A Systematic Literature Review (SLR) On The Beginning of Resume Parsing in HR Recruitment Process & SMART Advancements in Chronological Order

Aakankshu Rawat
Amity University
Siddharth Malik
Amity University
Seema Rawat
Amity University
Deepak Kumar (deepakdeo2003@gmail.com)
Amity University
Praveen Kumar
Amity University

Research Article

Keywords: Resume Parsing, Natural Language Processing (NLP), Information Extraction (IE), Named Entity Recognition (NER)

Posted Date: July 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-570370/v1

Abstract

Talent acquisition, also known as recruitment, is definitely amongst one of the most difficult decisions that an organization has to take. The workforce is the most crucial pillar of any organization and surely a deciding factor for its fate. Each organization/company/firm has a dedicated department in place for completely managing an employee's life cycle starting from recruitment till termination, known as the Human Resource (HR) department. National Cash Register Co. was the first company ever to include an HR department in 1900 for resolving conflicts among existing employees. Since then, the world of recruitment has grown rapidly with a lot of advancements to fast-track the hiring process. Realising the importance of hiring fitting and apt employees, HR units around the world have undergone substantial changes from traditional recruitment methodologies to network recruitment and finally to smart hybrid recruitment strategies for efficient hiring with a significantly less human workload. The purpose of this Systematic Literature Review (SLR) article is to shed light on all the advancements in the hiring process from a technical perspective. The article follows the defined format and all appropriate guidelines of an SLR and would provide an extensive study of various Machine Learning (ML) and Deep Learning (DL) approaches to facilitate hiring. The main emphasis has been given to "Resume/CV Parsing" as an enhancer of fast-track hiring. The SLR would be instrumental in providing various fast-track approaches to go with for resume parsing, the scope of improvement and also focus on various challenges/ethical considerations to keep in mind while automating the hiring process.

1. Introduction

The term "Human Resource" was first coined by John R. Commons in 1893 in his book "The Distribution of Wealth". Further expansion of the term with possible explanations was done by Peter F. Drucker in his pioneering work "The Practice of Management" in 1954 where he elucidated the singularity of human as a resource (Vitor M. Marciano, 1995). The initial emphasis of Human Resources Management (HRM) was to uplift the working conditions of employees. The Equal Pay Act and Civil Rights Act of 1963 & 64 respectively made it mandatory for organizations to emphasize compliance affairs as well. Studies in the psychological domain such as the Two-Factor Theory defining reasons for Job Satisfaction and Dissatisfaction proposed by Frederick Herzberg (Ewen et al, 1966) & the Self Determination Theory which explores the various intrinsic and extrinsic roots of motivational impulses (Deci & Ryan, 1980) compelled organizations to finally shift their focus from venturous business strategies to their most important resource, the employees. Things started to improve significantly and from 1990 onwards, recruitment of top-grade performers became pivotal with a shift from process-centric to a worker-centric mentality in most of the organizations. This shift along with the advent of the Internet (1989) led to cut-throat competition among prospective jobseekers as now HR managers were receiving oodles of applications from all over the world through e-mails. Organizations all around the world started dedicated career pages for displaying job openings. This led to the creation of Job Boards such as Monster Board (1994) and Netstart Inc (1995), now known as CareerBuilder followed by many others like Craigslist, CareerPath, and AOL. Job Boards were used by employers to post job openings in search of suitable candidates. An

initial form of Applicant Tracking System (ATS) came into scope in 1996 when Marin Ouelelett created the Via site. It was renamed Recruitsoft Inc and a proper Recruitment Management System (RMS) was launched by the name of Recruiter WebTop in 1999. Meanwhile, Monster Board did a merger with Online Career Center (OCC) to launch Monster.com (1999). Subsequently, many ATSs' came into the market with extended features of Boolean Searching and Resume Parsing. With so many technical advancements happening in the sphere of recruitment, it was ascertained that Resume Parsing was the most critical component to hire a suitable candidate. Resumes are the first source of impression for an employer to gauge any jobseeker and all HR units were comprehensively doubtful of the efficiency of ATSs' that did not include resume parsing. This led to extensive research in the field of Resume Parsing.

Resumes in general are unstructured in nature. Parsing involves converting resumes into a structured machine-readable format like JSON, CSV and XML. While parsing resumes, the text is analyzed and information is extracted about the different sections of a resume like personal details, education, work experience, skills, certifications, additional information etc. This segmentation of text into relevant sections can be achieved with the help of Machine Learning (ML). Natural Language Processing (NLP) is a subdomain of ML which provides various methodologies to work with natural human language. Since human language is constantly evolving and highly ambiguous in nature due to change in writing style, culture, context, demographic etc., it becomes exceedingly arduous to work with it. NLP can be considered as a sub-set of Computational Linguistics (CL) which focuses on utilizing human language to generate useful output in the form of algorithms or tools/applications (Tsujii, 2011). The SLR conducted in this article deals with numerous NLP approaches to solve the hurdle of parsing/screening resumes efficiently.

Table 1. Resume Parsers available in the Market

Product	Parent Company	Product Launch Date	Company Founded In	Headquarters
Sovren Parser	Sovren	1996	1996	USA
Lens Xray	Burningglass technologies	2001	1999	USA
Rchilli Resume Parser	Rchilli	2010	2010	USA
affinda Résumé Parser	Affinda	2013	2013	Australia
hiretual	Hiretual	2015	2015	USA
Textkernel Extract	Textkernel	2018	2001	The Netherlands
employa	employa	2019	2019	USA
DaXtra Parser	daXtra	NA	2002	UK
HireAbility Alex	HireAbility	NA	1999	USA
Sniper Ai	recruitment smart	NA	2015	UK
CVViZ Resume Screening	CVViZ	NA	2016	India

On examining various companies that provide resume parsing solutions, it is found that there are very few companies that provide parsing as a stand-alone service to their clients. Fig 1.1 lists a few of them. Most of the solution providers have included the functionality of parsing as a part of their Applicant Tracking System (ATS), Recruiting Software or Human Resource Management System (HRMS). The relation among these services is elucidated in Fig 1.2. Sovren was the first company to come up with their

parser in 1996. Since then, many other companies have worked on this domain and come up with parsing solutions in one form or the other. The latest parsing solution is by a company based in the USA by the name of “employa” (2019) which is a smart ATS parsing extension with extended functionalities of even detecting fraud and inconsistent resumes/profiles. Fig 1.3 & 1.4 demonstrate the geographical distribution & set up of implicit Parsing solutions for a better understanding of the niche development. The launch year of the services has been extracted either directly from the company’s website or articles, news, blogs, and videos about the product. “NA” has been used in case no information is available about the product launch date.

The rest of the article is divided into the following sections: Section 2 discusses the Systematic Literature Review (SLR) approach adopted along with the research questions, search criteria, quality assessment through inclusion and exclusion criteria and finally the data collected for review. Section 3 discusses the initial approaches to resume parsing along with certain major techniques and additions to basic parsing to enhance the results. Section 4 describes the most unique and effective approaches to resume parsing. Section 5 has been included to acquaint the readers with the challenges and ethical deliberations to be considered while automating recruitment entirely and finally the article is concluded in Section 6.

2. Methodology Adopted For Slr

The concept of SLR was first introduced in the field of medicine and is now being implemented in almost all domains to gather sufficient evidence for a subject matter. SLRs follow a fixed protocol that commences with defining Research Questions (RQ) to be investigated thoroughly over a multitude of databases/repositories while keeping in mind the search criteria (Kitchenham, 2004). It is then followed by filtering out the data collected in the form of research articles, conference papers, books etc using strict inclusion and exclusion criteria for ensuring the quality of the evidence collected. The same methodology is applied in this article considering that it has been applied by many in the technical field (Kitchenham et al, 2009).

2.1 Research Questions (RQ) to be investigated

Following are the research questions that are investigated and answered in this review article:

- RQ1: What are the major advancements in the field of HR recruitment?
- RQ2: What are the various parsing techniques developed till now?
- RQ3: How implementing ML frameworks for resume parsing/screening be an effective way to significantly reduce the hiring time?
- RQ4: How DL frameworks improve the efficiency of parsing?

2.2 Search Guidelines

The collection of data for conducting an SLR is always aided by several search guidelines/ criteria. A set of keywords are used for traversing the data repositories and it is ensured that those keywords appear in

the Title or Abstract of every search result. Keywords such as “Resume”, “Natural_Language_Processing”, “Ontology”, “Information_Retrieval” or combination of Keywords like “Machine_Learning + Computational_Linguistics”, “Deep_Learning + Named_Entity_Recognition” were used to retrieve unbiased data from various Data Repositories. Similar words such as “Artificial_Intelligence”, “Information_Extraction”, “Abstractive_Text_Summarization” were also used to gather as much data as possible. Fig 2.1 illustrates the various keywords used with the help of a Word Cloud. The bigger text represents more frequency of occurrence of that particular keyword in the search results (Filatova, 2016).

