediction stage.Several libraries exist, namely NLTK, Spacy & Stanford NER. Jing Li [4] provided a complete survey on deep learningbased NER solution which included the background oftheNER research, a brief of traditional approaches, current state-of-the-arts, and challenges and future research directions.In the current context,we also explored probabilistic models such as Conditional Random Fields (CRF), deep learning models based of Bi-LSTM's and finally spacy NER implementations.

C. Information Extraction using CRF.

Conditional Random Field is the class of statistical modelling method often applied in pattern recognition and machine learning and used for structured prediction.Conditional Random Fields is a class of probabilistic models best suited to prediction tasks where contextual information or state of the neighbors affect the current prediction.Dongyang Wang[3] proposed a multi-modal neural network that applies CRF for sentence annotation in sequence and it efficiently solved long-distance dependency of text semantics, shortening network training and predicted time.

It satisfies the property:

"When we condition the graph on X globally i.e. when thevaluesofrandom variablesin X isfixed orgiven, all the random variables in setY follow the Markov property$p(Y_u/X,Y_v, u\neq v) = p(Y_u/X,Y_x, Y_u\sim Y_x)$, where $Y_u\sim Y_x$ signifies that Y_u and Y_x are neighbors in the graph."Avariable's neighboring nodes or variables are also called the Markov Blanket of that variable.

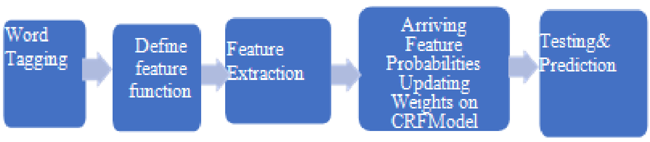


Figure 11: Feature Extraction and Prediction

Word Tagging:

An entity or a part of text that is of interest would be of great use if it could be recognized, named and called to identify similar entities. A CRF isa sequence modelling algorithm which is used to identify entities or patterns in text, such as POS tags.B. Veera Sekhar Reddy, Koppula SrinivasRao [7] used CRF and Active Learning Procedure for his research on NER and proved that it is both more efficient and requires less manually marked training samples.This mode lnotonly assumes that features are dependent on each other, but also considers future observations while learning a pattern.Interms of performance, it is considered to be the best method for entity recognition.

Since these models take into account previous data, we usefeatures which are modelled from the data to feed into the CRF. These feature functions express certain characteristics ofthe sequence that the data point represents, such as the tagsequencenoun -> verb -> adjective. When y is thehiddenstate and x is the observed variable, the CRF formula is givenby:

$$p(\mathbf{y}|\mathbf{x}) = \underbrace{\frac{1}{Z(\mathbf{x})}}_{\text{Normalization}} \prod_{t=1}^{T} \exp\left\{\sum_{k=1}^{K} \underbrace{\theta_k}_{\text{Weight}} \underbrace{f_k(y_t, y_{t-1}, \mathbf{x}_t)}_{\text{Feature}}\right\} \tag{4}$$

We have trained a CRF using feature functions to predict POStags and testing the model to obtain its accuracy and othermetrics. To train a CRF, we will be using the sklearn-crfsuitewrapper.

Feature Selection:

The features considered are:

- The word
- The word in lowercase
- Prefixes and suffixes of the word of varying lengths
- If the word is a digit
- If the word is a punctuation mark
- If the word is at the beginning of the sentence (BOS) or the end of the sentence(EOS) or neither
- The length of the word -no. of characters (since shorter words are expected to be more likely to be long to particular POS e.g., prepositions or pronouns).
- Stemmed version of the word, which deletes all vowels along with g, y, n from the end of the word,but leaves atleast a 2 character long stem.
- Features mentioned above for the previous word, thefollowing word, and the words two places before andafter
- Features are qualitative functions and can differ frompersontoperson.

Feature extraction:

Ayishathahira etal. [1] used neural networks and CRF to extract and segment details from the resume and the output was outperforming than other neural networks.

Here there are 2 components to the CRF formula:

Normalization: We observed that there are no probabilities on the right side of the equation where we have the weights and features. However, the output is expected to be a probability and hence there is a need for normalization.The normalization constantZ(x)is a sum of all possible statesequencessuchthatthetotalbecomes1.

Weights and Features: This component can be thought of asthelogistic regression formula with weights and the corresponding features. The weight estimation is performed by maximum likelihood estimation and the features are defined by us.

