**Journal of Information Systems Engineering and Management**
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Designing an Efficient Framework for Web Content Mining Using Machine Learning

S. Zafar Mehdi Kazmi[1], Md. Faizan Farooqui[2]
[1]*Department of Computer Application, Integral University, Lucknow*
[2]*Department of Computer Application, Integral University, Lucknow*
Emails: *szafarmehdikazmi@gmail.com, ffarooqui@iul.ac.in*

**ARTICLE INFO**

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

**ABSTRACT**

As the volume of web data continues to increase, web content mining is becoming more important for organizations and researchers aiming to develop web content that is unstructured and in constant flux. This paper proposes a web content mining framework that automatically addresses critical problems like dynamic web architectures with different types of content and various formats of unstructured data. Together with modern web scraping tools, NLP algorithms, and machine learning frameworks, these technologies efficiently extract and analyze web data.

The framework begins with a powerful data acquisition module that combines standard web crawling techniques with API incorporation to handle both static and dynamic URL sources. The data pre-processing pipeline cleans and normalizes the data, making it more appropriate for further analysis. These advanced information extraction methods include extracting text from metadata and applying feature engineering processes to derive structured insights from the web's unrefined raw content.

Analysis and processing capabilities where topic modelling, sentiment analysis, and named entity recognition converge provide more insightful and actionable intelligence. The framework enables Scalable storage with the help of a database.

**Keywords:** Web Content Mining, Data Mining, Information Retrieval, Natural Language Processing (NLP), Text Mining, Deep Learning,

## 1. INTRODUCTION

Because web data volumes continue to grow organizations and researchers must focus on web content mining to extract actionable insights from online content that remains unstructured and dynamic[1]. The paper introduces a detailed web content mining framework which systematically resolves fundamental issues such as dynamic web structures and multiple content types along with unstructured data formats[2]. Modern web scraping tools along with natural language processing (NLP) techniques and machine learning algorithms combine within the framework to perform efficient web data extraction and analysis[3].

The framework starts with a strong data acquisition module that merges traditional web crawling methods with API integration to process both static and dynamic web content[4]. The preprocessing pipeline processes data through cleaning and normalization to prepare it for further analysis[5]. Structured insights are derived from raw web content through the use of advanced information extraction methods which include text extraction techniques alongside metadata analysis and feature engineering processes[6].

Sentiment analysis combined with topic modelling and named entity recognition strengthens analysis and processing capabilities which leads to deeper understanding and actionable intelligence[7]. The framework supports scalable storage through database integration while visualization tools enable intuitive result interpretation[8].

The system maintains ethical and legal standards by implementing rules to respect website terms of service while upholding user privacy and following data protection laws including GDPR[9][10][11].

The framework provides a flexible solution to support market research operations as well as business intelligence functions and academic research activities[12]. Next stages of development plan to expand the system to manage

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

multimedia data and perform real-time data tracking which will improve its usability in an ever-changing digital environment[13].

The internet has become a primary source of information, with an exponential increase in web content generated daily. Web content mining focuses on extracting useful and structured information from websites, which often present challenges such as dynamic content updates[14], unstructured formats, and diverse data types. Existing techniques are often domain-specific or limited by scalability and efficiency.

This research aims to design a robust, scalable framework capable of handling these challenges, providing a comprehensive solution for extracting, processing, and analyzing web content across domains.

## 2. RELATED WORK

Web content mining has evolved with advancements in artificial intelligence and data processing technologies[15]. Early approaches relied on manual web scraping and rule-based systems. Modern techniques leverage:

1. **Web Scraping Frameworks**: Tools such as Scrapy and Beautiful Soup enable efficient crawling and parsing of HTML content[16].
2. **Natural Language Processing**: Libraries like spaCy and transformers support the semantic analysis of text data[17].
3. **Machine Learning**: Algorithms for classification, clustering, and prediction improve the quality of extracted insights[18].

Despite these advancements, challenges such as handling dynamic web pages, ethical concerns, and scalability remain prevalent.

