IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 2, Ser. 3 (Mar. – Apr. 2025), PP 70-82
www.iosrjournals.org

Scrutinizing the Collection of Datasets Dedicated to Resume Screening Leveraging NLP and Machine Learning

Ankit Rai[1], Gaurav Goel[2], Dinesh Kr. Singh[3], Shobhit Shukla[4]
[1]*Department of Computer Science & Engineering, Dr. Shakuntala Misra National Rehabilitation University Lucknow, India*
[2]*Assistant Professor, Department of Computer, Dr. Shakuntala Misra National Rehabilitation University Lucknow, India*
[3]*Assistant Professor, Department of Computer, Dr. Shakuntala Misra National Rehabilitation University Lucknow, India*
[4]*Assistant Professor, Department of Computer, Dr. Shakuntala Misra National Rehabilitation University Lucknow, India*

Abstract:
The employment market in India continues to expand rapidly. The Human Resources (HR) team
The recruitment process presents major difficulties to businesses which need to select ideal candidates for open positions. The search for the best candidate requiring resume screening spans a significant number of applications. The recruitment process becomes exponentially harder for companies extending mass positions. HR personnel face major challenges while screening numerous thousands of applications due to their requirement of locating top candidates. Various automated systems exist to speed up the effort of shortlisting candidates for different positions. HR managers achieve simplified applicant identification through machine learning models that analyze the large number of submissions using these tools. This study evaluates different market-available research resources with the goal of identifying shortcoming in existing solutions. [1][2] The process of identifying suitable candidates can be sped up through recognition of these performance weaknesses.
Key words: *NLP, Machine Learning models, resume, KNN. SVM*

I. Introduction

During the past few years the Indian hiring market expanded substantially because both job opportunities emerged and businesses needed competitively priced employees. Specialized firms now specialize in delivering qualified candidates solely to businesses through outsourcing the hiring process because the recruitment sector mirroring employment market expansion. [3] The approach of bulk recruiting demands substantial business resources that affect productivity so organizations need this outsourcing strategy. [4] Array-like tools allow talent acquisition firms to implement machine learning models that evaluate exceptional resumes thus simplifying employment selection and making work easier on HR departments.

Exploring the best candidates extends both time and expense because many jobseekers apply. The identification of top candidates becomes inefficient due to traditional manual resume reviews. [5] Job description matching pauses through automated systems using data-based methods such as cosine similarity in addition to other calculations. [6][7] Job candidates receive automated evaluation using K-Nearest Neighbors (KNN) algorithms to match their qualifications to job profiles. Text parsing through algorithmic testing demonstrates an 85% success rate together with a 92% ranking precision. [8] [9]

Data summary needs to become faster and more precise because the medical sector requires it as health records transition toward computerization. Numerous records become reviewable through this method which enables medical personnel to maintain high efficiency. The utilization of abstractive text summarization methods backed by Pegasus and BART and T5 together with preprocessing and data augmentation strategies summarizes patient progress notes in this case. Pegasus delivered the highest performance according to ROUGE-L score analysis with 0.2744 F1 score. [10][11]

The NLU industry achieved substantial progress across multiple NLU tasks following Masked Language Modeling (MLM) implementation. The conceptual basis behind MLM remains unclear despite its creation. The remedy to this situation can be found through the implementation of Descriptive Masked Language Modeling (DMLM). The term definition input of DMLM enhances the comprehension difficulty because it allows models to predict contextual words based on the descriptions they receive. The process enables semantic awareness in following tasks while outperforming standard MLM particularly on Semantic Role Labeling tasks. [12]

The identification of essential themes within discussions proves difficult during dialogue summarization work especially throughout normal informal dialogues. KADS represents a Keyword-Aware Dialogue Summarization System to address the mentioned issue. The key phrases extracted from conversations serve as input for KADS to create summaries based on transformer architecture. [13] The extensive testing across various datasets validated that KADS delivers superior results than baseline models especially when resources are restricted.

Software companies lead the expansion of IT recruitment through the hiring of college graduates. The HR department normally sustains responsibility for assigning work duties to newly hired employees through a manual process which requires substantial time and workload. Due to AI along with machine learning most companies now use automated systems that improve operational effectiveness. Through a voting classifier derived from ensemble learning a suggested resume classification program identifies appropriate subjects for candidates by considering their skills and work experience. [14]Stack Overflow APIs help determine suitable subjects for candidates who do not pass the classification confidence threshold. Topic modeling serves to generate new subjects for identification. The system generates better support for new domains through its automatically enhanced accuracy which is achieved through continuous retraining.

The development of resume parsing technologies has been affected due to recent advancement in distributed representations together with deep neural networks. This paper introduces a neural network which uses distributed embeddings together with an end-to-end pipeline to process resumes. The approach structures resume through classifiers that join line position data to text block semantic meanings for segmentation purposes. The detection of named entities within these segments depends on sequence labeling classifiers. The system's successful operation can be verified through comparative evaluations that identify BLSTM-CNNs-CRF as the most effective model for named entity recognition tasks. [15] This method achieves high parsing effectiveness according to resume parsers comparison results.

II. Related Work

A. About existing systems

The technology, models, and procedures now in use in medical record summarization, dialogue summarization, resume parsing, recruitment, and related disciplines. An examination of current systems based on the different domains you mentioned in your introduction is provided below:

1. Recruitment Systems

Recruitment systems in present times mostly rely on traditional approaches which involve both manual resume evaluations and in-person interview sequences. Organizations now seek help from outsourced recruitment agencies as their need for mass hiring expands because these agencies implement software tools alongside algorithms for higher efficiency. The technologies rely mainly on three elements including applicant tracking systems (ATS), rule-based systems and keyword matching mechanisms. Workday and Jobvite alongside other well-known applicant tracking systems (ATS) use predefined criteria to filter resumes through their established keyword systems. These algorithms face difficulties understanding the meaning and background information of candidates which exacerbates their struggle to select relevant candidates when dealing with big datasets. Machine learning models operate within specific systems in order to enhance candidate matching but these models still need better accuracy development. [16]