Now that we have a feature extraction function, we are nowready to pass the data in to the function. Which helps them toconvert it into sentences and we now proceed with training themodel. Let us train the CRF on the processed train set. c1 and c2 are the parameters for L1 and L2 regularization respectively, and they usually range from 0.01 to 0.01. They can bet weaked to give better results in model performance and the lowest loss were considered here.

Thevariousstagesanddata/informationextractionrepresentationofCRFmodelisgiven:

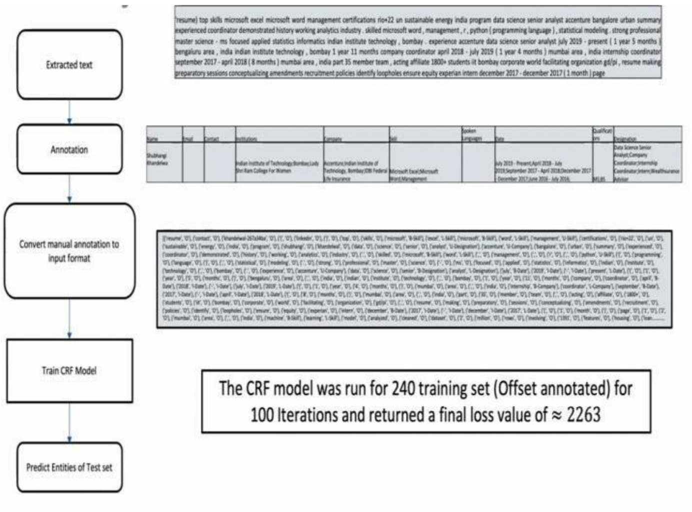


Figure12: CRFModel

The model is trained using L-BFGS algorithm with 240 resumes as training set and the model incurred a final loss of 2263.It had captured well for skills, company and designation. The output snapshot is provided below for better understanding.

	Entity	Tag
0	microsoft excel microsoft word management	U-Skill
1	analyst accenture	U-Company
2	microsoft word	L-Skill
3	python	U-Skill
4	accenture	U-Company
5	senior analyst july 2019 - present	L-Date
6	april 2018 - july 2019	L-Date
7	internship coordinator september 2017 - april ...	L-Date
8	december 2017 - december 2017	L-Date
9	machine learning	L-Skill
10	aid	U-Skill
11	2017 - 2019	L-Date
12	2014 - 2017	L-Date

Figure13: CRF Model

The f1 score is given below and it was found out to be around 77% good in extracting entities.

Entity	Precision(%)	Recall (%)	F1(%)
Skills	70	71	71
Date	98	98	98
Designation	90	81	85
Institutions	94	84	89
Company	96	82	88
Qualifications	96	61	75
Name	100	83	91

Test scores Accuracy and F1 values of themodel stood at 86%. We can see that the model has better accuracy of around 71% on the train set and 86% on the test set. Playing aroundwith the L1 and L2 regularization parameters might help giveusabetterperformanceonthetestsetandpreventoverfitting.

D. *NER using Bi-LSTM-CRF*

A bidirectional LSTM, or Bi-LSTM,is a sequence processing deep learning model that consists of two LSTMs. They are aspecial class of RNN's (Recurrent neural networks) consisting of 2 LSTM's one tracing the sequences in the forward direction (lefttoright) and another LSTM tracing in the reverse order (Right to left). This enables the model to learnnot only historical patterns in predicting the current entity butalso understand how the future or forward context can be used for the prediction. Tensor Flow was used for the below implementation, this requires several additional pre-processing steps to meet the Tensor Flow requirement.

After converting the pre-processed text, word tokens were created, and the following pre-training steps were carried out