2.3 Data Repositories referred to

The selection of databases/repositories is a crucial aspect of conducting an SLR. It is ensured that only standard databases are used to apply the above-defined search criteria. The search results consist of journal articles, pre-print, conference papers and books. No other information source is taken into consideration to ensure that the data retrieved is reliable. Following are the data repositories that are referred to:

- Springer (https://www.springer.com/in)
- arXiv.org e-Print archive (https://arxiv.org/)
- Elsevier (https://www.elsevier.com/en-in)
- Google Scholar (https://scholar.google.com/)
- ACM Digital Library (https://dl.acm.org/)
- MIT Press (https://direct.mit.edu/journals)
- IEEE Xplore (https://ieeexplore.ieee.org/Xplore)
- Taylor & Francis Online (https://www.tandfonline.com/)
- Google Books (https://books.google.co.in/)
- CiteSeer (Now CiteSeer X) (https://citeseerx.ist.psu.edu/index)
- Semantic Scholar (https://www.semanticscholar.org/)
- Academia.edu (https://www.academia.edu/)
- ACL Member Portal (https://www.aclweb.org/portal/)
- ResearchGate (https://www.researchgate.net/)

2.4 Final Selection of Data

Once the initial set of articles are retrieved, certain inclusion and exclusion criteria need to be applied as part of quality assessment so that the data collected is not skewed by several biases. Hence, after data collection the following criteria were set for further refining the set:

2.4.1 Quality Assessment

- ***Inclusion Criteria***

Only articles, papers & books written in the English language and about the search criteria defined were collected. Manual reading was conducted to retain articles that were expanding knowledge on various technical methodologies and frameworks required for resume parsing or improving the same. Also, articles dealing with resumes in languages other than English were considered for review if the proposed methodology did not get affected by the language itself (in case if an exclusive solution about a non-English language is proposed, then such articles were not considered).

- *Exclusion Criteria*

Data about languages other than English were removed. Even though the search criteria were applied, many papers were retrieved that were irrelevant to the subject matter. Hence, manual reading of the abstract and results was conducted to remove articles that did not focus on resume parsing and further implementation of improvements.

2.4.2 Data Collected

After following all the protocols, 317 articles were collected in total. An in-depth visualization has been done for the data collected in order to cover all of its dimensions.

Fig 2.2 represents the number of articles published in 5-year class intervals ranging from 1975-2025. As it is observed, a maximum number of researches were conducted between 2015-2020. Fig2.3 reflects upon the distribution of articles in repositories that were finalised. Fig 2.4 extends Fig 2.3 to demonstrate the count and volume of distribution of articles published in different journals that were finalised for review in context with the “year” class intervals. The maximum number of articles were published in 2017.

3. Coalescence Of Review Findings

Section 3 responds to three of the research questions (RQ1, RQ2, RQ3) defined in Section 2. The section is divided into three segments. The first segment discusses various approaches chronologically used for resume parsing and answers the 3RQs. Considering the inclusion criteria set in Section 2, the second segment is dedicated to algorithms and frameworks to improve the backend parsing process. Similarly, the third segment throws light on some auxiliary techniques to upgrade parsing.

3.1 Approaches to Resume Parsing in chronological order

Information Extraction (IE) from resumes with high recall and precision is a tough task owing to their heterogeneity due to factors like different formats, file extensions and writing style as well. One suitable method to extract information from resumes was proposed by (Yu, K., Guan, G. & Zhou, M., 2005) which constitutes a two-pass Hybrid Cascaded Model (HCM). In the first pass, Hidden Markov Model (HMM) was applied to extract general information by segmenting the resume into several blocks followed by annotating each block with a label.

Since HMM is a state-based model, extraction of information fields holding a strong order of sequence becomes much easier. In the second pass, detailed IE is carried out using Support Vector Machine (SVM) classifier within the boundary of each block as classification of information becomes viable if it is independent in nature as in the case with the segmented blocks except for the case with “Education” block. Since most resumes have the same sequence in their Education Section, HMM performs better in this case. SVM is used as the classifier due to its sturdiness to overfitting and high performance (Sebastini, 2002). Good Turing smoothing is applied in HMM for parameter estimation & a back-off schema given by Katz in 1987 is used for probability estimation as sparse training data could be a big hurdle (Gale, 1995; Katz, 1987). “One vs All” multi-class classification strategy is applied in the SVM model as it is a binary classification model by nature. Block Selection for the second pass is done using a fuzzy selection strategy to avoid non-boundary blocks being labelled as boundaries and to enlarge the search scope. Since HCM is a pipeline framework, the chances of error propagation from one pass to the next are very high. (Kopparapu, S. K, 2010) proposed a one pass alternative in which six major segments of a resume are extracted as described by the HR-XML Consortium. IE is done using the N-grams NLP model & a combination of heuristic rules are applied for extracting other segments. One of the major issues faced while doing IE using a predefined knowledge base is the creation of a knowledge base since it is time-consuming if done manually. A modification of the BASILISK algorithm (Thelen et al, 2002) proposed by (Pawar et al, 2012) which is unsupervised in nature works perfectly for domain-independent Named Entity Extraction (NEX) for automated gazette (knowledge base) creation. One such modification includes the addition of negative features to bring more clarity. As an example, depending upon the domain, “Role” and “Designation” might have the same meaning, or maybe not. This introduces ambiguity to the algorithm. A custom algorithm is also introduced for NER, where n-level indexes are prepared for the most important words of each named entity to leverage importance. These indexes are further used to map the occurrences with the gazette. This approach is much better and accurate than naïve regular expressions (regex) based methods such as (Kopparapu, S. K, 2010).

“Semantic Web” is a term coined by Tim Berners Lee and is an extension of WWW. It is a set of standards that lets internet data become machine-readable. Ontology is one of the pillars of Semantic Web standards and some of the ontology languages are Web Ontology Language (OWL) & Resource Description Framework (RDF). (Celik et al, 2012) incorporates Ontology Knowledge Bases (OKBs) to store various domain ontologies for each type of resume segment. These ontologies contain <Literal xml: lang> tag to deal with resumes in multiple languages. EXPERT uses ontology mapping to map resumes with job criteria using a custom mapping & similarity function (Kumaran et al, 2013). Another approach performing ontology mapping stores the resume in RDF graph database using a NoSQL data model & uses SPARQL for querying the database (Abirami et al, 2014; Bojārs et al, 2007). For implementing an ontology-based parser in J2EE, OWL API by Apache Jena can be utilized (Mohamed et al, 2018). Although still valid, the concept of the Semantic Web is deprecated since the rise of Artificial Intelligence (AI).

Initially and even now, there are many recruiting websites and software that provide an online form to fill out basic details like personal information, education details, certifications, work experience etc. This meta-data can be automatically analysed, but most of the candidates skip this part and directly upload

their resumes making screening of resumes mandatory. PROSPECT proposes a well-defined architecture for screening resumes (Singh, A., Rose, C., Visweswariah, K., Chenthamarakshan, V., & Kambhatla, N., 2010). The good aspects of this system include the use of shingling (Broder, A. Z.,2000) for detecting identical/plagiarised resumes, linear-chain Conditional Random Fields (CRFs) for IE & mining of named entities, TabClass & ColClass SVM classifiers for segregating multiple tables in a resume and further identification of columns in the table, respectively. CRFs perform better than HMMs in a linear chain model as they act as conditionally trained HMMs and avoid the label bias problem (Lafferty et al, 2001). But ColClass classifier is not applicable for 1-D tables having just rows. Also, data normalization here is dealt with using string-matching techniques which is quite rudimentary. The initial data annotation of the training set is conducted manually, and no automated alternative is defined. Segmentation and attribute extraction is done using multiple feature types like lexicon features, visual features etc. and still, IE is not satisfactory. Only tabular data is properly extracted. Scoring of resumes is done using Term Frequency – Inverse Document Frequency (TF-IDF) technique. Another framework proposed by (Farkas et al, 2014) also uses multiple feature types but the improvement here is the use of both manual and automatic annotation using a self-developed custom tool. A two-level annotation scheme is considered for IE due to the presence of complex data structures. Maximum-Entropy Markov Model (MEMM) & CRFs are applied for IE & MEMM is finally employed for use due to substantially less runtime. The process of data annotation can be added using the Dataturks tool & the spaCy library of python can be utilized to perform various NLP tasks such as tokenization, Part-Of-Speech tagging (POS), Named Entity Recognition (NER) etc. (Satheesh et al, 2020). A two-step extraction framework with adaptive segmentation as an intermediary using a basic classifier to obtain semi-structured data with Simple, KeyValue and Complex tags and further IE using Naïve Bayes algorithm works better than PROSPECT (Chen, J. et al, 2015). A standardized text-windows based approach is adopted by (Tikhonova et al, 2019) in which word-embeddings along with their TF-IDF weights are summed to generate a text field embedding which could then act as input to CatBoost classifiers for segmentation of resumes. Word-Embeddings could be constructed using algorithms such as FastText (Joulin et al, 2016; Bojanowski et al, 2017), Word2Vec (Mikolov et al, 2013; Mikolov, T., Yih, W. T., & Zweig, G., 2013) & GloVe (Pennington et al, 2014).