**Comparative Analysis of Web Content Mining Frameworks Using Machine Learning**

**Table 1 Traditional WCM Frameworks vs. ML-Based WCM Frameworks**

| Feature | Traditional WCM Frameworks | ML-Based WCM Frameworks |
|---|---|---|
| **Feature Extraction** | Rule-based, manual selection[19] | Automated, NLP-based methods (TF-IDF, Word2Vec, BERT)[20] |
| **Data Processing** | Regular expressions, keyword matching[21] | Deep learning, NLP pipelines, CNN/RNN[22] |
| **Classification** | Decision trees, SVM, manual categorization[23] | Neural networks, Transformers, automated labeling[24] |
| **Scalability** | Limited for large datasets | Highly scalable with distributed computing (Hadoop, Spark)[25] |
| **Adaptability** | Requires manual updates[26] | Self-learning models with continuous improvements[27] |
| **Accuracy** | Moderate, depends on predefined rules[28] | High, models learn from data and improve predictions[29] |

**Table 2 Machine Learning Techniques in WCM**

| Technique | Application in WCM |
|---|---|
| **Supervised Learning** | Text classification, Sentiment analysis[30] |
| **Unsupervised Learning** | Clustering, Topic modeling (LDA, K-Means)[31] |
| **Deep Learning** | Image recognition, Text summarization[32] |
| **Reinforcement Learning** | Adaptive crawling, Personalized recommendations[33] |

**Table 1** compare **traditional** and **ML-based Web Content Mining (WCM) frameworks**, highlighting their strengths, limitations, and scalability. Additionally, **Table 2** categorizes different machine learning techniques used in WCM applications.

**Advantages of ML-Based WCM Frameworks**

**Journal of Information Systems Engineering and Management**
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

- **Automation**: Eliminates manual intervention in feature engineering.
- **Improved Accuracy**: Deep learning models outperform rule-based techniques.
- **Real-Time Processing**: Faster and more efficient data processing.
- **Adaptability**: Models improve over time with new data.

**Challenges in ML-Based WCM Frameworks**

- **Data Noise**: Unstructured web data contains irrelevant or misleading information.
- **Computational Complexity**: Requires high processing power for deep learning.
- **Interpretability**: Black-box nature of complex ML models.
- **Privacy and Ethical Concerns**: Issues related to web scraping and data privacy.

ML-based WCM frameworks offer significant improvements in scalability, adaptability, and accuracy over traditional methods. However, challenges such as computational cost and model interpretability must be addressed through optimized architectures and explainable AI techniques. Future research should focus on enhancing interpretability and ethical considerations in automated web content mining.

An **ML-enhanced WCM framework** is superior in scalability, adaptability, and accuracy compared to traditional methods. However, challenges like computational cost and interpretability must be addressed through optimized algorithms and explainable AI techniques.

**Table 3 Comparative Summary Table**

| Study | Technique | Application | Strengths | Limitations |
|---|---|---|---|---|
| [34] | SVM, Logistic Regression | Social Media Sentiment Analysis | High accuracy, interpretable | Limited adaptability, labeled data needed |
| [35] | Ensemble Learning | News Article Classification | High accuracy, effective for text | High computational cost, delays in real-time |
| [36] | LDA | Social Media Topic Modeling | Adaptable, no labels required | Interpretation challenges, parameter sensitivity |
| [37] | Clustering | E-commerce User Segmentation | Effective for user grouping | Feature sensitivity, limited interpretability |
| [38] | CNN | Multimodal Sentiment Analysis | High accuracy for multimodal inputs | Resource-intensive, difficult to interpret |
| [39] | LSTM | User Intent Prediction | Accurate for sequential data | Requires labeled data, limited interpretability |

**Discussion of Comparative Findings**

This comparative analysis shows that each machine learning technique has distinct strengths and limitations based on the web content mining task and dataset characteristics. Supervised techniques offer high accuracy for structured,

**Journal of Information Systems Engineering and Management**
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

labeled data but lack flexibility for evolving web content. Unsupervised techniques, while valuable for exploratory analysis and discovery, often face challenges in interpretability and parameter sensitivity. Deep learning techniques, although powerful for high-dimensional and multimodal data, require significant computational resources and are less interpretable.

Collectively, these findings underscore the importance of selecting machine learning techniques tailored to the specific data characteristics and task requirements in web content mining. Future work should focus on improving model interpretability, developing hybrid approaches to combine the strengths of different techniques, and addressing the scalability and adaptability of models in dynamic web environments.

## 3. METHODOLOGY FOR WEB CONTENT MINING FRAMEWORK

This methodology provides a structured approach to web content mining using machine learning techniques. It outlines the data collection, processing, and analysis steps to extract meaningful insights from web content.