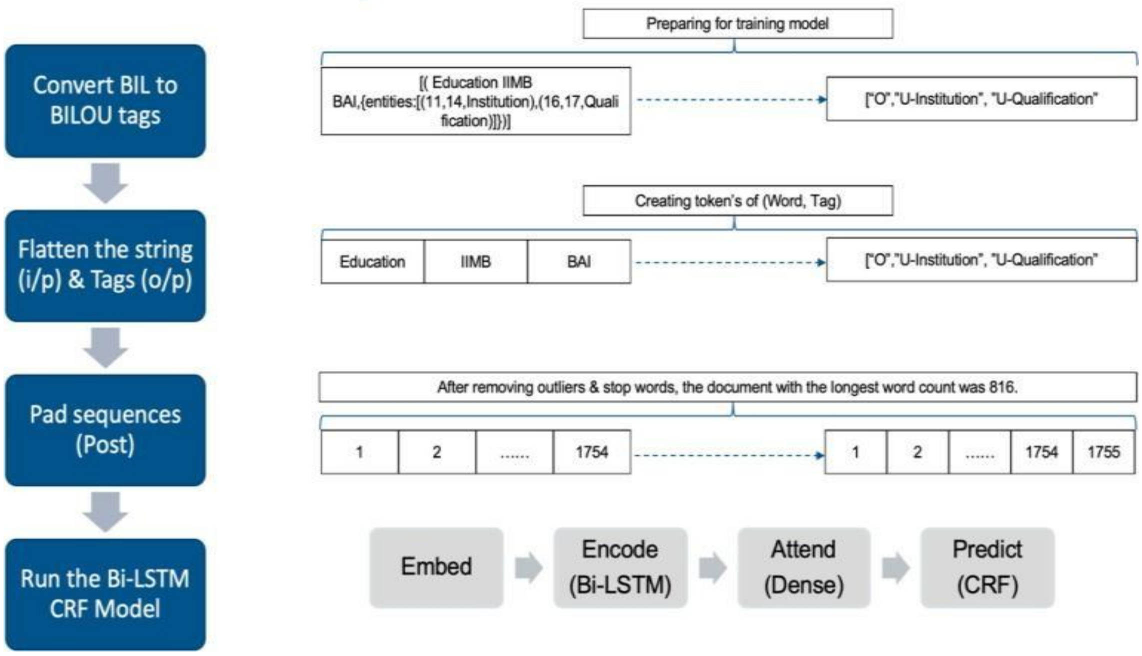


Figure 14: Trainingsteps forCRFModel

Step 1: The annotation tool built using python was to convertexcellabeledentitiesintooffsetmethod entityannotation.Thisneedstobeconverted intoBILOU form asLSTM'srequire sequence by sequence representation of token to tag.The Spacy'sGoldParse library was used for the conversion.Fewentitytoken–tagwerelostduetomis-alignmentissues.

Step2: Since Bi-LSTM's take in data as a flattened sequence, word tokens and taglist were converted to a list.

Step3: Tensor Flow accepts only equal length inputs, therefore, post-padding was done to standardize the inputs.The maximum length resume consisted of 1755 token/words.Each of the tokens are mapped to one of the 25 tags/entities.Everyotherresumehadtobepaddedwithdummylabels.

Step4: Run the Bi-LSTM-CRF model, this is representative ofthetransition-basedapproach.

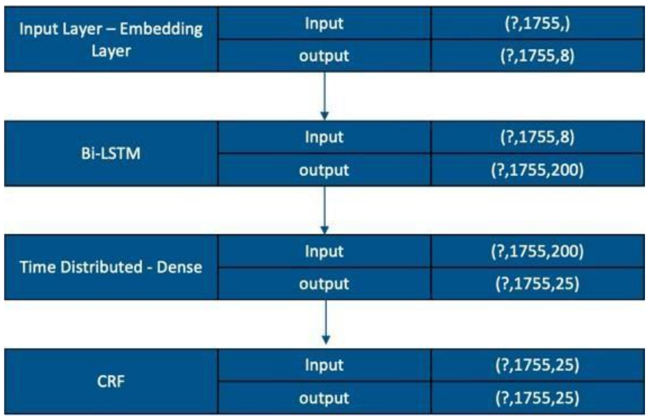


Figure15: Bi-LSTM-CRFModel

Embedding layer: Maximum layer specified was 1755 with the padded sequence.This layer will transform the input layer

into a vector of 8 dimensions (64, 120 & 180 dimensions weretestedtoo)

Bi-LSTM layer: Two separate LSTM's are used forward &Backward tracing. Both these take the embedded output andreturnasequencevector.

Time distributed– Dense layer: Aswe are dealing with an RNN with many to many relationships i.e. output for everyinput sequence. Time distributed layers allow dense operation for every output over every timestamp.If this layer is removed it will result in only one output for all the input sequence.

CRF layer: After the Bi-LSTM learn the intrinsic language and patterns, CRF models are used to extract certain constraints in the final predictions.