2. Resume Parsing Systems

The present resume parsing systems aim to extract essential information from resumes by handling contact data combined with educational backgrounds and professional experience and qualifications. Standard parsers using rule-based systems were designed to read pre-defined sections but do not handle unstructured or varied resume formats. These systems impose restrictions because they primarily rely on regular expressions and pattern matching in their processing of numerous resume formats and unorganized documents. The latest resume processing algorithms split resumes across multiple employment sectors and match them to position specifications through the combination of machine learning models with Natural Language Processing (NLP). Due to use of pre-determined keywords and basic classification models the existing advanced computing algorithms do occasionally struggle with sematic understanding and context. BLSTM-CNNs-CRF and neural networks represent deep learning models which function as effective named entity recognition solutions for resume parsing by providing superior precision than traditional approaches. [17]

3. Dialogue Summarization Systems

Two major implementations of dialogue summarization models currently exist as extractive and abstractive systems. Extractive summary selects key phrases as its main operation but fails to create meaningful or coherent summaries. The summary produced by abstractive technology does not need to repeat exact sentences from the original text but effectively represents essential information. The transformer-based BART and T5 and PEGASUS models produce human-like summaries that lead to exceptional outcomes in abstraction-based discourse summary. Almost all current algorithms struggle to work with informal language along with colloquial speech which results in poor summary quality. Many current system face limitations because they only have access to minimal training data of good quality. [18]

4. Medical Record Summarization Systems

Medical record summarizing systems today aim to decrease medical staff reading and interpreting patient documentation duration. Patient progress notes receive extraction summary processing by most systems that operate in this field. The growing use of abstractive summarization occurs because of its ability to generate adaptable summaries which are also highly succinct. The medical community uses PEGASUS BART and T5 along with clinical notes and medical records as their training datasets. Current medical systems have managed to achieve positive results yet still struggle with generating wrong summaries due to their difficulties handling domain-specific language and medical jargon. Closing analysis systems fall short concerning medical records understanding because they depend on basic rule-based extraction techniques from many such applications. [19]

5. Masked Language Models (MLM)

Multiple NLP tasks have gone through a transformation using Masked Language Models (MLM) including BERT and its subsequent models that understand word context. Without explicit semantic information in their training the MLM models can achieve lower success rates for deep linguistic tasks such as word sense disambiguation as well as semantic role categorization. The performance of existing MLM models deteriorates on domain-specific tasks without data-adjustments since they originally processed general corpora collections. The existing media information systems work with traditional MLM processes which display limited functionality during explicit semantic grounding requirements despite the proposed value of DMLM (Descriptive Masked Language Modeling). [20]

6. Project Allocation in Software Companies

The process of manual project assignment in new hires by HR departments in software organizations produces operational failures as well as increased staffing costs. Project assignment operates through alignment of candidates with projects based on qualification and expertise details that occur either manually or through simple algorithm rules. The manual solutions are unable to utilize AI technologies when adjusting project requirements and processing dynamic new data inputs. Voting classifiers in combination with ensemble learning methods are developing in popularity because they boost the accuracy of project assignments. The existing multiple systems face difficulties in adapting to changes in project domains and requirements which leads to inferior project distribution results. [21]

7. Topic Modeling and Knowledge Extraction

The existing systems for detecting topics and extracting knowledge from unstructured text utilize two techniques that are LDA with Dirichlet Allocation and TF-IDF. The techniques demonstrate success when locating patterns in big datasets even though they lack adaptability to new subjects or fields. More precise modeling of topics requires some systems to retrieve data from outside sources using APIs such as Stack Overflow REST APIs. Complete integration of these models into the hiring processes has not been achieved so far which results in reduced flexibility and limited real-time application. [22][23]

The present systems operating in these domains function effectively but restrict their capabilities to deal with unstructured or informal inputs and make adjustments dynamically and understand semantic content. Advanced systems implement machine learning models through ensemble classifiers and transformers and neural networks yet these technologies remain underutilized because more improvements need to occur to handle real-world needs of diverse and flexible large-scale tasks. The crucial areas of understanding data diversity and establishing semantic grounding need additional research innovation to resolve major problems. [24]

Table 1: Summarizing Table of Related Work.

Category	Description
Recruitment Systems	Traditional hiring methods involve manual resume screening and interviews. However, outsourcing and AI-powered applicant tracking systems (ATS) are gaining traction. These systems still struggle with semantic understanding and context.
Resume Parsing Systems	Traditional rule-based systems extract key details from resumes but struggle with unstructured formats. Advanced models like BLSTM-CNNs-CRF and deep learning methods improve named entity recognition.
Dialogue Summarization Systems	Extractive summarization selects key phrases but lacks deep understanding. Abstractive methods using BART, T5, and PEGASUS generate more human-like summaries. Challenges remain in handling informal language and limited training data.
Medical Record Summarization	Extractive methods are commonly used, but abstractive summarization with PEGASUS, BART, and T5 is growing. Domain-specific issues like medical jargon and incomplete records hinder performance.
Masked Language Models (MLM)	BERT and similar models improve NLP tasks but lack explicit semantic understanding. Descriptive Masked Language Modeling (DMLM) aims to enhance meaning comprehension in context.
Project Allocation in Software Firms	Traditional project assignments rely on manual matching. AI-based voting classifiers and machine learning improve dynamic allocation, though many systems lack real-time adaptability.
Topic Modeling & Knowledge Extraction	Traditional LDA and TF-IDF methods identify patterns but struggle with new domains. Neural-based topic models improve accuracy and adaptability for recruitment and project allocation.