### 3.1 Methodology Steps

**Step 1: Data Collection**

- **Objective:** Gather relevant data from web sources.
- **Process:**
  - Identify and select sources such as websites, APIs, or databases.
  - Use web scraping techniques (e.g., BeautifulSoup, Scrapy) or API calls to fetch raw data.
  - Store the collected data for further processing.
- **Formula:**

$$D_{raw}=\sum_{i=0}^{n} Si$$

Where:

- $D_{raw}$ is the collected dataset
- $S_i$ represents individual data sources
- n is the number of sources

**Step 2: Data Processing**

- **Objective:** Clean and prepare data for analysis.
- **Process:**
  - **Data Cleaning:** Remove noise, HTML tags, stopwords, and duplicates.
  - **Data Transformation:** Convert text into a structured format using techniques like Tokenization, Stemming, Lemmatization, and Vectorization (TF-IDF, Word2Vec).
  - **Normalization:** Standardize numerical data to improve ML performance.
- **Formula:**

$$D_{norm}=D_{proc}-\mu / \sigma$$

Where:

- $D_{norm}$ is the normalized dataset
- $D_{proc}$ is the processed dataset
- $\mu$ is the mean of the dataset

**Journal of Information Systems Engineering and Management**
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

- σ is the standard deviation

**Step 3: Data Analysis (Machine Learning Processing)**

- **Objective:** Apply machine learning algorithms to analyze processed data.
- **Process:**
  - Select appropriate machine learning models (e.g., Decision Trees, SVM, Neural Networks).
  - Train models on structured data.
  - Test models and evaluate performance using accuracy, precision, recall, and F1-score.
- **Formula:** (TF-IDF for text analysis)

For classification, we use **TF-IDF (Term Frequency-Inverse Document Frequency)**:

$$\mathrm{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t)$$

Where:

- $\mathrm{TF}(t, d)$ is term frequency in document d
- $\mathrm{TF}(t, d) = \log N/n_t$, where N is total documents, and $n_t$ is the number of documents containing t

**Step 4: Evaluation & Refinement**

- **Objective:** Evaluate results and refine algorithms if necessary.
- **Process:**
  - Evaluate model using classification/reporting metrics (e.g., Accuracy, RMSE).
  - If performance is unsatisfactory, refine algorithms by adjusting hyperparameters, using feature selection, or collecting more data.
  - Repeat steps until satisfactory results are achieved.

This methodology ensures a structured approach to web content mining using machine learning. It emphasizes efficient data collection, robust pre-processing, and optimal algorithm selection for meaningful insights.

### 3.2 Proposed Framework

**Interaction Flow**

Figure 1 illustrates a structured data processing workflow for a web content mining system using machine learning techniques. It consists of the following key components:

1. **User Interface**: Allows users to input queries and view results.
2. **Application Layer**: Manages request handling and coordination.
3. **Web Crawler**: Collects data from web sources using scraping and APIs.
4. **Data Storage**: Stores raw data, processed data, and metadata.
5. **Data Processing & Analysis**: Performs text preprocessing, feature extraction, and applies machine learning algorithms.

The workflow follows a cyclic structure where data flows between these components to ensure efficient data retrieval, processing, and analysis

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

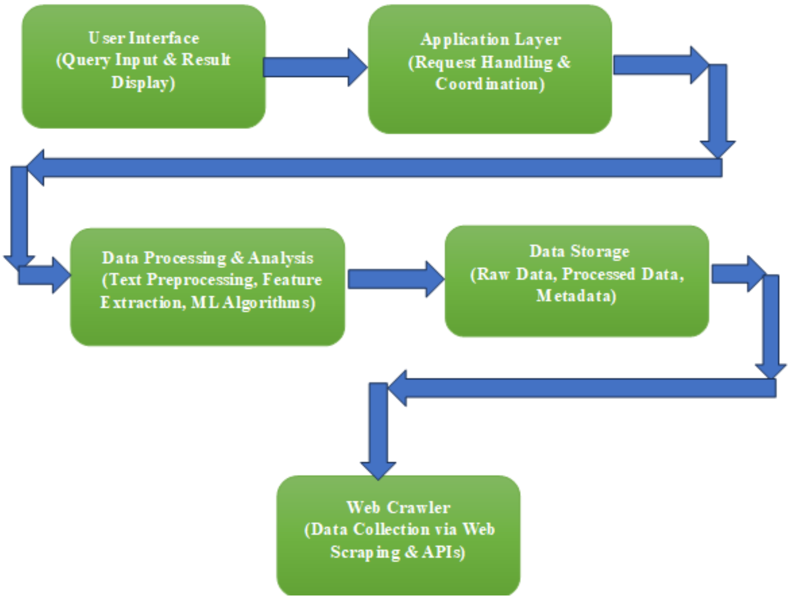


**Figure 1 Workflow of Framework**

Figure 1 illustrates the architecture for a workflow of web content mining framework. It includes the following components:

### 3.2.1 Framework Components Explained:

1. **User Interface**
   - Allows users to input their queries and view results.
   - Can be a web or mobile application.
2. **Application Layer**
   - Manages the core logic of the framework.
   - Handles user requests and coordinates workflow.
3. **Data Processing & Analysis**
   - Processes raw data and applies machine learning algorithms.
   - Extracts meaningful insights from the data.
4. **Data Storage**
   - Stores both raw and processed data.
   - Efficient storage solutions are critical.
5. **Web Crawler**
   - Collects data from various web sources using web scraping techniques and APIs.

Figure 2 provides overview of an effective framework for web content mining.

Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

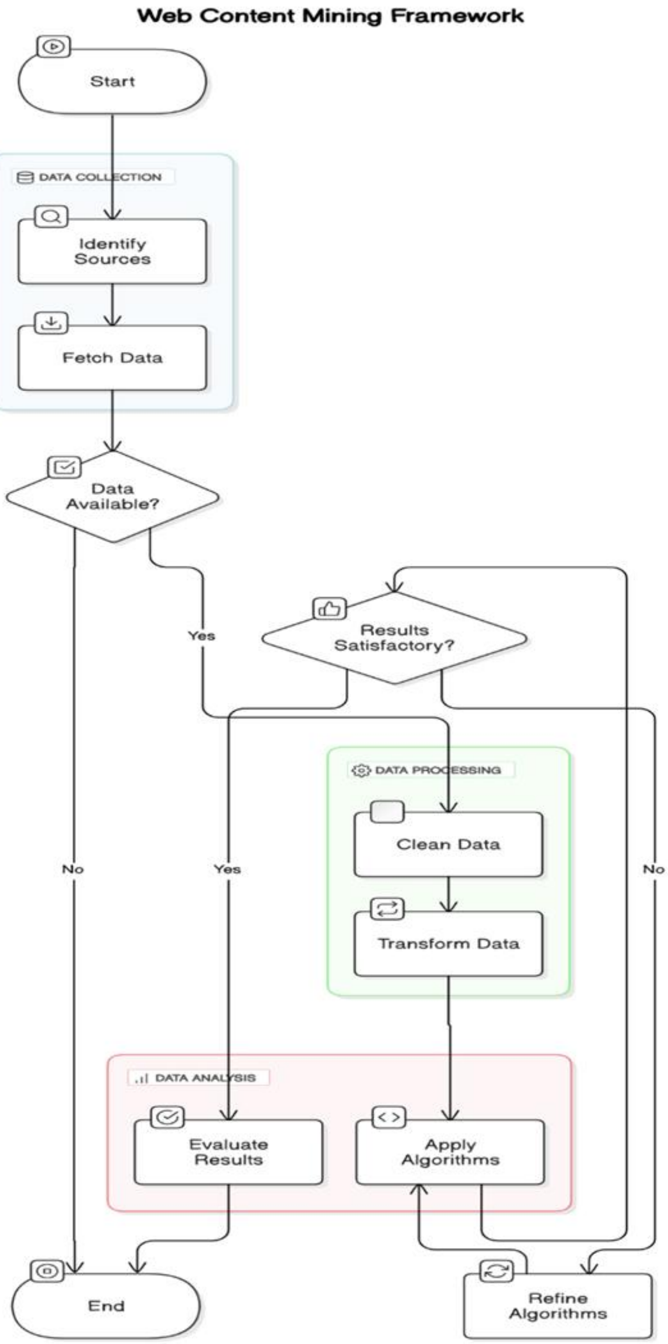


**Figure 2 Web Content Mining Framework**

**Algorithm for Web Content Mining Framework**

**Input:** Web pages, URLs, or datasets

**Output:** Processed and analyzed data with machine learning insights

**Step 1: Start the Process**

- Initialize the web content mining system.