E. NER using Spacy

Spacy is an open-source library for performing industrial strength Natural languageProcessing tasks including NER. It is specifically built for creating applications that process largevolumes of text. Darshita Kumar [5] developed a generalized NER framework which lets users build training models on top of the existing spaCymodels to allow for name dentity recognition on their text data.The framework takes a configuration file which contains model name, model size and hyperparameters, along with annotated data in JSON format asinput, and returnsa customized spaCy modelas output.Some of the features of spacy are tokenization, Part-of-speech tagging (POS), text classification & NER. The default spacyNER model is built to identify basic elements such as Person, Company, Time, Location, Organization etc. Apart from thisspacy also allows users to train a New NLP pipeline suited for customuse cases.

Spacy NER is based on the same principles of transition-basedapproach.

Embedding: Tokenized words are embedded using hashing trick or Bloom embedding which isa Compact embedding structure. This may result in colliding & potential same vector representation. This is avoided by, Repetition of embedding and the total of the seiterations are considered for training.

Encode: Post embedding Sequence of words are encoded intoa sentence / Sequence matrix. Context is used in building thesequencematrix.CNNisusedforEncoding.

Attend: Identify the informative section from the Sequencematrix.Returnproblemspecificrepresentation.

Predict: Use of deep ANN to gather inference and predict.

Training in spacy is an iterative process in which model predictions are mapped against reference annotations in order to estimate the gradient of loss. The gradient of loss is the n used to calculate the weights through back-propagation.

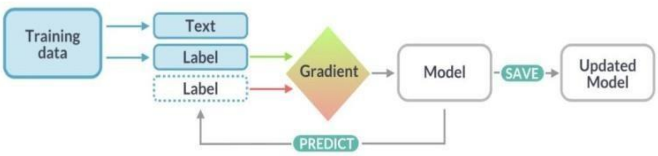


Figure16: Calculation of Gradient Loss

To custom train the spacy model, the manually annotated excel file is converted to a spacy offset annotation format. Theannotated & labelled data is passed to the spacy pipeline by first removing the remaining components such as Tokenizer,

POS tagger & text classifier. The list of entities is provided and trained with the annotated text to generate the model. Spacy allows users to tune the model the below hyper parameters.

Epochs: No. of passes or iterations the training data is used in updating the weights.

Dropout layer: To avoid over-fitting, several random neurons are dropped every epoch. This makes the model predictionharder.

Finally, new or unseen texts can be provided and based on thegradient loss the weights learn the pattern & context and make relevant predictions.

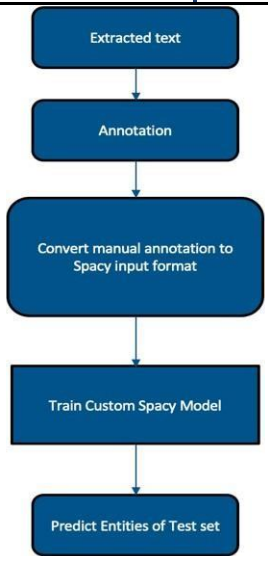


Figure17: Predictions based on Pattern & Context

Multiple runs using different combinations of Epochs & Dropout rate was considered.

Model with 30 Epochs & 0.5 dropout rate, provided the belowloss comparison chart (Training loss vs Test loss). From the chart below we can observe that with a high drop out rate the

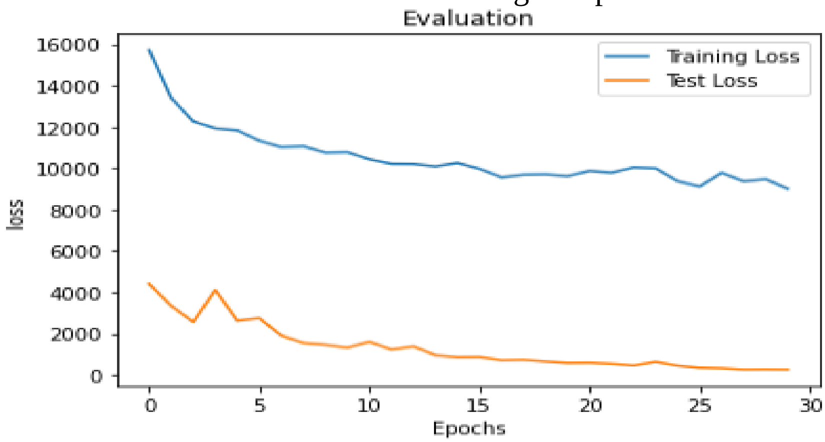


Figure 18: Training loss vs Test loss (high dropout rate)