Another approach deals with the concept of specialness/uniqueness to extract special skills from resumes (Maheshwari et al, 2010). This is based on the concept of product selection from e-commerce websites using special features (Maheshwari et al, 2009). It is assumed that a resume follows a two-layer structure when it comes to the skill segment i.e., skill type & value. Initial preprocessing is done on the text documents (resumes) using several rules implemented either manually or by using a combination of lists and hash tables. Then features sets are identified for the skill type and value using a custom algorithm with each feature represented by a tuple. A “Degree of Specialness (DS)” criteria is defined to score each feature from 0 to 1. Features with DS=0 & DS close to 1 are categorized as common and special features respectively. The rest of the tuples are categorized as common cluster features. Finally, a 3-level feature organization is done using a clustering algorithm with the III-level having the special features. A similar approach also extracts skills from the resume, not special features, but uses a skill ontology with more than 13,000 concepts to match skills (Chifu et al, 2017). The key takeaways from this article are the two

algorithms proposed: one for generating lexicalized contexts preceding a known skill using manually identified POS patterns to identify new skills unknown to the ontology, the other for suggesting taxonomic parents for the new skills detected using Wikipedia to store them in the ontology. Similar work is done by (Chandola et al, 2015) where a weighted knowledge base of skills (nouns, verbs & adjectives) is used to compare with the words extracted from the resumes after applying a POS Tagger & a chunker on top of it. The resumes are scored by adding up the weights of the matching skills. Finally, a categorization similar to sentiment analysis is performed based on the score to segment the resumes on a priority basis.

Table 2. Different approaches to Resume Parsing in chronological order

Author/s	Year	Techniques Used
Yu, K., Guan, G. & Zhou, M.	2005	Cascaded Hybrid Model using HMM & SVM
Singh, A., Rose, C., Visweswariah, K., Chenthamarakshan, V., & Kambhatla, N.	2010	CNNs for IE, TabClass & ColClass SVM classifiers for tabular data extraction, TF-IDF for scoring
Maheshwari, S., Sainani, A., & Reddy, P. K.	2010	Custom Framework proposed to extract special skills from a text-resume
Kopparapu, S. K.	2010	N-gram NLP model for skills extraction & a combination of heuristic rules to extract other details as specified by HR-XML Consortium
Pawar, S., Srivastava, R., & Palshikar, G. K.	2012	Proposed a modified BASILISK algorithm for NEX & A QUICK SEARCH custom algorithm for NER
Çelik, D., & Elçi, A.	2012	Semantic Web Ontology approach for custom parsing framework
Senthil Kumaran, V., & Sankar, A.	2013	Ontology mapping with a custom statistical mapping & similarity function
Farkas, R., Dobó, A., Kurai, Z., Miklós, I., Nagy, Á., Vincze, V., & Zsibrita, J.	2014	Resume conversion using Poppler PDF library, IE using MEMM & CRF
Abirami, A. M., Askarunisa, A., Sangeetha, R., Padmavathi, C., & Priya, M.	2014	Ontology mapping taking aid of TF-IDF & Cosine similarity
Chen, J., Niu, Z., & Fu, H.	2015	Tika for resume conversion, Naïve Bayes Classifier for IE
Chandola, D., Garg, A., Maurya, A., & Kushwaha, A.	2015	Chunker on top of POS Tagger for text mining, custom algorithm for skill mapping & resume categorization
Chifu, E. S., Chifu, V. R., Popa, I., & Salomie, I.	2017	NLP Software by Standford NLP Group for text-preprocessing and extracting nouns from phrases & two custom algorithms: one to generate lexicalized contexts preceding a known skill & other to suggest taxonomic parents of a new found skill
Ayishathahira, C. H., Sreejith, C., & Raseek, C.	2018	CNN for segmentation with GloVe for word embeddings, CRFs for sequence labelling
Pham Van, L., Vu Ngoc, S., & Nguyen Van, V.	2018	Rule Based Chunkers for NER, DNN consisting of CNN-Bi-LSTM-CRF to improve NER & SGD for Optimization
Maheshwary, S., & Misra, H.	2018	Siamese network of Twin CNNs for Job Matching, Doc2Vec for generating word embeddings
Tikhonova, M., & Gavrishchuk, A.	2019	CatBoost classifiers for classification of segments aided by TF-IDF
Satheesh, K., Jahnavi, A., Iswarya, L., Ayesha, K., Bhanusekhar, G., & Hanisha, K.	2020	SpaCy for NER, and Dataturks for aiding data annotation

Deep Learning (DL) is a sub-domain of ML which uses Artificial Neural Networks (ANNs) consisting of nodes mimicking the biological neurons. The depth of the ANNs enables them to differentiate between different segments/classes if given sufficient data. Also, since they store the input data in the nodes themselves, a loss of data from the source repository does not tamper with its performance. (Pham Van et al, 2018) applies a Deep Neural Network (DNN) for IE from resumes. In this framework, initial segmentation is performed using a data dictionary of common resume headings to match with the resumes. Dedicated rule-based chunkers for each segment are applied for NER. In order to find more named entities, a DNN comprised of CNN-Bi-LSTM-CRF layers from bottom to top is used. Convolutional Neural Network (CNN) is for word embeddings generation. CNN is efficient in encoding morphological information extracted from characters into neural representations (Dos Santos et al, 2014; Chiu et al, 2016). Recurrent Neural Networks (RNNs) can capture time-dynamics via cycles in the graph. Hence, they should be able to capture far-away dependencies, but they fail due to problems about gradient vanishing/exploding (Pascanu, R., Mikolov, T., & Bengio, Y., 2012). The Long Short-Term Memory (LSTM) model is an upgrade that can manage these issues. Hence, a Bi-directional LSTM (Bi-LSTM) model is used as it is recommended to have ingress for both past and future contexts to aid sequence labelling. Finally, the output vectors of Bi-LSTM layers are passed onto a CRF layer to generate the best possible sequence of labels due to its nature of learning the correlations between the outputs. A Stochastic Gradient Descent (SGD) is used for parameter optimization as it performs better than other optimization algorithms like RMSProp, AdaDelta and Adam. Early stopping can be implemented based on the performance of validation sets at each epoch (Caruana et al, 2001). Dropouts can be applied on both input & output vectors of the Bi-LSTM layer to mitigate overfitting (Srivastava et al, 2014). Another approach for IE using ANNs is done using the following techniques: a CNN model for segmentation using GloVe for word embeddings and CRFs for sequence labelling using CRF++. Bi-LSTM & Bi-LSTM-CNN models are used to compare with the above techniques for segmentation and sequence labelling, respectively. CNN performs better because of a pre-trained GloVe model and since CRFs are undirected in nature, they perform much better than the Bi-LSTM-CNN model because of their ability to access both past and future contexts (Ayishathahira et al, 2018). Job Matching using job descriptions (JDs) and resumes can be achieved using just CNNs through a deep Siamese network (Maheshwary et al, 2018). This approach uses a pair of CNNs with max pooling, repeating convolutions and leaky rectified linear unit layers covered by a fully connected layer on the top of the network. It helps to accurately obtain the underlying semantics by pushing away dissimilar resumes and JDs and projecting similar ones closer in the semantic space. LSTMs can also perform the same task but with much more computational cost. Parameter sharing in the siamese network reduces computational time. Doc2Vec model is used to generate word embeddings as the input for the network.

3.2 Algorithms & techniques to improve the backend process of parsing

It is now evident that segmenting and labelling sequence data is essential for parsing resumes. HMMs & stochastic grammars used for segmentation & labelling make strong independence assumptions owing to their generative nature. They define a joint probability over label sequences and observations which makes the models unmanageable due to the impracticality of representing long-range dependencies of

the observations or multiple interacting features. Maximum Entropy Markov models (MEMMs) are better than generative models but they along with other finite-state models suffer from label bias problem causing bias towards states with few successor states. The CRF framework works better than HMMs & MEMMs since it has a single exponential model for joint probability over label sequences, provided the observation sequence (Lafferty et al, 2001). Word vector representations were initially achieved using models like Latent Dirichlet Allocation (LDA) & Latent Semantic Analysis (LSA) further replaced by Feedforward Neural Net Language Model (NNLM) & Recurrent NNLM (RNNLM). (Mikolov et al, 2013) proposed two models: Continuous Bag-of-Words (CBOW) & Continuous Skip-Gram model (Skip-gram) having much more semantic and syntactic accuracy than NNLM and RNNLM models. CBOW predicts the current word based on context while Skip-gram predicts the neighbouring words based on the current word.