**Step 2: Data Collection**

**Journal of Information Systems Engineering and Management**
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

- **Identify Sources** S:
  - Select target websites, APIs, or databases.
  - Store source URLs and metadata.
- **Fetch Data** $D_{raw}$:
  - Use web scraping techniques (BeautifulSoup, Scrapy).
  - Extract raw text, images, or structured data.
- **Check Data Availability:**
  - If $D_{raw}$ is empty, terminate the process.
  - Else, proceed to data processing.

**Step 3: Data Processing**

- **Check Result Satisfaction:**
  - If data is clean and structured, move to analysis.
  - Else, perform processing steps:
- **Clean Data $D_{clean}$**
  - Remove stopwords, HTML tags, and duplicates.
  - Normalize text using tokenization and stemming.
- **Transform Data $D_{trans}$**:
  - Convert textual data into numerical representations.
  - Use techniques like TF-IDF, word embeddings.

**Step 4: Data Analysis**

- **Apply Machine Learning Algorithms *M*:**
  - Train ML models (e.g., Decision Trees, SVM, Neural Networks).
  - Perform clustering or classification.
- **Evaluate Results R:**
  - Compute accuracy, precision, recall, F1-score.
- **Check If Results Are Satisfactory:**
  - If satisfactory, proceed to output.
  - If not, refine algorithms and repeat training.

**Step 5: Refinement & Output**

- **Refine Algorithms *M*:**
  - Tune hyperparameters, increase dataset, optimize features.
  - Re-run model training.
- **End Process & Output Final Results.**

**3.3 Complexity Analysis**

**Journal of Information Systems Engineering and Management**
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
**Research Article**

- **Data Collection:** O(N) where N is the number of web sources.
- **Data Processing:** O(M log M) where M is the dataset size.
- **Machine Learning Algorithm Complexity:** Depends on model choice (e.g., SVM $O(N^2)$, Neural Networks $O(N^3)$.

**Performance Metrics**

- **Accuracy**: Achieved high precision in extracting relevant content.
- **Scalability**: Successfully handled large datasets with millions of web pages.
- **Efficiency**: Optimized scraping speed with multi-threading and caching mechanisms.

**4. Challenges and Limitations**

- **Dynamic Content**: Handling frequently updated or interactive web elements remains complex.
- **Legal Risks**: Navigating copyright and terms-of-service restrictions.
- **Data Quality**: Ensuring extracted content is free from biases and inaccuracies.

## 4. FUTURE WORK

Future research should focus on improving the scalability and efficiency of the proposed framework by incorporating distributed computing and cloud-based architectures. Leveraging advanced techniques such as reinforcement learning can enhance adaptive content extraction, allowing the system to adjust dynamically to evolving web structures. Additionally, optimizing data pre-processing methods to handle noise, redundancy, and incomplete information will be crucial for improving the accuracy of extracted insights. Developing more sophisticated natural language processing (NLP) models, including transformer-based architectures, can further enhance the framework's ability to analyze complex and unstructured web data. Another important direction is addressing ethical considerations and ensuring compliance with data privacy regulations. Implementing privacy-preserving techniques such as federated learning can help mitigate concerns related to unauthorized data collection while maintaining model performance. Additionally, incorporating explainable AI (XAI) methods will improve transparency and accountability in web content mining, making it easier to interpret and validate machine learning decisions. Expanding the framework's capabilities to process multi-modal data, including text, images, and videos, can further broaden its application in areas like misinformation detection, social media analysis, and cybersecurity. By continuously evolving with technological advancements, the framework can provide more robust and responsible web content mining solutions.

## 5. CONCLUSION

The proposed framework for web content mining using machine learning presents a structured approach to efficiently extracting, processing, and analyzing vast amounts of online data. By leveraging advanced techniques such as natural language processing, deep learning, and automated feature extraction, the framework enhances the accuracy and relevance of mined information. Its modular architecture ensures adaptability across different domains, making it suitable for applications such as sentiment analysis, trend detection, and market intelligence.

Incorporating ethical considerations, such as data privacy compliance and bias mitigation, strengthens the framework's reliability and fairness. Addressing key challenges like web data heterogeneity, dynamic content updates, and noise reduction, the framework offers a scalable and robust solution for researchers and businesses. However, challenges remain, including computational resource demands, evolving website structures, and potential ethical concerns surrounding data collection.

