Pre - Print Version

Intelligent Information Retrieval from Unstructured Data using Natural Language Processing

Muhammad Yusuf Khan[1], Syed Zain Ali[2], Muhammad Hassan Sohail[3], Muhammad Wasim[4],Lubaid Ahmed[5]*

***Abstract*-Companies and recruitment agencies required to go through tons of Curriculum Vitae every day to find suitable candidates, which is inefficient if done manually by a recruiter. In this paper, an automatic system is proposed for the selection of best candidate. This proposed model can take out all the vital information from the unstructured curricula vitae and transform them into the structured format. It will also allow recruiters to filter and search for only relevant data within the structured curricula vitae. This proposed model uses different techniques of data extraction, natural language processing and named entity recognition for converting unstructured information into the structured information.**

keywords: **Information Extraction; Natural Language Processing; Filtering unstructured Curriculum Vitae; Named Entity Recognition.**

INTRODUCTION

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) which focuses on spoken and written words in human language [1]. In recent years different NLP techniques such as rule-based, statistical or Artificial Neural Network (ANN) including deep learning are used for natural language processing. The increase in data generation has changed the way we have been dealing with data, data have been always the most important thing in every field in our lives for making better decisions [2], and increasing the efficiency and effectiveness of data is very popular and the contributions in this field have increased vastly [3]. In this paper, the idea is to make hiring process easier for recruiters. The problem is to cater the amount of unrelated data, the difference in formats and file type in different Curricula Vitae (CVs) to find suitable candidates. This proposed model process bulk amounts of data in the CVs and convert the details into one simple format by using Named Entity Relationship (NER). NER is a technique for classifying words, phrases from human language. NER performs what is known as surface parsing, which helps the machine to answer questions like Who, Where, When [4]. NER systems help machine recognize different entities, events, relationships [5]. The first generation of NER system rely on predefine linguistic rules and patterns [6]. In second generation, introduction of Conditional Random Fields (CRF) and Hidden Markov Models (HMM) machine learning algorithms used to learn patterns and features from labeled data [7]. In third generation, deep learning models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and Transformers used for NER tasks [8]. Lately with new era of transformer, fine-tuned pre-trained language models like Bidirectional Encoder Representations from Transformers (BERT) [9] proposed by Google AI team and Generative Pre-trained Transformer (GPT) has become a talk of the NLP domain.

In this paper, the proposed system of Curricula Vitae (CV) information extraction enables the extraction of relevant information from CVs which have unstructured format [10] and change it to structured information. The method is also known as Information Extraction (IE) [11]. Information or data extraction is to find and organize data from unstructured or semi-organized content [12]. This paper focuses on managing the data stacked from CVs and converting them to a format that can be use when required. To achieve this task, the proposed model starts from scratch, which involves converting documents to JavaScript Object Notation (JSON) [13] format. And then apply filtering which enables the opportunity for the most deserving candidates to get selected. To extract information spaCy library is used. It is an industrial-strength Natural Language Processor [14]. The library spaCy v2 uses the Convolutional Neural Network model, to have better accuracy with fewer resources [14, 15]. After data extraction, filtering of CVs can be achieved by changing the data in the lightweight data format [16] that can be easily filtered as per requirement. For example, if a recruiter wants to see the CVs with some special requirements, he/she won't have to go through each and every CV. He/she will only look through CVs which has required field(s).

BACKGROUND RESEARCH

Parsing human understandable language is related to the field of NLP. NLP is the study of how a computer can understand human language [17]. NLP helps computers to understand

[1-2-3-4-5]Department of Computer Science, Usman Institute of Technology Karachi
Country : Pakistan,
Email: *lahmed@uit.edu

and analyze the meaning from the human language [18]. CV Parsing is a challenge in NLP technique because the documents consist of a mixture of semi-structured and freeform text with variance in the data [19].