B. Types of systems

The term "types of systems" describes the many methods and strategies used in the disciplines of medical record summarization, dialogue summarization, resume parsing, recruitment, and related fields. A summary of the various system types now in use in these fields may be seen below:

1. Recruitment Systems

The following categories can be used to broadly classify recruitment systems:

- Applicant Tracking Systems (ATS): The central role of Applicant Tracking Systems (ATS) includes managing entire hiring operations. The system performs automatic execution of job posting activities followed by resume submission alongside candidate evaluation and interview organization. Jobvite and Greenhouse along with Workday operate as leading ATS platforms that filter candidate resumes based on job descriptions through keyword-based recognition systems. [25][26]
- Platforms for Outsourced Recruitment: These recruitment platforms allow businesses to hire external recruitment firms for managing their employee acquisition activities. Agencies reduce human workload in candidate selection by using specific automated tools alongside machine learning models as part of their evaluation systems. AI-based platforms that find suitable matches between candidates and job openings maintain two solutions known as HireVue and LinkedIn Recruiter. [27]
- Recruitment Chatbots: They utilize artificial intelligence to operate through messaging apps where they take over initial candidate evaluation tasks within the hiring process. These services provide immersive support which enables initial candidate examination while handling both initial screening and scheduling obligations. Two well-known recruitment solutions include XOR and Mya. [28]

2. Resume Parsing Systems

The following types constitute resume parsing systems:

These programs extract specified resume data thanks to established rules that apply regular expressions to recognize name and contact information and skills along with work history. The systems remain easy to operate but become unsuccessful when processing standard information and varied resume designs. [29]

The programs utilize supervised learning to train their models which identifies and pulls information from resumes. The models acquire accuracy improvement by processing labeled datasets. Three common algorithms in these systems are Neural networks together with Random Forests and Support Vector Machines (SVM). [30]

The advanced deep learning parsers use Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs) as deep learning models to accurately extract difficult information from resumes. Data parsers utilizing deep learning technology maintain higher performance rates when processing different resume structures and unorganized information. [31]

End-to-end resume classification systems use extracted information to place resumes into different job groups. Voting classifiers and stacking are the two ensemble approaches that achieve high classification precision by allowing two models to join together their output.[32][33]

3. Dialogue Summarization Systems
Dialogue summarizer software operates within three main categories.

Systems in the extractive summarizing category generate summary content through direct selection of crucial text portions from the input material. Although extractive summarization is efficient but it has its limitations since it fails to catch some important contents and fails to preserve the originality of the conversation organization

The abstractive summarization system functions to create essential summary content without linear sentence replication through paraphrased information retrieval. The systems perform context analysis and pick vital details to generate human-like summaries through NLP models including BART, PEGASUS and T5.

These systems use keyword extraction techniques to focus on essential discussion points for summary creation through a Keyword-Aware discussion Summarization System (KADS). They implement these keywords through their summary processes and produce summaries using pre-trained models and keyword extractors to select important subjects. [34]

4. Medical Record Summarization Systems
Medical record summary systems exist in two different categories.

Office automation tools utilize extractive summarization systems to directly remove essential information from medical patient data such as observations and diagnosis and treatment records. This method allows for simple extraction yet crucial background information might get unnoticed.

Two systems named PEGASUS along with BART create abstract summaries which take medical information to produce condensed outputs while retaining core elements. Medical terminologies need effective management along with high-quality data to achieve the goal of developing adaptable short summaries of medical content.

These systems give more precise summaries when extractive and abstractive techniques work together. When using extracted content for creating logical succinct summaries an abstractive model can be applied after selecting the most essential information. [35]

5. Masked Language Models (MLM)

The basic technology used in present-day NLP applications known as MLM shares connections with different systems.

Standard Masked Language Models (MLM) including BERT and RoBERTa among their members because they learn to predict missing words by analyzing phrase contexts. The models execute outstanding performance for tasks that include named entity recognition (NER) among other operations such as sentiment analysis and text classification.

Through descriptive masked language models (DMLM) the MLM technique gains semantic grounding by providing a word description in addition to contextual information. The designed system aims to improve model efficiency in semantic role labeling and context-based NLU work that requires detailed word meaning understanding. [36]

6. Project Allocation Systems
Software businesses organize their project allocation system into two categories:

The HR teams implement manual allocation systems that depend on human personnel to distribute work assignments to new hires based on their background experience as well as personal interests and qualifications. Manual project allocation turns out to be both time-intensive and ineffective when organizations conduct mass recruitments.

Such systems assign candidates to projects by using preset criteria between applicants' skills and project necessity. These systems supply better efficiency than manual approaches yet preserve fewer flexibility compared to manual strategies.

The procedural structure of machine learning-based project allocation systems selects proper projects for each applicant through analysis of competence specialties combined with skill potential and work performance history. Project allocation precision becomes enhanced through the implementation of ensemble learning methods and voting classifiers in many systems.

These systems possess built-in functionality to respond to project needs that alter or receive new information. Project allocation suggestions benefit from real-time feedback systems and continuous learning procedures used by these systems. The systems have the ability to automatically add new domains to their model framework while incrementally boosting accuracy through data retraining processes that use Stack Overflow REST APIs. [37]

7. Topic Modeling and Knowledge Extraction Systems
The different categories for topic modeling and knowledge extraction systems consist of three types.

The systems under this category use Latent Dirichlet Allocation (LDA) to discover themes within enormous text collections. As an unsupervised technique LDA transforms documents into theme collections although it fails to adequately deal with complex topics or overlapping themes.

The simplified analysis framework consists of two parts: Bag of Words (BoW) as well as term frequency-inverse document frequency (TF-IDF) models which identify key terms and concepts. Multiple systems generate proper outcomes in their various applications although they struggle to process semantic content and complex textual connections between elements.

Modern text subject modeling occurs through neural networks together with autoencoders as part of sophisticated neural Topic Modeling Systems. The models achieve better subject detection in extensive unstructured datasets through their ability to understand complex word-subject linkages. [38]

The different sectors use systems that provide varying degrees of functionality alongside complexity levels. Modern deep learning and machine learning approaches replace conventional systems because they offer better performance against the key limitations such as strict rule follow up and keyword matching and semantic understanding deficiencies. These advanced systems offer enhanced accuracy together with adaptability and flexibility which makes medical record summarization and resume parsing as well as dialogue summarization tasks more efficient. System performance continues to increase as ensemble models and hybrid approaches become increasingly popular because they unite the best components of multiple techniques.