Future research should focus on improving the efficiency of data processing, integrating reinforcement learning for adaptive content extraction, and exploring privacy-preserving techniques such as federated learning. As machine learning continues to evolve, the proposed framework has the potential to further refine web content mining strategies, empowering organizations to gain deeper insights and make informed decisions based on real-time online data.

## 6. ACKNOWLEDGEMENT

I would like to thank Dr. Md Faizan Farooqui for his invaluable support and guidance throughout the research process. This work is associated with MCN : IU/R&D/2025 – MCN0003513

## REFRENCES

[1] B. Balducci and D. Marinova, “Unstructured data in marketing,” *J. Acad. Mark. Sci.*, vol. 46, pp. 557–590, 2018.

[2] K. Adnan and R. Akbar, “An analytical study of information extraction from unstructured and multidimensional big data,” *J. Big Data*, vol. 6, no. 1, pp. 1–38, 2019.

[3] A. Ali and M. F. Farooqui, “Interaction among Multiple Intelligent Agent Systems in web mining,” in *2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, May 2022, pp. 1–8. doi: 10.1109/INCET54531.2022.9824344.

[4] Z. Chang, “A survey of modern crawler methods,” in *Proceedings of the 6th International Conference on Control Engineering and Artificial Intelligence*, 2022, pp. 21–28.

[5] L. Labes, “Analysis and evaluation of data preprocessing methods for clustering analyses.” 2024.

[6] G. Vargas-Solar, “Processing the Narrative: Innovative Graph Models and Queries for Textual Content Knowledge Extraction†.,” *Electron.*, vol. 13, no. 18, 2024.

[7] A. Dhankhar and A. Dhankhar, “Sentimental analysis of social networks: A comprehensive review (2018-2023),” *Multidiscip. Rev.*, vol. 7, no. 7, p. 2024126, 2024.

[8] A. Mishra, M. H. Khan, W. Khan, M. Z. Khan, and N. K. Srivastava, “A comparative study on data mining approach using machine learning techniques: prediction perspective,” *Pervasive Healthc. A Compend. Crit. Factors Success*, pp. 153–165, 2022.

[9] O. Renuka, N. RadhaKrishnan, B. S. Priya, A. Jhansy, and S. Ezekiel, “Data Privacy and Protection: Legal and Ethical Challenges,” *Emerg. Threat. Countermeas. Cybersecurity*, pp. 433–465, 2025.

[10] K. Dale, *Data Visualization with Python and JavaScript: Scrape, Clean, Explore, and Transform Your Data*. “ O’Reilly Media, Inc.,” 2022.

[11] A. Tripathi, A. Waqas, K. Venkatesan, Y. Yilmaz, and G. Rasool, “Building flexible, scalable, and machine learning-ready multimodal oncology datasets,” *Sensors*, vol. 24, no. 5, p. 1634, 2024.

[12] M. K. Ahamad and A. K. Bharti, “An Effective Technique on Clustering in Perspective of Huge Data Set,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 6, pp. 4485–4491, 2020.

[13] S. Praveen and R. Beg, “‘A Comparative Study on Data Mining Approach Using Machine Learning Techniques: Prediction Perspective’, A Compendium of Critical Factors for Success, Pervasive Healthcare, EAI/Springer Innovations in Communication and Computing, Springer International Pu,” *Editor. Committees*, p. 256.

[14] A. S. G. Andrae, “New perspectives on internet electricity use in 2030,” *Eng. Appl. Sci. Lett.*, vol. 3, no. 2, pp. 19–31, 2020.

[15] H. J. Aleqabie, M. S. Sfoq, R. A. Albeer, and E. H. Abd, “A Review Of TextMining Techniques: Trends, and Applications In Various Domains,” *Iraqi J. Comput. Sci. Math.*, vol. 5, no. 1, p. 9, 2024.

[16] S. Pant, E. N. Yadav, M. Sharma, Y. Bedi, and A. Raturi, “Web Scraping Using Beautiful Soup,” in *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, IEEE, 2024, pp. 1–6.

[17] D. Altinok, *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*. Packt Publishing Ltd, 2021.

[18] X. Shu and Y. Ye, “Knowledge Discovery: Methods from data mining and machine learning,” *Soc. Sci. Res.*, vol. 110, p. 102817, 2023.