In [10], text data mining is discussed, as a process of analyzing text to extract information that is useful for specific purposes. It enables unstructured text data for analysis [18, 20]. Text Mining is the extraction of information from text [21] that tries to find patterns from large datasets. Mining text data introduces an importance of the text analytic field of study and is contributed by international researchers and practitioners [22]. It includes structuring of input using text mining which deals with text data that are inherently unstructured [19]. Text mining is also known as Intelligent Text Analysis [23]. There are a few NER frameworks on the planet. For example, GATE, CRFClassifier, OpenNLP and Stanford NLP [40

It is advent due to advancement in computing power, ANN is likely to be used in NLP and will become more widespread in different fields. In recent years, this can be seen in the shape of ChatGPT [24] which allows human-like conversations using chatbot. Dall.E 2 [25] is an AI system that uses natural language and create an realistic image and Whisper [26] is a pre-trained based on encoder-decoder transformer architecture which can be used to automatically recognized human speech. Information extraction from different sources is the key to natural language processing. In other domain, extracting relevant information from patients' clinical note is discussed a method to complete the questionnaire [27]. Building of gene bank from knowledge graph was proposed that will automatically integrate multi-sourced with different characteristics retired mechanical product information [28]. An innovative approach [29] of Generative Pre-Trained Transformer 4 (GPT-4) [24] is used to automate transformation of a large and diverse dataset of radiology reports while complying standards. [30] outlines CIAD's approach to Named Entity Recognition (NER) handle classification task using two approaches namely Bidirectional Encoder Representations from Transformers (BERT) models and a Graph Convolutional Network (GCN) to connect words with same content to build graph. In [31], comprehensive overview for information extraction is discussed. Information from invoices such as proof of purchase, transaction date, etc. Curricula Vitae (CVs) Parsing represents an intriguing test to current NLP procedures, on the grounds that the CV archives comprise of a blend of semi-organized and free shape content with difference in the information [32].

In [33], information extraction is discussed that the extraction of relevant information from unstructured CVs format and change it into structured. This process of CV information extraction is also discussed in [34]. The parsing of CVs text and convert it into structured is discussed in [10]. Different types of methods used in CV information extraction are also discussed in [10]. Named-entity-based information extraction method identify certain words present in phrases. Rule-based information extraction is based on grammatical rules for identification of information. [33] discussed the learning-based classification algorithms for the extraction of information from a document. In [35], resume overloading problem is discussed. In [36], Part Of Speech (POS) tagging is used for parsing, so if there is a noun tag then that will be a name, relating it to the context it can be identified of what this name can be. For example, [37] checks tag with a noun. With prefix or postfix of the University, therefore University names can be found in a document. [38] uses approach of applying grammar induced extraction patterns on Wikipedia corpus to extract relations between named entities.

In this paper, unstructured curricula vitae are processed and transform into structured data. The proposed model uses data extraction technique, natural language processing and named entity recognition for processing and conversion. This paper proposes an information extraction tool especially designs for recruiters for profiling and better analysis of candidates. Its primary task is automatically extracting structured information of the candidate from different unstructured text format CV (i.e. docx, pdf, odt, etc.) written in plain English.

METHODOLOGY

Companies and hiring agencies receive thousands of CVs from job applicants every day. Extracting information from CVs with high precision is not an easy task [37]. This is because they vary in types of information, their order, writing style, etc. This also means that the CVs received are written in various formats like '.txt', '.pdf', '.doc', '.docx', '.dot', '.rtf' etc [33]. In the manual scan of the CV, a recruiter looks for educational qualifications, work experience, job titles, and other personal information. Extracting this information from CVs is required which is largely based on recruiter requirements and also used for filtering CVs.

First, the proposed system collects CVs of all the candidates and save it in the mail box (i.e. database) of respective recruiter. All the collected CVs are then sent to the deployment server for processing. Deployment server is the main processing unit. It consists of two parts:

handling of unstructured data to be transformed in structured data,
finding relevant information such as name, email address, phone etc. of the candidates. Figure 1 shows the complete architecture of the proposed system.

Figure 1: System flow diagram of the proposed model

The proposed model for processing of data is divided in two steps.
a) Conversion of unstructured data into simple text.
b) Identifying relevant entities.

a) Conversion of unstructured CV data into simple text:
The data from CVs are converted into a simple text format by using Apache Tika [39]. Then the CV is segmented in the different section [40] and related information such as personal details like name, email, phone etc. are handled which are shown in Table 1.

b) Identifying relevant entities:
To identify the relevant entities, spaCy v2 library is used. It works best for extracting information from computer science and software engineering CVs [36, 41]. Figure 2 shows the highlighted text results.

To pursue my career in a reputable company Avanza Solution as Software Engineer Key skills Graduation bachelor's in computer science from University of Karachi of Engineering & Technology Relevant skills Programming Languages C, C++, Java, C#, Objective C Concepts OpenGL, iOS, Android Experience sep 2013 - Jun 2015 as an Android developer at iTap Studios Software Engineer in Avanza Solution for duration July 2013 to June 2015

Figure. 2: Shows highlighted text entities in the sample CV

This highlighted text are the different fields that should be extracted and recognized. Entities are extracted and identified by the proposed trained NER model, where each entity extracted is separated as per requirement shown in Figure 3, each entity is extracted by the proposed model and very few of them are not identified.