ML algorithms require some sort of text representation as input since text cannot be directly fed to the algorithms. One such method of text representation is the fixed-length Bag-of-words (BoW) model. But BoW loses a lot of information like word order and sentence grammar. (Mikolov et al,2014) proposed an unsupervised algorithm, Paragraph Vector, that is trained using stochastic gradient descent and backpropagation. It utilizes text of variable length to learn fixed-length feature representations. Paragraph Vector outperforms the bag-of-word model by about 30% on a text classification task.

A huge amount of feature engineering and lexical information is required for efficacious NER. Taking inspiration from (Collobert et al, 2011), a novel neural network-based architecture is proposed to minimize the need for the same (Chiu et al,2016). The model incorporates a character and word-level features-based hybrid Bi-LSTM and CNN model to attain state-of-the-art performance for NER, minimizing the need for feature engineering. Similarly, a DNN architecture named CharWNN employs character and word-level representations to perform POS Tagging (Dos Santos et al,2014). State-of-the-art results are achieved using a convolutional layer to extract character level features without using handcrafted features.

Table 3. Advanced algorithms and frameworks used in Resume Parsing

Author/s	Year	Work done	Contribution
Lafferty, J., McCallum, A., & Pereira, F. C.	2001	Proposed CRF framework	Better than hidden and maximum entropy markov models in segmenting and labelling sequence data
Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P.	2011	Proposed a unified neural network architecture: SENNA in C language that can be applied to almost all nlp tasks	Performs much better than various benchmark frameworks and algorithms for POS Tagging, Chunking, NER etc.
Mikolov, T., Chen, K., Corrado, G., & Dean, J.	2013	Proposed two architectures for computation of word vector representation	Word similarity is calculated at a much lower computational cost
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J.	2013	Proposed extensions of Skip-gram model to work with idiomatic phrases	Subsampling frequent words results in fast training of word vector representations
Pennington, J., Socher, R., & Manning, C. D.	2014	Proposed GloVe model	Much better at calculating word similarities but not effective for documents
Le, Q., & Mikolov, T.	2014	Proposed Paragraph Vector model	Better than bag-of-words-model
Dos Santos, C., & Zadrozny, B.	2014	Proposed CharWNN, a DNN for POSTagging	POSTaggers produced for english language had 97.32% accuracy
Chiu, J. P., & Nichols, E.	2016	Proposed a novel neural network architecture for automatic detect of word and character level features, novel method to encode partial lexicon matches	A hybrid Bi-directional LSTM AND CNN helps in eliminating most of the feature engineering

3.3 Auxiliary techniques to aid parsing

Table 4. Auxiliary techniques to aid parsing

Author/s	Year	Techniques Used
Jacob, F., Javed, F., Zhao, M., & Mcnair, M.	2014	Indexing using Lucene, J48 classifier for filtering & Ignore case equals + Lucene Levensthein for string comparision
Sayfullina, L., Malmi, E., Liao, Y., & Jung, A.	2017	Domain adaptation using Word2Vec model followed by CNN with a soft-max layer
Zhang, C., & Wang, H.	2018	Naïve bayes for realizing career patterns, Apriori for explicit relation mining & cosine similarity for implicit relation mining
Mittal, V., Mehta, P., Relan, D., & Gabrani, G.	2020	Stanford CoreNLP, regex for NER, TF-IDF followed by Logistic Regression for job domain prediction

This segment presents some techniques that could facilitate the parsing of resumes in a better way with a bit of effort to obtain better results. Normalizing academic organization names would improve the parsing process by efficient clustering of candidates from the same academic background and a better understanding of the market dynamics in terms of the pay scale of candidates in the same college/university. sCool is a framework designed for CareerBuilder (CB) to achieve the same (Jacob et al, 2014). It is achieved using two major steps: firstly, the database is initialized for normalization by creating a mapping from names using MediaWiki API and the existing CB database. Then the mappings are merged, and duplicate & invalid mappings based on similarity measure are removed followed by indexing of the valid mappings using Lucene (McCandless et al, 2010). The second step is to perform normalization on the institute names. This step involves the removal of unwanted or invalid names using the J48 Classifier developed using Weka (Hall et al, 2009). Normalized institutes can be filtered out using sCooL's search query efficiently as it allows the user to select from a range of string-comparison algorithms such as N-gram, Jaro Wrinker, Levenshtein etc. plus a combination of algorithms. A combination of ignoring case equals and Lucene Levensthein works best with their system. A concept similar to job recommendation which is used to suggest suitable jobs to job-seeking candidates can be used by the recruiters as well to classify resumes in their database into different domains or job categories so that candidates irrelevant to their field can be removed. This helps in efficient database capacity utilization. (Sayfullina et al, 2017) uses job descriptions for training the dataset and tests it on a set of resume summaries using a fastText classifier and a custom CNN model. fastText is used due to its outstanding performance without utilizing GPU and reportedly better performance than DNNs like CNN and char-CNN. The custom CNN used receives input in form of a matrix formed using concatenating word vector representations (using word2vec model) by rows. In order to capture the most important feature max-pooling is applied after the convolution. Finally, a soft-max layer is used to obtain a probability distribution over classes. The custom CNN model outperforms the fastText classifier in predicting the job domains. An alternative method performs parsing on resumes using Stanford CoreNLP,

regex, and pattern-matching operations (Mittal et al, 2020). TF-IDF is used to extract features from the parsed information and logistic regression is applied to assign a job domain to the resumes. The classifier is trained using a manually curated training set of skills along with their job domains. There is hardly any research in the field of resume visualization. Extracted information from resumes if visualized, can make the job of recruiters much more diverting. An attempt at visualizing resumes on government officials is made by ResumeVis (Zhang et al, 2018). This visualization tool displays three major graphs: a statistical histogram for comparing career trajectories using naïve Bayes classifier; an ego-network based spiral graph for displaying interpersonal relationship among candidates using Apriori algorithm for mining frequent resume sets from a basket dataset of resumes and organizations, a custom matching algorithm to measure similarity among the resumes in the set and cosine similarity to compute implicit relationships; & an organizational individual mobility map among various sectors facilitating the hiring decision.

4. Distinct And Efficacious Approaches For Resume Parsing

This section introduces several unique parsing methods cumulated after reviewing articles published from 2015 and onwards. Table 5 lists all the unique articles which are elucidated in the section. A completely ML-based data-driven solution is proposed by (Lin et al, 2016) using unsupervised feature extraction and a custom bagging ensemble method. Manual features are listed in a dictionary and retrieved using numerical keys while training; semantic features are extracted using Word2Vec model; and similar semantic features which are otherwise left out, are here extracted using K-means algorithm and LDA for text and document features, respectively. Finally, a custom bagging method comprising of Random Forest (RF) & XGBoost (XBG) shallow estimators and LSTM & CNN deep estimators named IBagging is used to testing the model with resumes. Another approach takes full advantage of the hierarchal structure of PDF documents for IE (Chen et al, 2016). Preprocessing of documents is done using heuristic rules based on character position in the document. Higher-level blocks are segmented using a heuristic rule-based recursive bottom-up algorithm and further classified using LIBSVM, which is an optimized application of the SVM classifier (Chang et al, 2011). SVM extends to multi-class classification using the “one versus one” strategy. Both content-based and layout-based features are extracted using CRFsuite (Okazaki, 2007). Rule-based post-processing is performed to improve the results of specific entities. Considering the hierarchical structure improves the performance by more than 20% in terms of F1-score. Similarly, a statistical generative model could also be used to extract the structure of a resume. (Ravindranath et al, 2019) propose a Gibbs sampling algorithm-based model to extract structure as well as meaning from resumes based on seven key assumptions. Sampling from a joint posterior probability distribution returns the most probable parent node and class for each text block. Markov Chain Monte Carlo” theory suggests that stationary distribution of samples approach a true joint distribution (Gilks, W.R., 2005). So, by iteratively sampling from posterior conditional distributions until converge to mimic a true joint distribution. A parse tree is obtained using this approach from which information can easily be extracted. An algorithm for detecting horizontal lines and shading is also proposed with the help of pdf2image library and canny edge detector (Canny, 1986). But relying on

hierarchical structure might not be the best choice as people nowadays present their resumes in idiosyncratic ways in order to stand out from the crowd. (Chen et al, 2018) proposes a two-step framework for IE incorporating text block classification to get an intermediate semi-structured output followed by facts identification to get fully structured output. Apache Tika is used to process the resumes into raw text and a heuristic-based algorithm is used to perform trim, split or merge-operations. All the lines in the raw text are tagged as Simple, KeyValue or Complex lines. Simple lines are used to figure out block titles. KeyValue lines are used to compute cosine similarity using TF-IDF & similar values are clustered using the K-means algorithm. A framework called writing style is designed to extract the hidden syntax information of a line. Hidden syntax information is the local format (geography-based, domain-based, etc.) used by people to write a particular block. It takes into input the word and punctuation index, lexical attributes, and classification results trained on entity names collected from the internet using the Naive Bayes classifier. The classifier helps to gain a probability distribution of each phrase in a line about a class. The classification results are mapped with the standardized cluster attributes to get the structured output. The same can be achieved by a neural networks-based pipeline using Attention Bi-LSTM for text block segmentation and a DNN comprised of Bi-LSTM-CNN's-CRF for resume facts identification (Zu et al, 2019). It is observed that the CNN layer is an efficient text feature extractor by conducting an ablation study. Also, BERT emerges as the best algorithm for generating word embeddings among GloVe, BERT, random initialization approach and Word2Vec.