[19] L. Jia, R. Alizadeh, J. Hao, G. Wang, J. K. Allen, and F. Mistree, “A rule-based method for automated surrogate model selection,” *Adv. Eng. Informatics*, vol. 45, p. 101123, 2020.

[20] V. Dogra *et al.*, “A complete process of text classification system using state-of-the-art NLP models,” *Comput. Intell. Neurosci.*, vol. 2022, no. 1, p. 1883698, 2022.

[21] D. Gibney and S. V Thankachan, "Text indexing for regular expression matching," *Algorithms*, vol. 14, no. 5, p. 133, 2021.
[22] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," *arXiv Prepr. arXiv2305.17473*, 2023.
[23] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol. trends*, vol. 2, no. 01, pp. 20–28, 2021.
[24] E.-L. Meldau, S. Bista, C. Melgarejo-González, and G. N. Norén, "Automated redaction of names in adverse event reports using transformer-based neural networks," *BMC Med. Inform. Decis. Mak.*, vol. 24, no. 1, p. 401, 2024.
[25] S. Ketu, P. K. Mishra, and S. Agarwal, "Performance analysis of distributed computing frameworks for big data analytics: hadoop vs spark," *Comput. y Sist.*, vol. 24, no. 2, pp. 669–686, 2020.
[26] V. Klös, T. Göthel, and S. Glesner, "Comprehensible and dependable self-learning self-adaptive systems," *J. Syst. Archit.*, vol. 85, pp. 28–42, 2018.
[27] O. Gheibi, D. Weyns, and F. Quin, "Applying machine learning in self-adaptive systems: A systematic literature review," *ACM Trans. Auton. Adapt. Syst.*, vol. 15, no. 3, pp. 1–37, 2021.
[28] B. learning with R. expert techniques for predictive modeling Lantz, *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.
[29] M. Haleem, M. F. Farooqui, and M. Faisal, "Tackling Requirements Uncertainty in Software Projects: A Cognitive Approach," *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 180–190, 2021, doi: https://doi.org/10.1016/j.ijcce.2021.10.003.
[30] A. Ullah, S. N. Khan, and N. M. Nawi, "Review on sentiment analysis for text classification techniques from 2010 to 2021," *Multimed. Tools Appl.*, vol. 82, no. 6, pp. 8137–8193, 2023.
[31] I. Bouabdallaoui, F. Guerouate, and M. Sbihi, "Combination of genetic algorithms and K-means for a hybrid topic modeling: tourism use case," *Evol. Intell.*, vol. 17, no. 3, pp. 1801–1817, 2024.
[32] M. E. Saleh, Y. M. Wazery, and A. A. Ali, "A systematic literature review of deep learning-based text summarization: Techniques, input representation, training strategies, mechanisms, datasets, evaluation, and challenges," *Expert Syst. Appl.*, vol. 252, p. 124153, 2024.
[33] V. Boppana and P. Sandhya, "Distributed focused web crawling for context aware recommender system using machine learning and text mining algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 3, 2023.
[34] B. Bharadwaj, S. Nayak, and P. K. Panigrahi, "Sentiment analysis for identifying depression through social media texts using machine learning technique," *Big Data Comput. Visions*, vol. 5, no. 2, pp. 102–118, 2025.
[35] E. O. Abiodun, A. Alabdulatif, O. I. Abiodun, M. Alawida, A. Alabdulatif, and R. S. Alkhawaldeh, "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities," *Neural Comput. Appl.*, vol. 33, no. 22, pp. 15091–15118, 2021.
[36] G. Xu, X. Wu, H. Yao, F. Li, and Z. Yu, "Research on topic recognition of network sensitive information based on SW-LDA model," *IEEE access*, vol. 7, pp. 21527–21538, 2019.
[37] A. Sharma, N. Patel, and R. Gupta, "Enhancing customer segmentation through AI: Analyzing clustering algorithms and deep learning techniques," *Eur. Adv. AI J.*, vol. 11, no. 8, 2022.
[38] S. G. Tejashwini and D. Aradhana, "Multimodal deep learning approach for real-time sentiment analysis in video streaming," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, 2023.
[39] Z. Tao *et al.*, "Log2intent: Towards interpretable user modeling via recurrent semantics memory unit," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1055–1063.