Entities
Company name →Avanza Solution, iT3ap Studios
Major → Software Engineer, Computer Science
Institutes→ University of Karachi
Skills→ C, C++, Java, C#, Objective C, OpenGL, iOS, Android developer, C, C#, C++, OpenGL
Duration→ July 2013 to June 2015)

Figure. 3: Entities are shown that are extracted and identified

In the final stage, JSON formatted data [13], [42] from the entities are generated. This will be filtered and stored in the structured database according the fields shown in Table 1.

Table 1: Segment and sub-segments that are extracted from the unstructured CV

Segments	Sub-Segments
Personal Details	Name
	Email
	Phone no
Experience	Job Title
	Company
	Duration
Education	Institute
	Major/degree
Skills	Skills

RESULTS

In this proposed system, dataset shown in Table 2 is used. The training dataset is of 64% of total dataset and for testing purpose we use 36% of total dataset. The proposed model needs to be trained that will find the entities in the documents. After training the achieved the results can be seen in Figure 4, which shows that after 81 iterations the losses are reduced to minimal. The percentage of misses at the beginning of the model training was very high and after completion of training it reduced to 0.15% using training dataset. This value is calculated using the following eq.

Table 2: Shows the distribution of dataset

Data Distribution	No of CVs	Percentage
Training Dataset	56	64%
Testing Dataset	32	36%
Total	88	100%

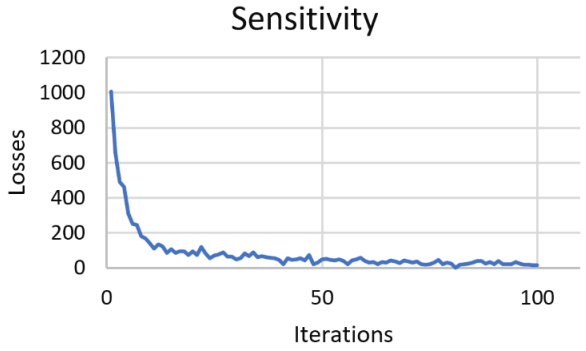

Figure. 4: Shows the misses of the entities while training

Table 3: Shows the sensitivity and precision for all extracted Entities

Entities	Values	Results
Institutes	True Positive = 47 False Positive = 5 False Negative = 7 True Negative = 4839	Sensitivity = 87% Precision = 90%
Major	True Positive = 73 False Positive = 3 False Negative = 8 True Negative = 4814	Sensitivity = 90% Precision = 96%
Company	True Positive = 34 False Positive = 18 False Negative = 11 True Negative = 4835	Sensitivity = 75% Precision = 65%
Experience	True Positive = 16 False Positive = 24 False Negative = 6 True Negative = 4856	Sensitivity = 72% Precision = 40%
Job Title	True Positive = 48 False Positive = 16 False Negative =15 True Negative =4819	Sensitivity = 76% Precision = 75%
Skills	True Positive = 298 False Positive = 54 False Negative =41 True Negative = 4505	Sensitivity = 87% Precision =84%

The precision and the sensitivity is calculated by using confusion matrix [43], [44] and shown in Table 3. The results can be visualized in the Figure 5. In this Figure 5, statistics can be visualized and compared to different entities that are extracted from the CVs using the proposed NER model. The results for entity "Major" is better than the results of entity "Experience". In Figure 6, it can be seen that as the scopes of entities while training and testing were similar, good prediction results can be achieved. This can be seen in Figure 6, where "Major" field is showing good results. When entity "Experience" prediction results are compared, the results are not good. This is due to training issue as in training the model pick dates for both "Experience" and "Education" as dates for one entity. This confuses the learning algorithm to differentiate between the two different entities, resulting in reduced predictions. The results of this experiment are shown in Figure 7.