Table 3:

	Test Set		
Entity	Precision	Recall	F1
Company	99.29	99.28	99.29
Skill	95.52	100.00	97.71
Name	100.00	100.00	100.00
Designation	98.82	98.91	99.41
Qualifications	97.83	100.00	98.90
Institutions	100.00	100.00	100.00
Overall	97.76	100.00	98.86

Model fails to generalize the predictions across the training & test set. Due to random neurons being dropped every pass, the model misses out on meaningful information. Also, the test loss seems to be considerably lower than training loss, suggesting unknown fit which is caused as a result of moreeasier prediction cases on test set when compared to training set (55%ofthetestsetresumewerelinked-inprofiles).

Model with 30 Epochs & 0.1 dropout rate, provided the belowloss comparison chart (Training loss vs Test loss). From thechart below we can observe that the model generalizes well,though there is still some underfitting.

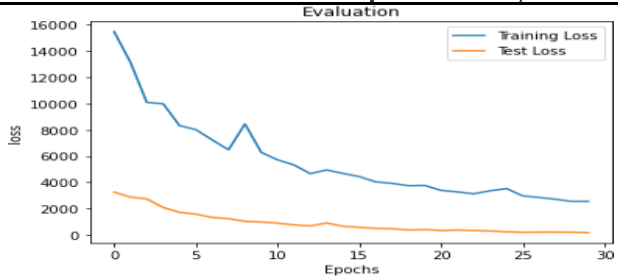


Figure 19: Training loss vs Test loss(generalized well)

Table3:

	Test Set		
Entity	**Precision**	**Recall**	**F1**
Company	100.00	100.00	100.00
Skill	99.48	99.48	99.48
Name	100.00	100.00	100.00
Designation	98.81	98.81	98.81
Qualifications	97.83	100.00	98.90
Institutions	100.00	100.00	100.00
Overall	99.38	99.58	99.48

The above model with 30 epochs & dropout rate of 10% werechosen as the hyper-parameters for the final model. The said model averaged an f1-score of 95%.

VI. RESULTS

Base Model:

The average f1-score using Regex, Rule based & dictionaries was around 55%.Regular expression for searching and extracting entities with standard form at is preferred but performs poorly when the entity is dynamic in its occurrence.While searching words from dictionaries is directly proportional to the depth of dictionary. Lesser the dictionary size less the overall accuracy in detection.

Information Extraction using CRF:

The average f1-score was around 77%. This is most widelyused due to the fact it can handle multiple input features suchas cases, parts of speech tags, data type etc. to build context.The major drawback in CRF is its inability to identify context from future word occurrences.

NER using Bi-LSTM-CRF:

Different Resumes have different length, hence padding are resume based on longest sentence creates bias while predicting shorter corpus. Tensor Flow & Word 2 VecGlove Embedding work well the definite or limited corpus. Out ofvocabulary words are tagged with same vector, hence will beidentical for the model to make prediction. Bi-LSTM'saregood at understanding intrinsiclanguage; however, a resume is a semi structured sentence with structure not similar to POS. Performs poorly with small training set. Performs well with BILOU tags, since for the current exercise Offset tags were converted to BILOU few important tagswere lost in conversion.As this model failed to produce required predictions due to the problems stated, this method was put on hold for the current run down.

NER using Spacy:

Spacy NER uses Bloom embeddings that remove the effects caused by non-vocabulary words and reserves the vectorization to a pre-determined range.To get her coupled with 1D CNN to gather or learn patterns make sit a good contender for the best NER model. Dropout rate =0.1% works best.30 Epochs with a final test loss of 275 & train loss of 3000.Mis aligned text were dropped during scoring.Hence,test set has higher accuracy in comparisons with train set.Also ,the average f1- score was recorded at 95%.

An ensemble model of Regex & Spacy NER will be used to extract the required entities to feed data into the recommendation system.

VII. RECOMMENDER SYSTEM

On the practical aspect, as the next step, the extracted entities are incorporated into an application. Recommender systems are a nessential part of today's businesses. The recommendation model is designed to take job description and resume as input and provide the list of resume which are closest to the provided job description.With the comprehensive requirements for a job, the extracted entities of candidates, the prediction of how likely a candidate is a good fit for the job is done. The recommendation engine assigns a fit score to each candidate and ranks them. Amruta Mankawade[6],developed and used a recommendation system that uses cosine similarity algorithm for online job searching to lessen this tedious task.In this study,an architecture that used has been proposed to extract the most suitable professions based on the resume of the individual.