Ranking of resumes is of the essence when it comes to hiring candidates for a specific position. Most of the articles surveyed in this review rank resume using string-matching algorithms from a “skill” knowledge base about a fixed domain. A unique technique independent of the job domain to select candidates based on the job description is to use the Bidirectional Encoder Representations from Transformers (BERT) algorithm (Bhatia et al, 2019). This can be achieved by extracting the work experience (WEX) from a candidate’s resume using any parsing technique and creating positive and negative samples based on the combinations of different WEX of a candidate & combinations of different WEX of different candidates to train BERT. This trained model would then be able to compare a job description with a candidate’s resume and so the candidates can now be ranked based on the degree of similarity between the job description and candidates’ WEX which is quite impressive. A firefly driven optimization algorithm can be used to substantially decrease the time complexity of ranking resumes (Deepak et al, 2020).

Table 5. Effective approaches for Resume Parsing (2015 onwards)

Author/s	Year	Uniqueness	Techniques Used
Lin, Y., Lei, H., Addo, P. C., & Li, X.	2016	Completely ML based data-driven solution, does not use any automatic semantic tool, covers a wide range of job domains and uses a testing set of approx 47k resumes	Word2vec to extract semantic features, K-means & LDA to extract similar features,Ensemble Solution consisting of RF, XGBoost, LSTM & CNN
Chen, J., Gao, L., & Tang, Z.	2016	Specifically deals with parsing of PDF format resumes utilizing it's hierarchical structure to the most	LIBSVM for block segmentation & CRFsuite for sequence labelling
Chen, J., Zhang, C., & Niu, Z.	2018	Does not rely on hierarchical structure of resumes, proposes Writing Style: a feautre to extract hidden syntax information, avoids manual annotation of data	Naïve bayes classifier for text classification, cosine similarity based on TF-IDF & K-means for attribute cluster
Bhatia, V., Rawat, P., Kumar, A., & Shah, R. R.	2019	Proposes a new way that could be used to select candidates based on job description	BERT for sequence-pair classification
Ravindranath, V. K., Deshpande, D., Girish, K. V. V., Patel, D., Jambhekar, N., & Singh, V.	2019	Proposes a statisical generative model to extract both structure and meaning from a resume	Gibbs Sampling Based Algorithm for IE, Canny Edge Detector based algorithm for Heading Detection
Zu, S., & Wang, X.	2019	Has done extensive research on text block segmentation and resume facts identification using muliple neural-networks	TextCNN,RCNN,Adversarial LSTM,Attention BLSTM and Transformer for text block segmentation; Bi-LSTM-CRF,Bi-GRU-CRF,IDCNN-CRF,BLSTM-CNNs-CRF for resume facts identification
Deepak, G., Teja, V., & Santhanavijayan, A.	2020	Proposes an upgrade of firefly optimization algorithm for increase speed of ranking resumes	A custom firefly drive algorithm for ranking resumes with an astounding accuracy of 94.19%

5. Challenges/ethical Deliberations

This section is out of the scope of the SLR and has been included to acquaint the readers with the possible challenges of automating recruitment. It is quite evident that technical advancements including the application of ML in almost every business sector including HR have benefitted recruiters a lot. Apart from the benefits of managing a lot of applications efficiently within a fixed time constraint, another motive for automating recruitment is to remove any sort of bias during the hiring process (Langer et al, 2019). But whether automating recruitment using algorithmic hiring acutely increases bias or not is a question that induces many reservations. Many companies provide platforms using algorithmic decision-making for hiring such as IBM, Microsoft, SAP etc. and many others that utilize such platforms. Algorithmic decision making if not conducted prudently might steer to biasing affecting all echelons in a

corporate's hierarchy. Though the chance of biasing is much higher in automated pre-employment assessment tests as the inputs to such algorithms are ambiguous in nature (Raghavan et al, 2020), resume screening can also suffer from biasing depending upon the dataset used for training. Years of organizational audits have revealed that employees face a lot of implicit and explicit discrimination in terms of gender, race, caste, colour, and other factors (Bendick Jr et al, 2012). So, in case a company knowingly or unknowingly has a biased employee demographic that is used to train the model to hire candidates based on their skill set, university, work experience etc., they might end up hiring a highly skewed batch of employees failing the whole purpose of automating resume parsing. One such incident took place with Amazon Inc in which male dominance in the company ensured a bias against women in their newly developed automated recruiting system. The system developed using an ML model taught itself to favour male candidates over females. Such a state of affairs where a segment of people is adversely affected is referred to as disparate impact (Perry, P. L., 1990). Apart from disparate impact, differential validity is another cause for bias. In simple terms, a knowledge base used to parse resumes might be valid for one group of candidates and not for the other. Another aspect to look at is the automated advertisements employed by hiring portals to seek candidates. Advertisements use predictive analytics on the backend to discover a fitting pool of candidates. This again is based on previous hiring decisions leading to historical bias. Hence, the objective of achieving impartiality conflicts with the impulse to personalize hiring (Burke et al, 2018).

6. Conclusion

It is now apparent that the recruiting world has forged ahead by dint of technological advancements remarkably & is evolving rapidly owing to the extensive research in the field of ML and DL. The SLR conducted can corroborate the significance of automating the parsing of resumes. A combination of ML frameworks along with existing DL based algorithms are used predominantly in most of the research work. Utilizing neural networks to their utmost potential by using custom combinations can open the door to many new possibilities. Also, more work needs to be done in constructing optimized universal parsing solutions that do not rely on a domain-restricted knowledge base and if it does, it should be automated. Much research is required on optimizing parsing models to remove bias. After coalescing all the findings of the review, it can be concluded that extracting valuable information from resumes using ML & DL eases out the hiring process to a great extent and neural networks reduce the time complexity by a huge margin.

Declarations

Funding

No funding was received for conducting this study.

Conflicts of interest/Competing interests

The authors have no financial or proprietary interests in any material discussed in this article. The authors have no relevant financial or non-financial interests to disclose.

Availability of data and material

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability

Not Applicable

Authors' contributions

Dr Seema Rawat conceived and designed the study, Mr Aakankshu Rawat & Mr Siddharth Malik performed the research, Dr Deepak Kumar analyzed the data, and Dr Praveen Kumar contributed to editorial input. All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [Mr Aakankshu Rawat], [Mr Siddharth Malik] and [Dr Deepak Kumar]. The first draft of the manuscript was written by [Mr Aakankshu Rawat] and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval

Not Applicable

Consent to participate

Not Applicable

Consent for publication (include appropriate statements)

Not Applicable

References

1. Marciano VM (1995, August) THE ORIGINS AND DEVELOPMENT OF HUMAN RESOURCE MANAGEMENT. In *Academy of Management proceedings* (Vol. 1995, No. 1, pp. 223–227). Briarcliff Manor, NY 10510: Academy of Management
2. Ewen RB, Smith PC, Hulin CL (1966) An empirical test of the herzberg two-factor theory. Journal of applied psychology 50(6):544
3. Deci EL, Ryan RM (1980) Self-determination theory: When mind mediates behavior. *The Journal of mind and Behavior*, 33–43
4. Tsujii JI (2011, February) Computational linguistics and natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 52–67). Springer,

Berlin, Heidelberg

5. Kitchenham B (2004) Procedures for performing systematic reviews (Joint Technical Report). *Software Engineering Group, Department of Computer Science, Keele University and Empirical Software Engineering National ICT Australia Ltd*
6. Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering–a systematic literature review. Inf Softw Technol 51(1):7–15
7. Filatova O (2016) More than a word cloud. Tesol Journal 7(2):438–448
8. Yu K, Guan G, Zhou M (2005, June) Resume information extraction with cascaded hybrid model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 499–506)
9. Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surveys 34(1):1–47
10. Katz S (1987) Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans Acoust Speech Signal Process 35(3):400–401
11. Gale WA, Sampson G (1995) Good-turing frequency estimation without tears. Journal of quantitative linguistics 2(3):217–237
12. Broder AZ (2000, June) Identifying and filtering near-duplicate documents. In *Annual Symposium on Combinatorial Pattern Matching* (pp. 1–10). Springer, Berlin, Heidelberg
13. Lafferty J, McCallum A, Pereira FC (2001) Conditional random fields. Probabilistic models for segmenting and labeling sequence data
14. Singh A, Rose C, Visweswariah K, Chenthamarakshan V, Kambhatla N (2010, October) PROSPECT: a system for screening candidates for recruitment. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 659–668)
15. Satheesh K, Jahnavi A, Iswarya L, Ayesha K, Bhanusekhar G, Hanisha K (2020) Resume Ranking based on Job Description using SpaCy NER model
16. Maheshwari S, Reddy PK (2009, August) Discovering special product features for improving the process of product selection in e-commerce environment. In *Proceedings of the 11th International Conference on Electronic Commerce* (pp. 47–56)
17. Maheshwari S, Sainani A, Reddy PK (2010, March) An approach to extract special skills to improve the performance of resume selection. In *International Workshop on Databases in Networked Information Systems* (pp. 256–273). Springer, Berlin, Heidelberg
18. Chifu ES, Chifu VR, Popa I, Salomie I (2017, September) A system for detecting professional skills from resumes written in natural language. In *2017 13th IEEE international conference on intelligent computer communication and processing (ICCP)* (pp. 189–196). IEEE
19. Çelik D, Elçi A (2012, November) An ontology-based information extraction approach for résumés. In *Joint international conference on pervasive computing and the networked world* (pp. 165–179). Springer, Berlin, Heidelberg