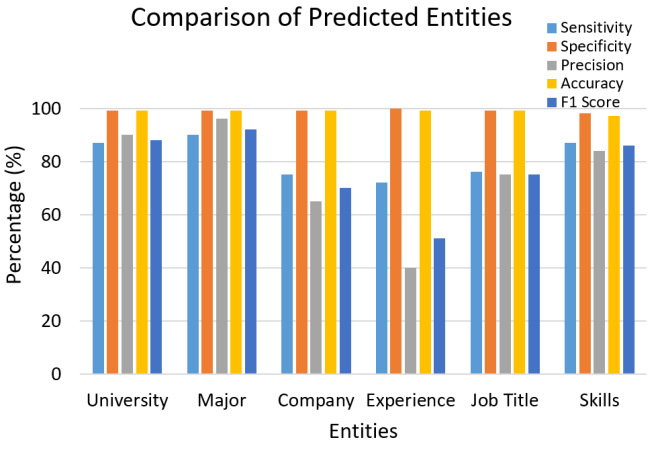

Figure 5: Shows comparison of testing results

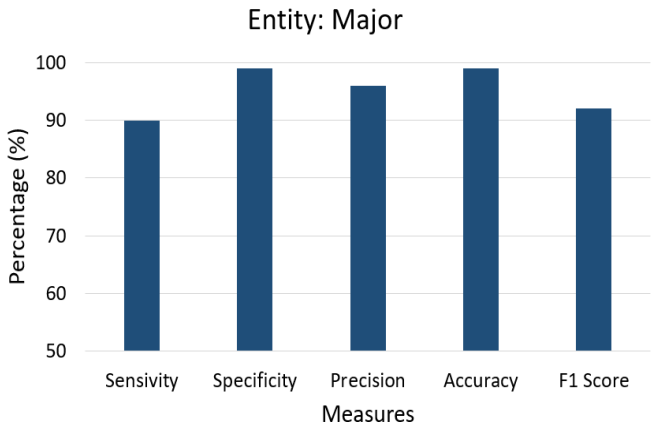

Figure 6: Prediction for "Major" entity

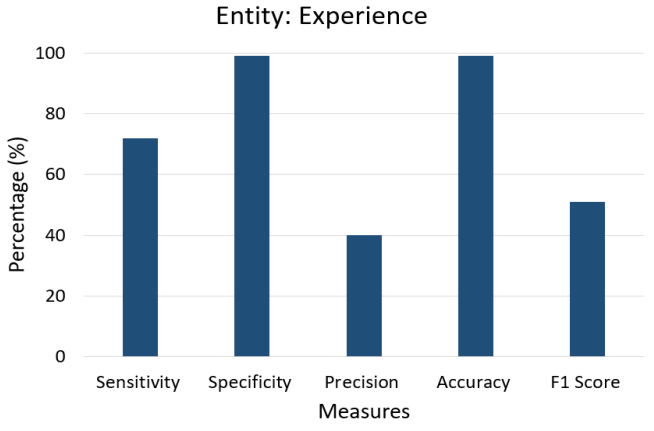

Figure 7: Prediction for "Experience" entity

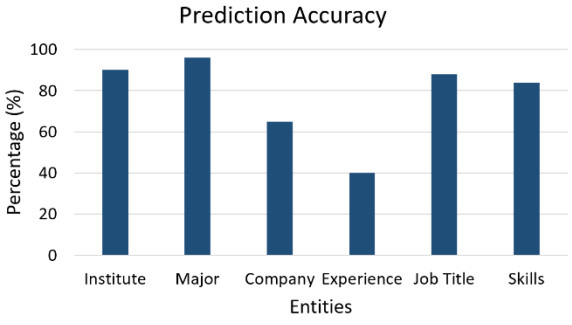

Figure 8: Accuracy of predicted entities

The comparison of the total number of entities in the testing dataset and the total number of correct predictions of each entity field from the testing dataset is shown in the Table 4. Figure 8 shows the graphical representation of Table 4, focusing on the overall accuracy of the extracted field using the proposed model.

Table 4: Shows the accuracy of prediction

Field	Total in Field	Total Found	Accuracy
Institutes	52	47	90%
Major	76	73	96%
Company	52	34	65%
Experience	40	16	40%
Job Title	54	48	88%
Skills	352	298	84%