A. Architecture

The suggested recommendation tool uses the ensemble modelas designed earlier to extract the important information from resumes & job descriptions. Information retrieved from resume include Email ID's, dates & contacts using Regex and Names, skills, qualifications, Institutions, Companies from Spacy NER. The same NER model is used to extract skills & qualification requirement from the JD.

Finally, a similarity score such as cosine similarity calculate show close the Job requirements is with the individual profiles.Cosine similarity is the measure of similarity between two vectors, by computing the cosine of the angle between two vectors projected into multi dimensional space.It can be applied to items available on a dataset to compute similarity to one another via keywords or other metrics. Similarity between two vectors (A and B) is calculated by taking the dot productof the two vectors and dividing it by the magnitude value asshown in the equation below. Cosine Similarity score of two vectors increases as the angle between them decreases.

$$cos(\theta) = \frac{A.B}{||A||.||B||} = \frac{(\sum_{i=1}^{n} Ai.Bi)}{\sqrt[2]{\sum_{i=1}^{n}(Ai)^2}\sqrt[2]{\sum_{i=1}^{n}(Bi)^{\wedge}2}}$$

To illustrate the above, consider the JD requirements as shown below:
JD=['python','java','spark','AWS','Regression','Neural

Network']
Below are skills extracted from candidates 1,2&3 and the corresponding cosine similarity scores.
Candidate_1= ['AWS' ,'Regression','NeuralNetwork']Score=75.6%
Candidate_2= ['python','java','spark','AWS','Regression']Score=84.5%
Candidate_3= ['finance','excel','project management']Score=0%

Therefore, we can see that Candidate 1 & Candidate 2 both have good skill match when compared to Candidate 3.

Using the above application of NER & Cosine similarity score the proposed recommendation model architecture is shown below:

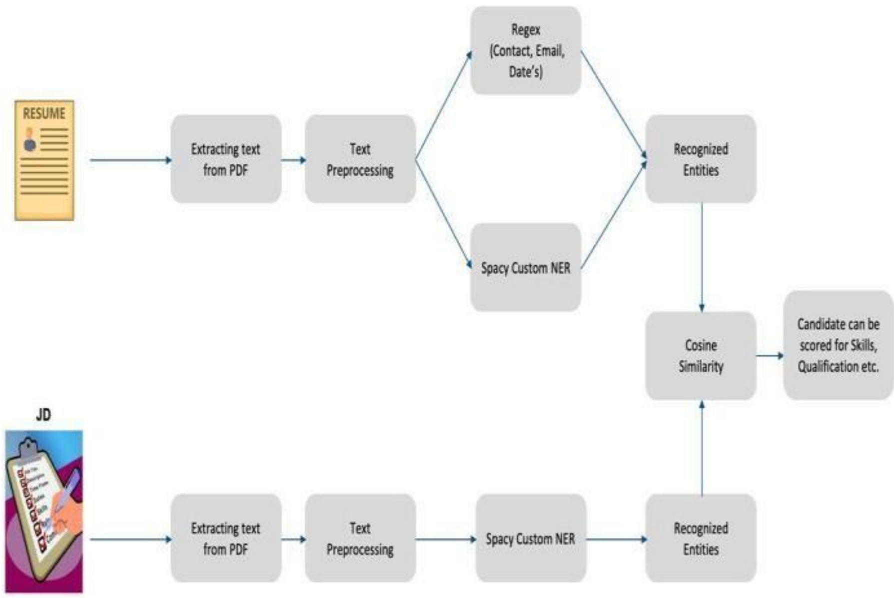


Figure20:Proposed Recommendation model

The recommender engine solves the tedious task of identifyingthenecessaryinformationandcomparingitwithjobdescription. These way recruiters can quickly identify their topcandidatesandengagethemfaster.

B. DeploymentPlan

The recommendation engine built was deployed as a HTML web tool using Flask application. Flask is a small light weight Python web frame work that provides useful tools and features that help increating web applications in Python easier. The flask application allows us to show case the power and functionality of the recommendation engine and communicate its importance to the company. Below is the screen shot of the landing page,here the user has two options.