20. Senthil Kumaran V, Sankar A (2013) Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (EXPERT). Int J Metadata Semant Ontol 8(1):56–64
21. Abirami AM, Askarunisa A, Sangeetha R, Padmavathi C, Priya M Ontology based ranking of documents using Graph Databases: a Big Data Approach
22. Mohamed A, Bagawathinathan W, Iqbal U, Shamrath S, Jayakody A (2018, December) Smart Talents Recruiter-Resume Ranking and Recommendation System. In *2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)* (pp. 1–5). IEEE
23. Bojārs U, Breslin JG (2007, January) ResumeRDF: Expressing skill information on the Semantic Web. In *1 st International ExpertFinder Workshop*
24. Chandola D, Garg A, Maurya A, Kushwaha A (2015) Online Resume Parsing System Using Text Analytics. *Journal of Multi Disciplinary Engineering Technologies (JMDET) vol*, *9*
25. Kopparapu SK (2010, December) Automatic extraction of usable information from unstructured resumes to aid search. In *2010 IEEE International Conference on Progress in Informatics and Computing* (Vol. 1, pp. 99–103). IEEE
26. Thelen M, Riloff E (2002, July) A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002)* (pp. 214–221)
27. Pawar S, Srivastava R, Palshikar GK (2012, January) Automatic gazette creation for named entity recognition and application to resume processing. In *Proceedings of the 5th ACM COMPUTE Conference: Intelligent & scalable system technologies* (pp. 1–7)
28. Farkas R, Dobó A, Kurai Z, Miklós I, Nagy Á, Vincze V, Zsibrita J (2014) Information extraction from Hungarian, English and German CVs for a career portal. In: Mining intelligence and knowledge exploration. Springer, Cham, pp 333–341
29. Chen J, Niu Z, Fu H (2015, June) A novel knowledge extraction framework for resumes based on text classifier. In *International Conference on Web-Age Information Management* (pp. 540–543). Springer, Cham
30. Tikhonova M, Gavrishchuk A (2019, November) NLP methods for automatic candidate's CV segmentation. In *2019 International Conference on Engineering and Telecommunication (EnT)* (pp. 1–5). IEEE
31. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5:135–146
32. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*
33. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*
34. Mikolov T, Yih WT, Zweig G (2013, June) Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the*

association for computational linguistics: Human language technologies (pp. 746–751)

35. Pennington J, Socher R, Manning CD (2014, October) Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543)
36. Pham Van L, Ngoc Vu, S., & Van N, V (2018) Study of Information Extraction in Resume. Conference
37. Dos Santos C, Zadrozny B (2014, June) Learning character-level representations for part-of-speech tagging. In *International Conference on Machine Learning* (pp. 1818–1826). PMLR
38. Chiu JP, Nichols E (2016) Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics 4:357–370
39. Pascanu R, Mikolov T, Bengio Y (2012) Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, *2*(417), 1
40. Caruana R, Lawrence S, Giles L (2001) Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Advances in neural information processing systems*, 402–408
41. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15(1):1929–1958
42. Ayishathahira CH, 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). IEEE
43. Maheshwary S, Misra H (2018, April) Matching resumes to jobs via deep siamese network. In *Companion Proceedings of the The Web Conference 2018* (pp. 87–88)
44. Jacob F, Javed F, Zhao M, Mcnair M (2014, May) sCooL: A system for academic institution name normalization. In *2014 international conference on collaboration technologies and systems (CTS)* (pp. 86–93). IEEE
45. McCandless M, Hatcher E, Gospodnetić O, Gospodnetić O (2010) Lucene in action, vol 2. Manning, Greenwich
46. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11(1):10–18
47. Sayfullina L, Malmi E, Liao Y, Jung A (2017, July) Domain adaptation for resume classification using convolutional neural networks. In *International Conference on Analysis of Images, Social Networks and Texts* (pp. 82–93). Springer, Cham
48. Mittal V, Mehta P, Relan D, Gabrani G (2020) Methodology for resume parsing and job domain prediction. Journal of Statistics Management Systems 23(7):1265–1274
49. Zhang C, Wang H (2018) Resumevis: A visual analytics system to discover semantic information in semi-structured resume data. ACM Transactions on Intelligent Systems Technology (TIST) 10(1):1–25
50. Lin Y, Lei H, Addo PC, Li X (2016) Machine learned resume-job matching solution. *arXiv preprint arXiv:1607.07657*

51. Chen J, Gao L, Tang Z (2016) Information extraction from resume documents in pdf format. Electronic Imaging 2016(17):1–8
52. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM transactions on intelligent systems technology (TIST) 2(3):1–27
53. Okazaki N (2007) Crfsuite: a fast implementation of conditional random fields (crfs)
54. Chen J, Zhang C, Niu Z (2018) A two-step resume information extraction algorithm. *Mathematical Problems in Engineering*, *2018*
55. Bhatia V, Rawat P, Kumar A, Shah RR (2019) End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT. *arXiv preprint arXiv:1910.03089*
56. Ravindranath VK, Deshpande D, Girish KVV, Patel D, Jambhekar N, Singh V (2019, September) Inferring structure and meaning of semi-structured documents by using a gibbs sampling based approach. In *2019 International Conference on Document Analysis and Recognition* Workshops *(ICDARW)* (Vol. 5, pp. 169–174). IEEE Computer Society
57. Gilks WR (2005) M arkov Chain M onte C arlo. *Encyclopedia of biostatistics*, *4*
58. Canny J (1986) A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 679–698
59. Zu S, Wang X RESUME INFORMATION EXTRACTION WITH A NOVEL TEXT BLOCK SEGMENTATION ALGORITHM
60. Deepak G, Teja V, Santhanavijayan A (2020) A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. Journal of Discrete Mathematical Sciences Cryptography 23(1):157–165
61. Langer M, König CJ, Papathanasiou M (2019) Highly automated job interviews: Acceptance under the influence of stakes. International Journal of Selection Assessment 27(3):217–234
62. Raghavan M, Barocas S, Kleinberg J, Levy K (2020, January) Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 469–481)
63. Bendick M Jr, Nunes AP (2012) Developing the research basis for controlling bias in hiring. J Soc Issues 68(2):238–262
64. Perry PL (1990) Two Faces of Disparate Impact Discrimination. Fordham L Rev 59:523
65. Burke R, Sonboli N, Ordonez-Gauger A (2018, January) Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency* (pp. 202–214). PMLR
66. Le Q, Mikolov T (2014, June) Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188–1196). PMLR
67. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P (2011) Natural language processing (almost) from scratch. Journal of machine learning research 12(ARTICLE):2493–2537

68. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*

Figures

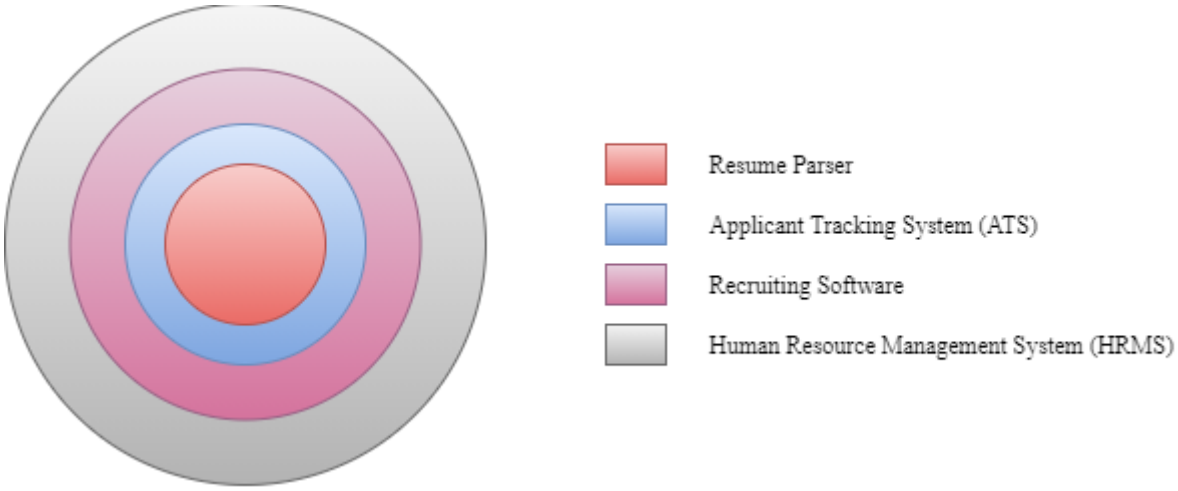

Figure 1

Venn Diagram of Various HR solutions (software) explaining the interrelationship amongst them