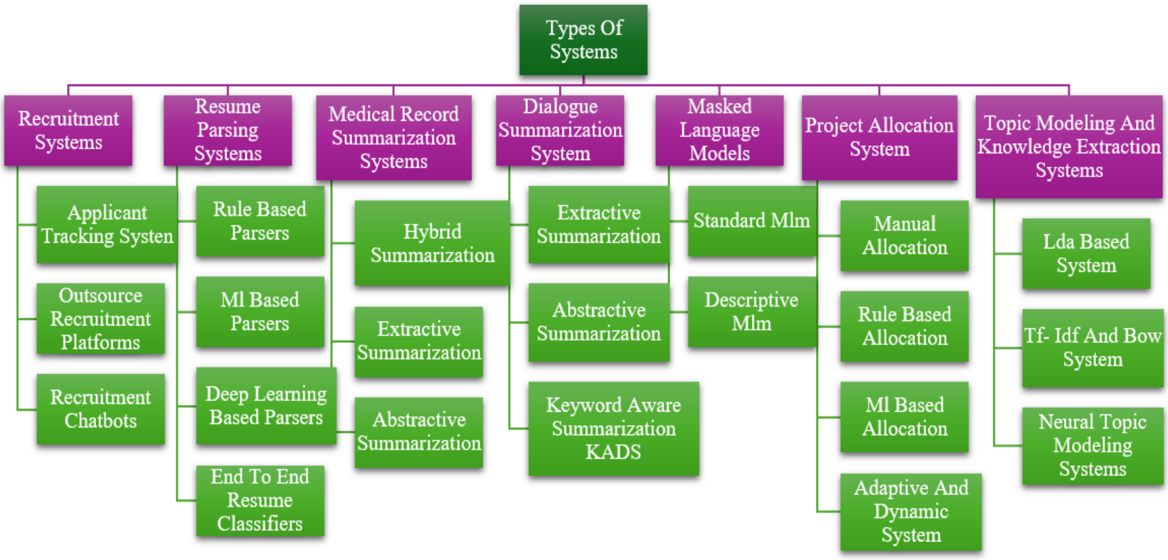

Figure 1: Flowchart of Types of System.

C. History of technology
Resume parsing, dialogue summarization, medical record summarization, and other related fields are examples of how techniques and tools for automating and improving these procedures have evolved in the context of hiring. An examination of how technology has advanced in several areas is provided below:

1. Recruitment Technology
The early automation technologies of the late twentieth century launched the historical development of recruitment technology.
During the time period of 1960s through the 1980s mainframe systems began to automate job posting functions and application and candidate record management acted as the initial phase in automated recruitment processes. Such early system processing required people to handle applications manually which created extensive screen perfection needs.
ATS Applicant Tracking Systems became popular throughout the 1990s because database management systems combined with internet technologies. Recruitment platforms obtained the ability to advertise job positions through online platforms while receiving applications and conducting basic keyword search comparisons of resumes due to these technological developments. The first systems which automated resume screening and candidate monitoring were developed by Taleo and then followed by Kenexa.
A revolutionary development in recruiting technologies occurred when artificial intelligence (AI) integration with machine learning (ML) entered the scene during the 2000s. Hiring platforms like HireVue and LinkedIn Recruiter

utilized AI technology to provide assistance during resume screening while conducting initial interviews and conducting better applicant evaluations according to job requirements. The efficiency of applicant assessments and system automation were improved through the adoption of deep learning together with natural language processing (NLP) by these platforms. [18]
The recruitment process has experienced a dramatic change throughout the 2010s to present because organizations now outsource their hiring operations while employing specialized recruitment chatbots and automated scheduling applications. The recruitment assistants operated by Mya Systems and XOR utilize AI technology to reduce human resource staff commitment to candidate screening through manual processes.

2. Resume Parsing Technology

Resume parsing technologies developed due to increasing business demands for resume data automation and organization processes.

Rule-Based and Manual Systems operated between the 1980s and 1990s as the first resume parsing technologies using basic rule-based approaches and manual procedures. Organizations needed people along with rigid decision rules to extract standard resume data items including names, educational details and abilities from these initial systems. The OCR process converted resumes stored in scanned paper format into text. [12]

Pattern matching along with keyword-based parsing systems emerged in the 2000s because digital resumes gained popularity during this period. These systems identified specific categories such as name and abilities through regular expressions together with basic algorithmic evaluations. This type of resume-processing system demonstrated inflexibility because it failed to adapt to various format requirements and linguistic complexity.

Machine Learning together with NLP-Based Parsers revolutionized resume parsing during the 2010s and continuing up to the present. The algorithms obtained better understanding of resume structure alterations through this method of learning from identified data. NER and other technologies achieved widespread adoption to extract detailed sophisticated information from resumes. The combination of deep learning models including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) was implemented to enhance parsing accuracy according to the literature. [30]

Through end-to-end system development from 2015 to the present-day classifiers emerged to detect project domains and job roles combined with resume parsing through formatting while applying advanced data adaptation and prediction accuracy algorithms.

3. Dialogue Summarization Technology

Machine learning technologies supported by natural language processing have brought continuous advancement to dialogue summarization systems throughout history.

During the 1950s to 2000s most dialogue summary systems operated as extractive systems through statistical techniques to identify essential dialogue sentences for their summaries. These systems found the most important components of discussions through the use of frequency-based models or basic TF-IDF algorithms. During the early period of text summarization research the main objectives involved identifying crucial topics and restructuring them based on protocol guidelines. [26]

During the 2000s rule-based algorithms along with statistical methods extracted vital text elements from conversations because both natural language processing and computational strength continued to develop. Before colloquial language understanding became prevalent these systems only processed short exchanges that remained formal at the same time.

Systems that generate summaries through original information paraphrasing or rewording entered the scene during the 2010s with abstractive summarization. Attainments in deep learning development amplified abstractive system performance particularly through transformer-based and encoder-decoder implementation modalities. Sequence to Sequence models were produced by the first researchers who achieved significant progress in dialogue summarization systems.