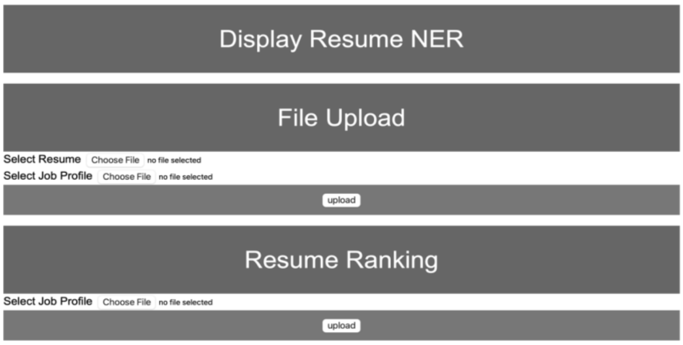


FIGURE21: CRF MODEL

File upload:

This is an 1v1 comparison window. The recruiter or the enduser can select aresume of interest and the respective JD. The model extracts the relevant entities in this scenario the Skills and calculate and display the similarity between the two. The rightside of the window displays the JD entities such as skills & Qualification, the right side the candidate resumes entities.

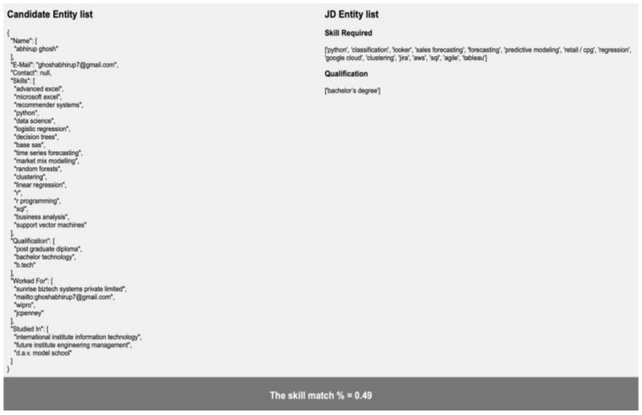


Figure22: Skill Match similarity score bottom.BestCandidateselect:

This is the best feature and is able to extract entities from any number of candidate's resumes and match their Skills & Qualifications with the JD of interest. The skill & Qualification required as per the JD havebeen assigned the same weights.In case of any customization of importance can be easily incorporated. The result displays a table with all candidate information along with Skill & qualification scores. The best candidates can be selected based on informative criteria and further analyzed using the 1v1 view as mentioned above.

Candidate Entity list

	Filename	Phone	Email	EXP	Company	Institutions	Skill	Qualification	Skill_score	Quali_score	Total_score
0	Candidate 2.pdf	None	[ghoshabhirup7@gmail.com, mailto:ghoshabhirup7@gmail.com]	2	[wipro, mailto:ghoshabhirup7@gmail.com, jcpenney, sunrise biztech systems private limited]	[international institute information technology, future institute engineering management, d.a.v. model school]	[clustering, business analysis, decision trees, base sas, time series forecasting, r programming, random forests, r, recommender systems, market mix modelling, microsoft excel, logistic regression, support vector machines, sql, python, linear regression, advanced excel, data science]	[post graduate diploma, bachelor technology, b.tech]	0.490	0.707	1.197
3	Candidate 4.pdf	None	[Information Not Found]	2	[ntt ltd, mit, ntt ltd., kochar, ibm]	[iit-hyderabad, central institute plastics engineering technology, indian institute technology]	[programming language, bulandshahar, predictive analytics, machine learning, python]	[bachelor technology, master technology, b.tech]	0.336	0.707	1.043
2	Candidate 1.pdf	None	[Information Not Found]	2	[goldman sachs]	[motilal nehru national institute technology, dr. virendra swarup education centre]	[natural language processing, big data, software development, nlp]	[btech, bachelor technology]	0.000	0.707	0.707
4	Candidate 5.pdf	None	[vkwayne4@gmail.com, mailto:vkwayne4@gmail.com]	1	[]	[kendriya vidyalaya ballygunge, narula institute technology]	[k-means clustering, data visualization, tableau, statistics, probablity, machine learning, logistic regression, data analysis, sql, python, linear regression]	[b.tech]	0.503	0.000	0.503