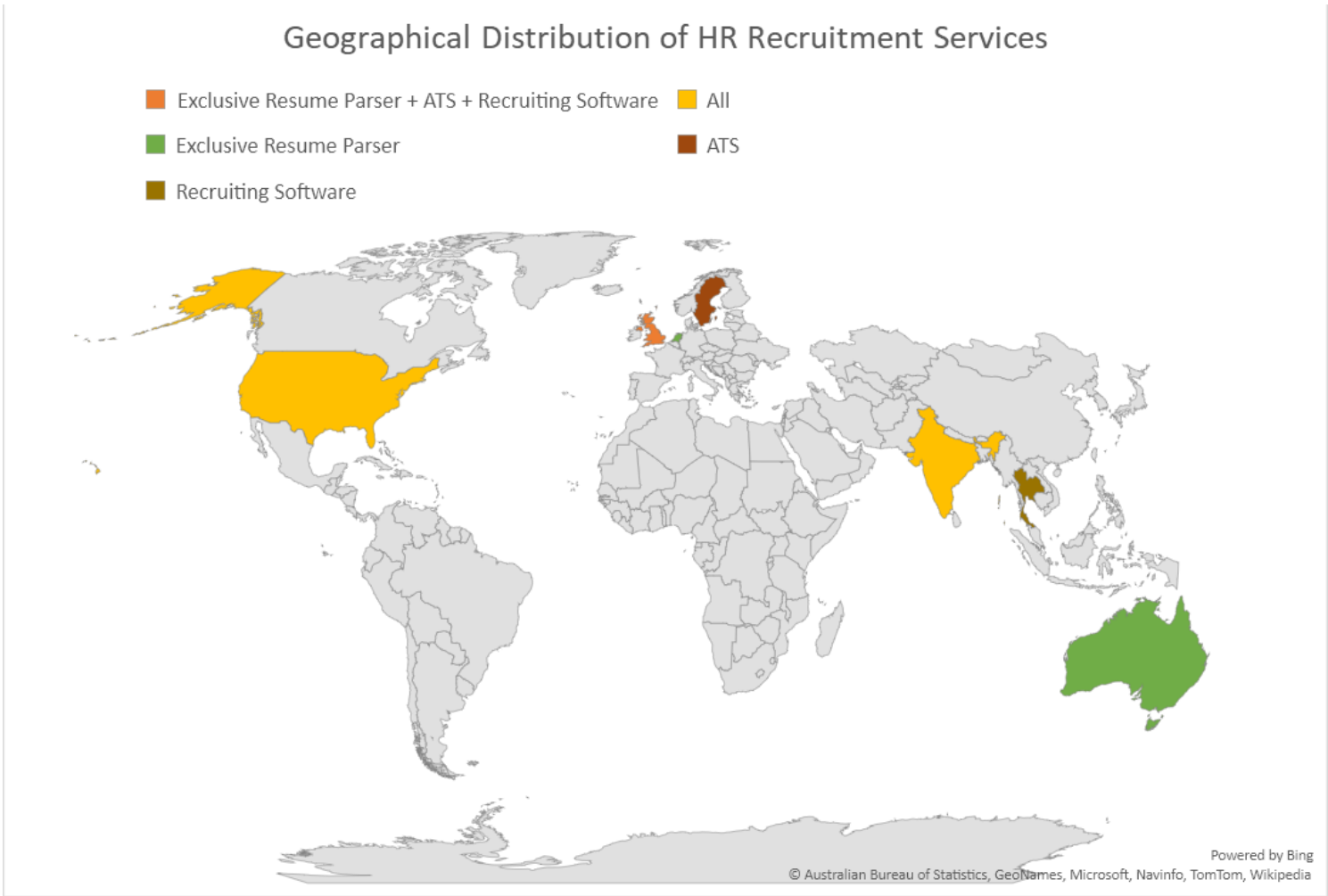

Figure 2

Geographical Distribution of HR Recruitment Services (Implicit Parsing Solutions)

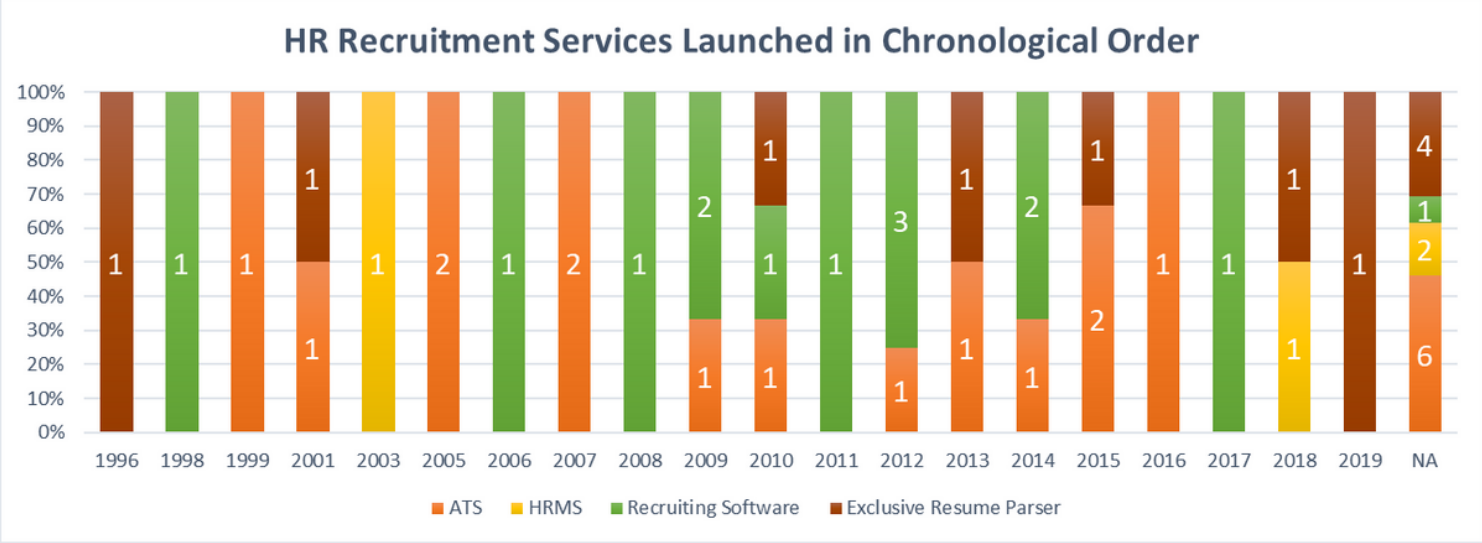

Figure 3

100% Stacked Column Chart of HR Recruitment Services (Implicit Parsing Solutions) launched in Chronological Order

Figure 4

"Search Criteria" adopted for Data Retrieval

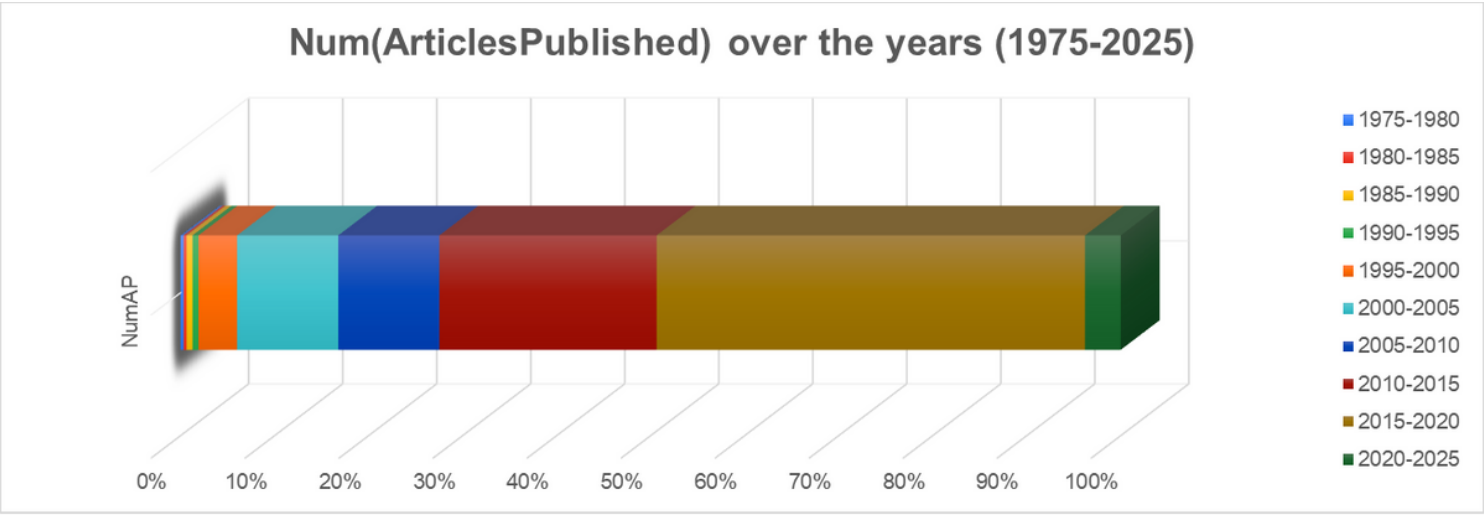

Figure 5

Number of articles published over the years (1975-2025)