Transformers alongside BERT and GPT and T5 and PEGASUS emerged during the 2020s through Present to enhance dialogue summarization performance. This model architecture enables better contextual understanding for talking summaries as well as logical and effective summary generation. Keyword-aware dialogue summarization enabled systems to grasp topics better through its focus on essential keywords just like the proposed KADS method. [22]

4. Medical Record Summarization Technology

The advancement of artificial intelligence alongside health record digitization formed the basis for medical record summarization progress.

Doctors performed manual summaries of detailed progressive notes alongside diagnoses together with treatment history records before medical record digitization took place (Before 1990s). Manual summary generation took considerable time while exposing itself to human errors.

Throughout the 1990s and 2000s the medical field saw the adoption of early automated tools which helped extract significant healthcare data from electronic health records (EHR) when these records started gaining popularity. The systems operated by established rules which extracted data from structured formats such as patient diagnosis data and demographic information.

Medical records underwent extractive and abstractive model processing in the 2010s through present times by applying natural language processing (NLP) and machine learning to extract important information. These models kept their limitations to extracting only basic informational data. PEGASUS and T5 brought about advanced abstractive summary models that provided medical practitioners with better methods to summarize patient records and clinical notes. These models boosted accuracy numbers through their use of transformers and their extensive pre-trained language models. [11]

The development of clinical natural language processing techniques occurred throughout 2015 up to the present to address the complicated nature of medical language. Medical jargon and domain terms and abbreviations receive better management through clinical BERT together with other purpose-made models trained with healthcare data.

5. Masked Language Models (MLM)

Natural language processing achieved its most important development through the use of Masked Language Models (MLM:

During the period from the 1950s to 2010s N-gram and basic probabilistic machines achieved language modeling through word prediction based on existing words to establish language modeling technologies. Such models lacked proper representation of language intricacies because their operations depended on straightforward systems.

During the 2010s the natural language models accepted Deep Learning and Neural Networks through Recurrent Neural Networks and Long Short-Term Memory networks which enabled them to work with extended context and dependency abilities. [9]

The 2018 Google BERT transformer system revolutionized the entire industry through its adoption of masked language model operations. BERT performs word prediction by first obstructing some words before analyzing phrase context to achieve more accurate predictions. The development of BERT transformers led to the creation of RoBERTa and ALBERT which resulted in improved specific language understanding.

Descriptive Masked Language Models have appeared in the 2020s to handle the growing need for semantic understanding through integration of word understanding with meaning data within predictive algorithms. These advanced models gained the capability to perform semantic role labeling procedures that depend on semantic foundation. [1][2]

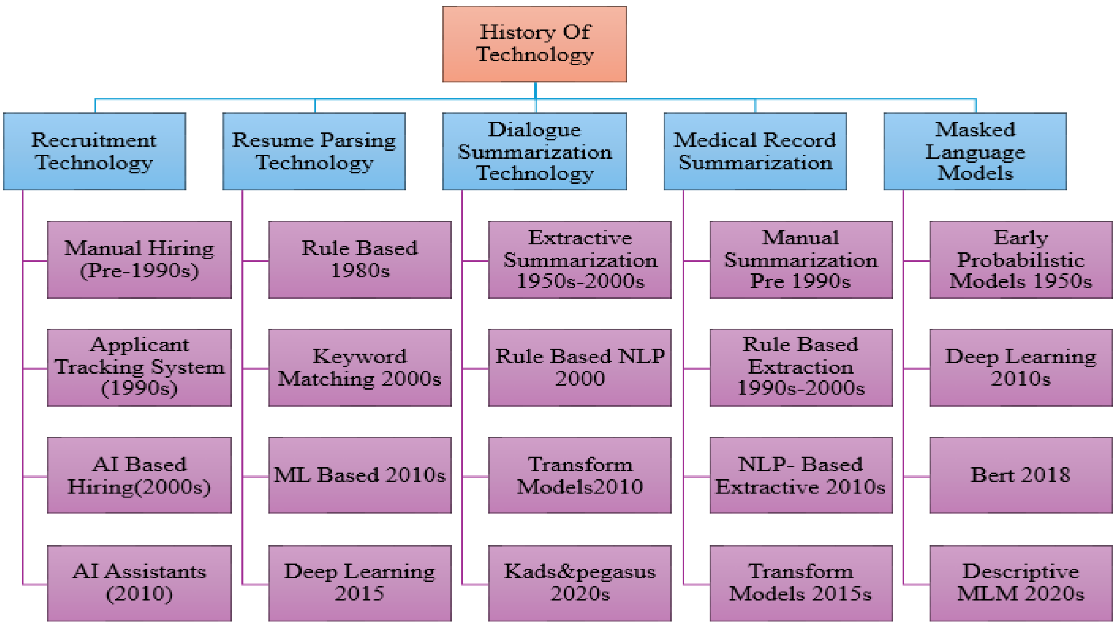

Figure 2: Flowchart of History of Technology

The evolution of technology in various fields demonstrates how advancements have consistently improved the automation and effectiveness of procedures such as medical record summarization, dialogue summarization, resume parsing, and hiring. Technology has been essential in changing the way these jobs are performed, from the early manual approaches and rule-based systems to the emergence of machine learning, deep learning, and transformer models. Accuracy, adaptability, and scalability have significantly increased with the advent of AI-powered systems, such as machine learning algorithms, natural language processing models, and domain-specific solutions, setting the stage for future developments.

Table 2: Summarizing Table of History of Technology

Category	Evolution Summary
Recruitment Technology	Began with manual hiring → ATS automation (1990s) → AI-based hiring platforms (2000s) → AI-driven recruitment assistants (2010s-Present).
Resume Parsing Technology	Started with rule-based parsing (1980s) → Keyword-based (2000s) → ML-based parsers (2010s) → Deep learning & ensemble classifiers (2015-Present).
Dialogue Summarization Technology	Initially extractive (1950s-2000s) → Rule-based NLP (2000s) → Abstractive summarization & transformer models (2010s) → KADS & PEGASUS (2020s-Present).
Medical Record Summarization	Manual summarization (pre-1990s) → Rule-based extractive methods (1990s-2000s) → NLP-driven extractive models (2010s) → Transformer models like PEGASUS (2015-Present).
Masked Language Models (MLM)	Early probabilistic models (1950s-2010s) → Deep learning (2010s) → BERT revolutionized NLP (2018) → Descriptive MLMs (2020s).