Figure 23: CandidateEntityList

VIII. CONCLUSIONS

The goal of the study was to tackle 3 basic problems of extracting, accuracy and adding value to the current HR process and has been done successfully. We summarize all the major findings and recommendations.The problem of extracting entities from a document in this case a Resume is a difficult task given the issue of False positives and False Negatives. The current process of using Regex & Rule based approach works well given each and every occurrence of the given entities are captured and represented in Regex or Dictionary. However, as with the current scenario the overall population for Skills, academics, companies etc.keep getting wider that will result in the seen entities being missed while extracting. Named entity recognition is the perfect solution tosolveanyusecases related to information extraction. It is dynamic & not limited in its design. CRF is a widely usedtechnique to locate and classify entities; however it lacks theability to identify future patterns. This problem is eliminated by using Deep learning techniques such as Bi-LSTM's & Spacy NER. Spacy NER in particular performs exceptionally well due to its architecture of hashing tricks or Bloom embeddings and Implementation of 1D CNN layers. For the current scenario, the spacy NER reached an f1-score of 95%.This was achieved through a very small training set and an in-houseannotated text. Oncethe entities are extracted this can be further used to rank the candidates rather than just use them to summarize the resume. Similarity scores such as Cosines similarity works well while comparing to string vectors. Based on the company benchmark weightages are assigned for each entity score to rank and select the best candidate.

IX. FUTURE SCOPE

Training model is very crucial for overall generalization. Theapplication can be extended to train the model over 1000+ unique resumes. Open CV API could be utilized to split the document (pdf, word, image etc.) into different sections.(Summary, Education, Work Experience etc.) and then perform entity recognition operation. Bi-LSTM's CRF with BILOU tags and larger training data has to be explored though Spacy model performs well but has fewer hyper parameters to train.

REFERENCES

[1]Ayishathahira, C.H., Sreejith,C., & Raseek,C. (2018,July). Combination of Neural Networks and Conditional Random Fields for Efficient Resume Parsing.In 2018 International CET Conference on Control, Communication, and Computing (IC4) (pp.388-393).

[2]Dwivedi, R. K., Aggarwal, M., Keshari, S. K., & Kumar,A.(2019). Sentiment analysis and feature extraction using rule-based model (RBM).InInternational Conference on Innovative Computing and Communications (pp.57-63).Springer,Singapore.

[3]Dongyang Wang, JunliSu, Hongbin Yu (2020, Feb).Feature Extraction and Analysis of Natural LanguageProcessing for Deep Learning English Language. In 2020 IEEE Access (Vol.8pp.46335–46345)

[4]JingLi, AixinSun, Jianglei Han,Chenliang Li(2022, Jan). A Survey on Deep Learning for Named Entity Recognition. IEEE Transactionson Knowledge and Data Engineering (Volume:34, Issue:1,01 January2022,pp.50-70)

[5]Darshita Kumar; Shambhavi Pandey; Pooja Patel; Kshitija Choudhari; Aparna Hajare; Shubham Jante(2021,Aug). Generalized Named Entity Recognition Framework. In2021 Asian Conference on Innovation in Technology (ASIANCON) IEEE.

[6]Amruta Mankawade, Vithika Pungliya,Roshita Bhonsle, Samruddhi Pate, Atharva Purohit, Ankur Raut(2023,April).Resume Analysis and Job Recommendation. In2023 IEEE 8th International Conferencef or Convergence in Technology(I2CT)

[7]B.VeeraSekharReddy, Koppula Srinivas Rao,Neeraja Koppula(2022,Dec).Named Entity Recognition using CRF with Active Learning Algorithm in English Texts.In 2022 6th International Conference on Electronics, Communication and Aerospace Technology.

[8] S Bharadwaj,Rudra Varun, PotukuchiSreeram Aditya, Macherla Nikhil, G. Charles Babu (2022, Jul).Resume Screening using NLP and LSTM. In 2022 International Conference on Inventive Computation Technologies (ICICT).

[9]Corbin Petersheim, Joanna Lahey, Josh Cherian, Angel Pina, Gerianne Alexander, Tracy Hammond(Volume: 66, Issue: 2, April 2023). In IEEE Transactions on Education (pp.130-138).

[10]Dipendra Pant, Dhiraj Pokhrel, Prakash Poudyal(2022,Mar)Automatic Software Engineering position Resume Screening using Natural Language Processing, Word Matching, Character Positioning, and Regex. In 2022 5th International Conference on Advanced Systems and Emergent Technologies (IC_ASET).

[11]Muskan Sharma, Gargi Choudhary, Seba Susan (2023, Jan) Resume Classification using Elite Bag-of-Words Approach.In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)