The results are generated in JSON format and are shown in Figure 9 after extraction of all required entities. All the segments (i.e. Personnel Details, Experience, Education and Skills) and their sub-segment (i.e. Name, Email, Phone no., Job Tile, Company, Domain, Institute, Major/degree, and Skills) are extracted and stored with different key value. Note, if the NER model failed to extract the experience from sub attribute. The result shows a miss match and the not found entities will be replaced with "Null" value.


```
{
    "pk": 2,
    "fields": {
        "CVdata": {
            "skills": [
                "html",
                "css",
                "js",
                "python"
            ],
            "Metadata": {
                "Name": "Muhammad Yusuf khan",
                "email": "myusuf95@gmail.com",
                "phone no": "+923412138798"
            },
            "Experience": [{
                    "Company": "Gadittek",
                    "JobTitle": "Graphic Designer"
                },
                {
                    "Company": "creative chaos",
                    "JobTitle": "Web Designer"
                }
            ],
            "Qualification": [{
                    "Majors": "Computer Science",
                    "Institute": "little folks School"
                },
                {
                    "Majors": "Computer Science",
                    "Institute": "Usman Institute of Technology"
                }
            ]
        }
    }
}
```


Figure 9: Extracted fields in the JSON format

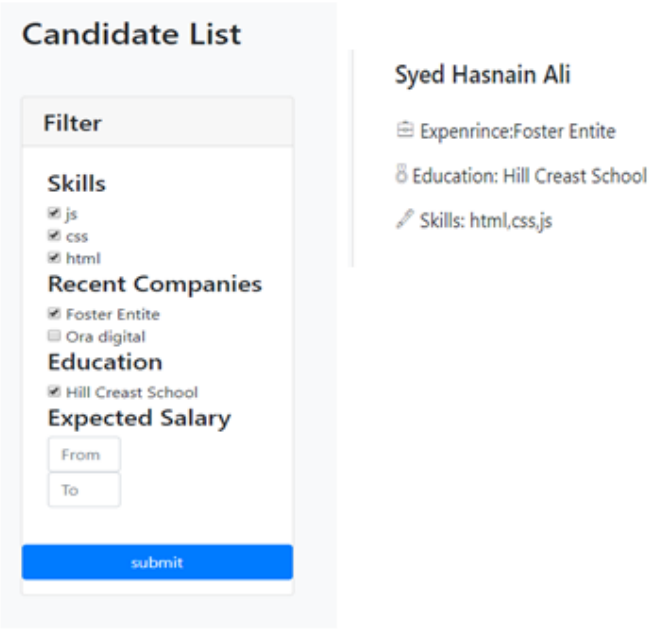

Figure 10: Graphical User Interface for Recruiters

A complete web application is developed for this proposed system. In Figure 10, a Graphical User Interface (GUI) for the recruiter is shown. In this web-application, the recruiter can see the relevant tags that exists in the CVs and used them for filtering the candidate with required entities/fields.

CONCLUSION

While training the proposed model it was observed that there is not much difference in the model error reduction after the 80th iteration as shown in the Figure 4. This process takes a good amount of time, therefore training time can be reduced to half the numbers of iterations for fairly similar results. In future, it can be made a lot better with better training. At this time the process is working fine with good results except for the entity "Experiences" where most of the dates are marked as the experience entity which is not correct. In CVs there are date periods for entity "Education" is also provided so the model also extracts those dates as "Experience". To resolve this issue, the proposed model uses regular expression method.

This proposed system provides an effective and efficient recruiting method. By providing a way of parsing through tons of CVs for finding the required candidates that may have been missed if parsed the CVs manually.

REFERENCES

Sander, P., et al., Machine Learning and Artificial Intelligence in Radiation Oncology. Natural language processing in oncology, ed. T.R. John Kang, Barry S. Rosenstein. 2023: Academic Press. 137-161.

Manogaran, G., C. Thota, and D. Lopez, Human-computer interaction with big data analytics, in Research Anthology on Big Data Analytics, Architectures, and Applications. 2022, IGI global. p. 1578-1596.

Pouyanfar, S., et al., Multimedia big data analytics: A survey. ACM computing surveys (CSUR), 2018. 51(1): p. 1-34.

Ghosh, S., S. Roy, and S.K. Bandyopadhyay, A tutorial review on Text Mining Algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 2012. 1(4): p. 7.

Hirschberg, J. and C.D. Manning, Advances in natural language processing. Science, 2015. 349(6245): p. 261-266.

Collins, M. and Y. Singer. Unsupervised models for named entity classification. in SIGDAT conference on empirical methods in natural language processing and very large corpora. 1999.

Patil, N., A. Patil, and B. Pawar, Named entity recognition using conditional random fields. Procedia Computer Science, 2020. 167: p. 1181-1188.

Chiu, J.P. and E. Nichols, Named entity recognition with bidirectional LSTM-CNNs. Transactions of the association for computational linguistics, 2016. 4: p. 357-370.