D. Limitations and challenges

Different constraints such as technological needs and operational and data-related restrictions affect the limitations and challenges of these targeted technologies which include medical record summarization and resume parsing as well as dialogue summarization, recruitment, and masked language models. This section explores the primary limitations along with difficulties which occur in these technological areas:

1. Recruitment Technology

AI-based hiring tools often develop biases from biased information present in their training datasets. Past hiring data containing biases can reinforce them during employee recruitment because they appear in available information. When recruiting practices become unfair and workforce diversity drops this is anticipated to occur.

The system of unstructured data including non-standardized resumes as well as casual language or unique qualifications creates problems for ATS but structured data including job titles and dates can be handled effectively. Despite efforts data extraction from this type of information remains a difficult task. [1]

Many machine learning models employed for hiring applications including candidate rating systems and resume filter models function as unexplainable systems that limit recruiters' understanding of selection procedures. The inability to interpret decision-making processes from automated technologies generates concerns especially during job-related decisions.

The recruiting platforms demonstrate solid performance for businesses with medium size and smaller yet they may struggle to meet the processing requirements of large corporate hiring and handling thousands of candidate applications. The systems encounter major challenges because of the complicated process of linking them with existing HR information systems and technology platforms. [4]

2. Resume Parsing Technology

The built-in systems aiming to pull useful information from resumes contain multiple challenging weaknesses.

The large obstacle comes from irregularities between resumes which use different formatting standards and language styles. Resumes with non-traditional typefaces and photos along with tables or unusual layouts create obstacles for parsing algorithms to extract data properly thus leading to improper data processing.

Ambiguous terms together with synonyms present an issue because resume parsing systems may detect crucial credentials differently from how human reviewers would see them. The parsing system would fail to link "software developer" and "programmer" roles together when screening candidates.

Resume parsers often fail to understand whole contexts of item references which causes them to extract isolated data points like educational background and skills and work experience details. The absence of contextual understanding by the system causes candidates to be grouped under wrong categories and profile contents fail to be completed properly.

Sensitive information contained in resumes presents major privacy and security risks during the handling process through which personal details such as names and contact information remain unprotected. The compliance challenge for GDPR-like data protection standards remains unresolved for organizations even today [5].

3. Dialogue Summarization Technology

The process of dialogue summarization faces multiple limitations and challenges mainly through the usage of abstractive systems.

A system must comprehend colloquial language and slang because informal dialogue uses both informal linguistic expressions and unconventional grammar rules. Traditional summarizing algorithms with advanced capabilities face difficulties in processing casual communication due to which they generate inadequate or mistaken summaries.

The change of subject matter and broken dialogues by multiple speakers creates challenges for dialogue summarization systems to maintain proper context coherence. Summary content can become impractical because it fails to connect important details and divides the presented information.

During multi-speaker dialogue handling the process becomes especially complex because the system needs to understand conversation flow patterns to properly attribute speech to each participant. When information is incorrectly attributed in the summary process it creates illogical or misleading content.

Huge dialogue pieces prove hard for contemporary dialogue summarization algorithms to process because their maximum input capacity limits their performance. During summarization many crucial information tends to disappear because of poor connecting between discussion parts. [17]

4. Medical Record Summarization Technology

Medical record summarization continues to advance but multiple important hurdles continue to exist.

The medical data is subject to strict privacy rules because HIPAA standards establish stringent requirements for patient records which maintain their confidential nature. The verification of rule compliance becomes difficult during AI summary processing of patient records.

Multiple health institutions maintain patient medical records across separate database and information locations. The lack of complete or consistent data elements specifically regarding challenged medical diagnosis information or insufficient patient records affects summary accuracy and summary creation proves difficult to execute.

Several medical NLP systems face challenges when processing surgical medical terminology because of its specialized professional terminology. Clinical BERT was created to address medical vocabulary complexity but did not resolve the extensive amount of clinical terminology that makes parse and summary operations difficult.

General NLP models perform better today yet their inconsistency in medical domain processing stems from medical terms that are too specialized. The development of specific medical domain models faces difficulties when optimizing precise medical terminology and clinical situation analysis. [6]

5. Masked Language Models (MLM)

The employment industry in NLP has undergone major changes because of Masked Language Models starting with BERT and its multiple versions yet their deployment faces significant obstacles:

The training process of multilevel marketing algorithms depends on extensive data mining of content which often includes harmful or prejudicial or otherwise objectionable content from the Internet. The model receives biases from multiple sources that cause the generation of problematic outcomes particularly in medical care and recruitment systems.

Big MLM training involves expensive computer resources including large datasets along with GPUs for high-performance operation. Malevolent impact on smaller firms prevents them from training such models or to effectively implement them.

The syntactic abilities of BERT-type MLMs are strong yet they frequently lack the capability to understand semantic meaning explicitly. The structure of sentences becomes clear enough for models but they cannot completely interpret word context properly leading to errors on complex comprehension tasks.

The size of BERT does not protect the model from Out-of-Vocabulary (OOV) issues because some words or specialized phrases from their training data remain undetected during processing. The incorrect or incomplete processing of the input text occurs because of this issue.

Despite using black box mechanisms many multilevel marketing enterprises maintain high levels of decision-landscape obscurity for their users. The medical field needs full explainability since a lack of it brings criticism to summary processes. [14]

6. General Challenges Across All Systems

These technologies face generic limitations including the ones outlined before domain-specific challenges.

High-quality labelled data has to be available whenever machine learning model training occurs. Effective model training becomes complex since numerous domains (such as medical records and recruitment) consistently present their information as inconsistent or faulty or incomplete data. The lack of accessibility to extensive labelled datasets becomes an issue because of resource limitations together with privacy restrictions.