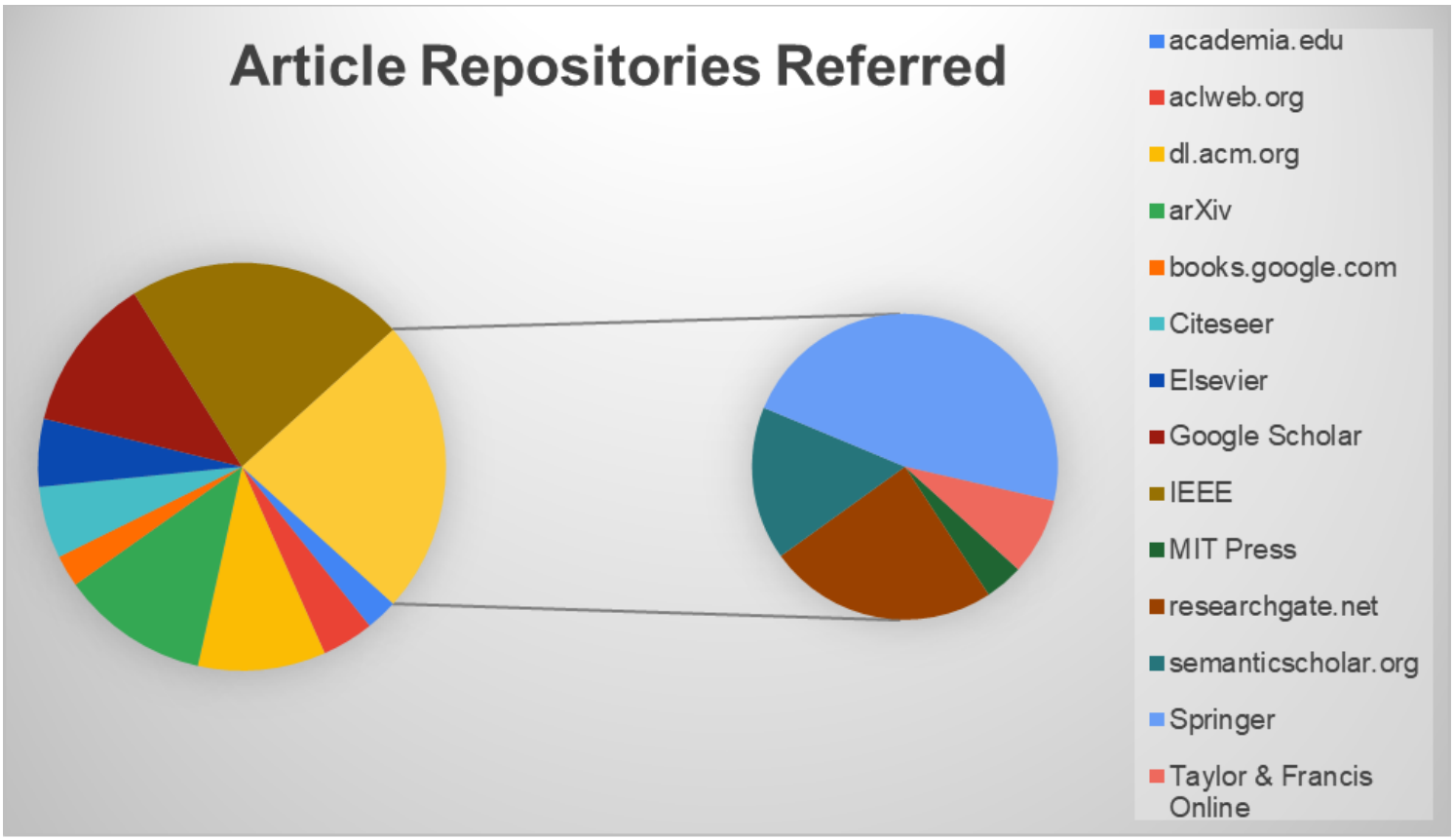

Figure 6

Distribution of articles in repositories referred

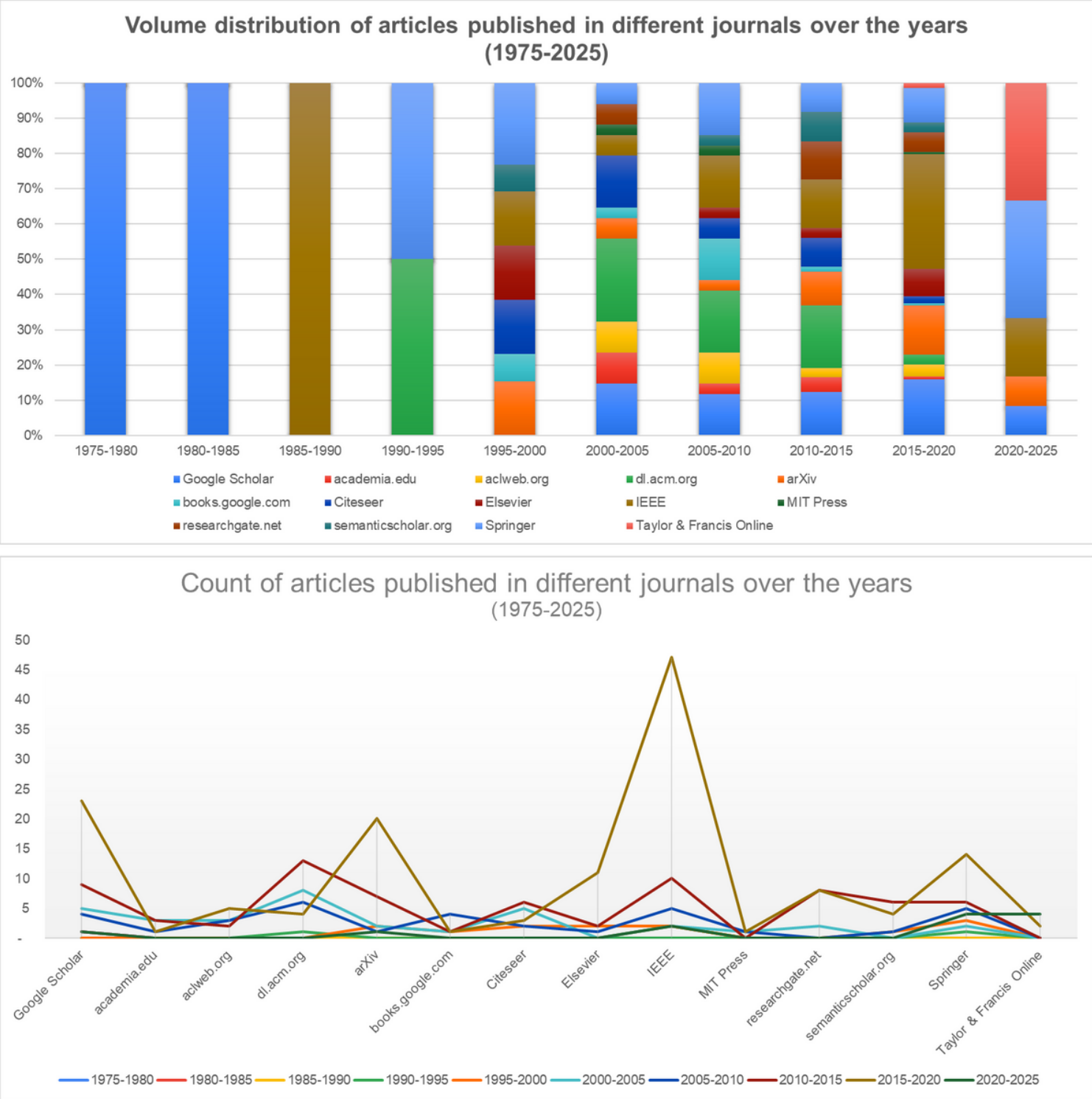

Figure 7

(a) Volume distribution of articles published in different journals over the years (b) Count of articles published in different journals over the years