Devlin, J., et al., Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Siefkes, C. and P. Siniakov, An overview and classification of adaptive approaches to information extraction. Journal on Data Semantics IV, 2005: p. 172-212.

Huang, C.-C., et al. Cross-lingual information to the rescue in keyword extraction. in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.

Patil, N., et al., Candidate recruitment system by using keyword based searching. International Research Journal of Engineering and Technology, 2017. 4(3): p. 24-26.

Bourhis, P., et al. JSON: data model, query languages and schema specification. in Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems. 2017.

Introducing spacy. 2016 [cited 2024 February 19]; Available from: http://https://spacy.io/.

Andor, D., et al., Globally normalized transition-based neural networks. arXiv preprint arXiv:1603.06042, 2016.

Li, Y., et al., Mison: a fast JSON parser for data analytics. Proceedings of the VLDB Endowment, 2017. 10(10): p. 1118-1129.

Joshi, A.K., Natural language processing. Science, 1991. 253(5025): p. 1242-1249.

Kumar, L. and P.K. Bhatia, Text mining: concepts, process and applications. Journal of Global Research in Computer Science, 2013. 4(3): p. 36-39.

Tan, A.-H. Text mining: The state of the art and the challenges. in Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases. 1999. Citeseer.

Francis, L. and M. Flynn, Text Mining Handbook Casualty Actuarial Society E Forum. 2010, spring.

Gupta, V. and G.S. Lehal, A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence, 2009. 1(1): p. 60-76.

Aggarwal, C.C., Mining text streams, in Mining text data. 2012, Springer. p. 297-321.

Gabriel, R., P. Gluchowski, and A. Pastwa, Data warehouse & data mining. 2009: W3l GmbH.

ChatGPT. November 30, 2022 [cited 2024 February 19]; Available from: http://https://spacy.io/.

DELL.E 2. April 6, 2022 [cited 2024 February 19]; Available from: http://https://spacy.io/.

Whisper. September 21, 2022 [cited 2024 February 19]; Available from: http://https://spacy.io/.

Yashodhya, V.W., et al., A phrase-based questionnaire–answering approach for automatic initial frailty assessment based on clinical notes. Computers in Biology and Medicine, 2024. 170: p. 108043.

Yuyao, G., et al., Integrated modeling for retired mechanical product genes in remanufacturing: A knowledge graph-based approach. Advanced Engineering Informatics, 2024. 59: p. 102254.

Pan, Y.a.F., JinXia and Zhu, ChunTing and Li, Minda and Wu, HuiQun, Towards an Automatic Transformer to Fhir Structured Radiology Report Via Gpt-4. SSRN, 2024.

Armary, P., et al. CIAD System for Geographical Entity Detection at TextMine'24. in TextMine'24. 2024. Dijon, France.

Saout, T., F. Lardeux, and F. Saubion, An Overview of Data Extraction From Invoices. IEEE Access, 2024. 12: p. 19872-19886.

Jiang, J., Information extraction from text. Mining text data, 2012: p. 11-41.

Sanyal, S., et al., Resume parser with natural language processing. International Journal of Engineering Science, 2017. 4484.

Ma, L., Information extraction from unstructured document. 2004: University of New South Wales.

MURZABULATOV, A., The problem of resume overload during talent acquisition. 2015.

Jigyasa Nigam, S.S., Fast and Effective System for Name Entity Recognition on Big Data. International Journal of Computer Sciences and Engineering, 2015. 3(2): p. 31-35.

Khan, I.A., et al., Efficient Techniques for Improved Data Classification and POS Tagging by Monitoring Extraction, Pruning and Updating of Unknown Foreign Words. 2015.

Iftene, A. and A. Balahur-Dobrescu. Named Entity Relation Mining using Wikipedia. in LREC. 2008.

Apache tika. 2007 [cited 2024 February 19]; Available from: https://tika.apache.org/.

Stadermann, J., D. Jager, and U. Zernik, Hierarchical information extraction using document segmentation and optical character recognition correction. 2020, Google Patents.

Zhou, G. and J. Su. Named entity recognition using an HMM-based chunk tagger. in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002.

Huang, D., et al., Means to process hierarchical json data for use in a flat structure data system. 2017, Google Patents.

Han, J., J. Pei, and M. Kamber, Data mining: concepts and techniques. 2011: Elsevier.

Kantardzic, M., Data mining: concepts, models, methods, and algorithms. 2011: John Wiley & Sons.