AI systems managing increasing sensitive data create new privacy and ethical issues that become major concerns because of their FormatException: Failed to fetch. The response is missing necessary information. functions to process sensitive information. AI tools that screen medical records together with resumes have the potential to violate patient privacy through unintentional misuse of personal data.

The integration of AI-driven technology into standard operational systems which mainly consist of medical record platforms and communication channels and traditional HR tools remains a hard challenge. The lack of integration between different systems produces barriers for the organizations to effectively adopt novel technological solutions.

AI systems face obstacles when trying to adapt their operations to changing environments that involve modifications in user conduct and market styles and legal framework adjustments. The dependence on historical data in these systems leads to eventual obsolescence when they lack regular update or retraining procedures that diminishes system performance effectiveness. [13]

The combined effects of data weaknesses and model prejudice along with computational processing challenges create substantial barriers which technologies encounter while performing in hiring systems and resume sorting as well as dialogue expansion and medical case note extraction and masked language modelling operations. The obstacles in front of these systems need better data management practices and refined model development alongside a moral approach to open AI applications. The ongoing research within deep learning together with machine learning and natural language processing indicates that numerous present restrictions will be solved to generate more dependable and efficient systems.

Table 3: Limitations of Different Systems

Category	Key Limitations
Recruitment Technology	Bias in AI models, difficulty handling unstructured data, lack of transparency, scalability challenges.
Resume Parsing Technology	Struggles with inconsistent resume formats, synonym/ambiguity issues, limited contextual understanding, privacy concerns.
Dialogue Summarization Technology	Challenges in understanding slang, maintaining context, handling multi-speaker dialogues, and processing long conversations.
Medical Record Summarization	Data privacy & compliance issues, fragmented medical records, difficulty handling medical jargon, lack of domain-specific models.
Masked Language Models (MLM)	Training data biases, high computational cost, limited semantic understanding, out-of-vocabulary (OOV) problems, lack of transparency.
General AI Challenges	Data quality issues, ethical concerns, integration with existing systems, adaptability to new trends, computational inefficiencies.

III. Outcomes

Several important conclusions have been drawn from the examination of several NLP and machine learning models used in resume screening:

1. Efficiency of Automated Resume Screening

Machine learning tools namely KNN and SVM significantly boost the capability of resume filtering systems.

The accuracy of candidate ranking gets upgraded by both ensemble classifiers and cosine similarity methods. [1]

2. Advancements in Resume Parsing

Named entity recognition benefits from better results produced by BLSTM-CNNs-CRF along with other neural network-based parsing models.

Rule-based parsers with conventional approaches need enhancement of semantic understanding since they struggle to process resumes without structured formats.

3. Challenges in Recruitment Automation

AI-based hiring systems encounter three main difficulties that consist of bias issues alongside interpretability problems alongside incompatibility challenges when connecting to existing applicant tracking systems.

The installation of automated candidate selection systems encounters ongoing challenges because of privacy issues surrounding data protection.

4. Potential of Dialogue Summarization in Recruitment

The method of dialogue summarization through PEGASUS, BART and T5 models demonstrates favourable results in HR interview settings.

The hiring process through chatbots becomes stronger with keyword-aware summarization (KADS). [4]

5. Application in Other Domains

The flexibility of abstractive models for medical record summarization needs specific domain-related modifications.

Descriptive masked language models (DMLM) help users understand semantics better than typical MLMs. [6]

IV. Conclusion Table

This is a well-organized synopsis of the key findings from the study.:

Aspect	Findings
Recruitment Systems	The increased productivity from resume screening with Artificial Intelligence comes with problems related to bias as well as transparency.
Resume Parsing	In the case of structured data extraction, deep learning models do perform better than conventional parsers.
Dialogue Summarization	Chatbot-based hiring and interview analysis can be improved by using NLP models such as PEGASUS and KADS.
Medical Record Summarization	Transformer-based models present accurate summarization even though they need to be trained specifically for a particular domain.
Masked Language Models	Although they need a lot of training data, descriptive MLMs do increase comprehension.
Challenges	The technology reaches its limit due to three main operational challenges: data privacy issues, scalability requirements and interpretability difficulties.

References

[1]. Kinge, B., Mandhare, S., Chavan, P., & Chaware, S. M. (2022). Resume Screening using Machine Learning and NLP: A proposed system. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, *8*(2), 253-258.

[2]. Tejaswini, K., Umadevi, V., Kadiwal, S. M., & Revanna, S. (2022). Design and development of machine learning based resume ranking system. *Global Transitions Proceedings*, *3*(2), 371-375.

[3]. Kolhatkar, G., Paranjape, A., Gokhale, O., & Kadam, D. (2023, July). Team Converge at ProbSum 2023: Abstractive text summarization of patient progress notes. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks* (pp. 510-515).

[4]. Barba, E., Campolungo, N., & Navigli, R. (2023, July). DMLM: Descriptive Masked Language Modeling. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 12770-12788).

[5]. Yoo, C., & Lee, H. (2023). Improving Abstractive Dialogue Summarization Using Keyword Extraction. *Applied Sciences*, *13*(17), 9771.

[6]. Gopalakrishna, S. T., & Vijayaraghavan, V. (2019). Automated tool for Resume classification using Sementic analysis. *International Journal of Artificial Intelligence and Applications (IJAIA)*, *10*(1).

[7]. Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020). A Machine Learning approach for automation of Resume Recommendation system. *Procedia Computer Science*, *167*, 2318-2327.

[8]. Zu, S., & Wang, X. (2019). Resume information extraction with a novel text block segmentation algorithm. *Int J Nat Lang Comput*, *8*(2019), 29-48.

[9]. Mukherjee, A. (2024, February). Resume Ranking and Shortlisting with DistilBERT and XLM. In 2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE) (pp. 301-304). IEEE.

[10]. Arokodare, O., Wimmer, H., & Du, J. (2024). Clinical Text Summarization using NLP Pretrained Language Models: A Case Study of MIMIC-IV-Notes. In Proceedings of the ISCAP Conference ISSN (Vol. 2473, p. 4901).

[11]. Süsstrunk, N., & Weichselbraun, A. (2024, June). Harnessing LLM's for generating Patient Discharge Reports. In Proceedings of the 9th edition of the Swiss Text Analytics Conference (pp. 173-173).

[12]. Barba, E., Campolungo, N., & Navigli, R. (2023, July). DMLM: Descriptive Masked Language Modeling. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 12770-12788).

[13]. Yoo, C., & Lee, H. (2023). Improving Abstractive Dialogue Summarization Using Keyword Extraction. Applied Sciences, 13(17), 9771.

[14]. Varadharajan, S. T. G. V. ENSEMBLE LEARNING BASED VOTING MODEL FOR DYNAMIC PROFILE CLASSIFICATION AND PROJECT ALLOTMENT.

[15]. Kabir, E., Bakos, J. D., Andrews, D., & Huang, M. (2024, November). ProTEA: Programmable Transformer Encoder Acceleration on FPGA. In SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 521-530). IEEE.

[16]. Navarro, G. (2025). Fair and Ethical Resume Screening: Enhancing ATS with JustScreen the ResumeScreeningApp. Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2(1), 1-7.

[17]. Thangaramya, K., Logeswari, G., Gajendran, S., Roselind, J. D., & Ahirwar, N. (2024, May). Automated Resume Parsing and Ranking using Natural Language Processing. In 2024 3rd International Conference on Artificial Intelligence for Internet of Things (AIIoT) (pp. 1-6). IEEE.

[18]. Joseph Sirrianni, P. H. D. Evaluation of a Digital Scribe: Conversation Summarization for Emergency Department Consultation Calls.

[19]. Arokodare, O., Wimmer, H., & Du, J. (2024). Clinical Text Summarization using NLP Pretrained Language Models: A Case Study of MIMIC-IV-Notes. In Proceedings of the ISCAP Conference ISSN (Vol. 2473, p. 4901).

[20]. Su, Z., Li, J., Zhang, Z., Zhou, Z., & Zhang, M. (2023, December). Efficient continue training of temporal language model with structural information. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 6315-6329).
[21]. Shoushtari, F., Daghighi, A., & Ghafourian, E. (2024). Application of Artificial Intelligence in Project Management. International journal of industrial engineering and operational research, 6(2), 49-63.
[22]. Digha, A. F., Obasi, C. M. E., & Ajao, W. B. Analysing Whatsapp Group Chat Using Advanced Natural Language Processing (NLP) Techniques.
[23]. Oprea, S. V., & Bâra, A. (2025). Is Artificial Intelligence a Game-Changer in Steering E-Business into the Future? Uncovering Latent Topics with Probabilistic Generative Models. Journal of Theoretical and Applied Electronic Commerce Research, 20(1), 16.
[24]. Hochmair, H. H., Juhász, L., & Li, H. (2025). Advancing AI-Driven Geospatial Analysis and Data Generation: Methods, Applications and Future Directions. ISPRS International Journal of Geo-Information, 14(2), 56.
[25]. WAŻNA, E. (2024). THE ROLE OF ARTIFICIAL INTELLIGENCE IN THE EMPLOYEE RECRUITMENT PROCESS. Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie, (208).
[26]. Balcerak, A., & Woźniak, J. (2023). ICT-BASED RECRUITMENT AND SELECTION TOOLS: THE RECRUITERS'PERSPECTIVE. Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie, (178).
[27]. Sloane, M. (2023, May). Automation and Recruiting: Understanding the Intersection of Algorithmic Systems and Professional Discretion in the Sourcing of Job Candidates. In Proceedings" Automation by Design" Conference, University of Minnesota, Forthcoming.
[28]. Hovland, I. V. (2021). Artificial Intelligence for innovating recruitment and selection processes: evidence from Scandinavian companies (Master's thesis, NTNU).
[29]. Gulati, V., Gupta, I., Firdous, F., & Narwal, R. (2024). Resume Analyzer Using Natural Language Processing (NLP).
[30]. Zu, S., & Wang, X. (2019). Resume information extraction with a novel text block segmentation algorithm. Int J Nat Lang Comput, 8(2019), 29-48.
[31]. Ch, S., & Thenmozhi, M. RESUME ANALYSIS USING DEEP LEARNING. Artificial Intelligence Language Learning And Communication: Exploring The Intersection Of Technology And Education, 106.
[32]. Bhoir, N., Jakate, M., Lavangare, S., Das, A., & Kolhe, S. (2023). Resume Parser using hybrid approach to enhance the efficiency of Automated Recruitment Processes. Authorea Preprints.
[33]. Sajid, H., Kanwal, J., Bhatti, S. U. R., Qureshi, S. A., Basharat, A., Hussain, S., & Khan, K. U. (2022, January). Resume parsing framework for e-recruitment. In 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-8). IEEE.
[34]. Yoo, C., & Lee, H. (2023). Improving Abstractive Dialogue Summarization Using Keyword Extraction. Applied Sciences, 13(17), 9771.
[35]. Mishra, R., Bian, J., Fiszman, M., Weir, C. R., Jonnalagadda, S., Mostafa, J., & Del Fiol, G. (2014). Text summarization in the biomedical domain: a systematic review of recent research. Journal of biomedical informatics, 52, 457-467.
[36]. Patel, R., Brayne, A., Hintzen, R., Jaroslawicz, D., Neculae, G., & Corneil, D. (2024). Retrieve to Explain: Evidence-driven Predictions with Language Models. arXiv preprint arXiv:2402.04068.
[37]. Jegede, O. Enhancing Healthcare Delivery: Process Improvement via Machine Learning-Driven Predictive Project Management Techniques.
[38]. Wahid, J. A., Shi, L., Gao, Y., Yang, B., Wei, L., Tao, Y., ... & Yagoub, I. (2022). Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response. Expert Systems with Applications, 195, 116562.
[39]. Shukla, C., Goel, G., Yadav, S., & Gaur, A. (2024). Comparative Analysis of Machine Learning Model Performance Under Identical Condition. Available at SSRN 5035925.