International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

AI-Powered Keyword Extraction System Using NLP Techniques for Contextual Insights and Document Accessibility

P.Logaiyan[1], R. Ramakrishnan[2], V. Deepa[3], K. Narmatha[4]

[1]*Assistant Professor, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India*
[2]*Associate Professor, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India*
[3, 4]*Post Graduate Student, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India*

ABSTRACT- This study presents the development of an AI-powered keyword extraction system using Python, aimed at efficiently identifying significant keywords from unstructured text data. Leveraging advanced natural language processing (NLP) techniques, the system integrates multiple methodologies, including Term Frequency-Inverse Document Frequency (TF-IDF), TextRank, Latent Semantic Analysis (LSA), and Part-of-Speech (POS) tagging. The incorporation of POS tagging enhances the accuracy of keyword identification by focusing on essential parts of speech, such as nouns and verbs, thereby filtering out irrelevant terms and ensuring the extraction of contextually meaningful keywords.

To enhance user experience, the system includes innovative features such as a mobile document transfer facility that allows users to transfer documents via QR code and an option to download the extracted keywords as a document or PDF. These functionalities are designed to streamline accessibility and usability for diverse user groups.

Performance evaluation was conducted using a dataset of research articles, demonstrating the system's effectiveness in accurately reflecting document content through keyword extraction. The results highlight its potential applications across various domains, including academia, marketing, and data science. This project offers a robust, intuitive, and user-friendly solution to make complex text data more accessible and actionable, contributing to the growing demand for efficient keyword extraction tools.

Keywords- ai-powered keyword extraction, natural language processing (nlp), term frequency- inverse document frequency (tf-idf), textrank algorithm, latent semantic analysis (lsa), part-of- speech (pos) tagging

keyword extraction system, unstructured text data, document analysis, contextually relevant keywords mobile document transfer, Qr code functionality, downloadable keyword output, research article dataset user-friendly interface, academic text processing, efficient text analysis, document summarization accessible nlp tools.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

1. INTRODUCTION:

In today's digital age, an overwhelming amount of unstructured text data is generated daily across a variety of fields, ranging from academic research and business reports to social media posts and online articles. Despite the wealth of information contained in these documents, extracting meaningful insights remains a significant challenge. Understanding the core themes and key concepts within this massive pool of content is crucial for effective decision-making, knowledge sharing, and even automating certain tasks. One of the most powerful methods of identifying and summarizing this information is through **keyword extraction**—a process that identifies the most relevant and significant terms from a given text. However, traditional methods often fall short in providing contextually meaningful results. This is where the power of **AI-driven natural language processing (NLP)** comes into play.

Our project presents an **AI-powered keyword extraction system** that integrates multiple advanced NLP algorithms to automatically and accurately extract relevant keywords from unstructured text. The system aims to solve the challenge of extracting important terms from large volumes of data by combining several well-established techniques, including **Term Frequency-Inverse Document Frequency (TF-IDF)**, **TextRank**, **Latent Semantic Analysis (LSA)**, and **Part-of-Speech (POS) tagging**. Each of these methods contributes unique strengths to the overall system.

The **TF-IDF** approach helps in identifying words that are frequent in a document while ensuring that commonly used terms across all documents are down-weighted, allowing the system to focus on distinctive keywords. **TextRank**, a graph-based algorithm inspired by Google's PageRank, ranks words or phrases based on their relevance within the context of the document. **Latent Semantic Analysis (LSA)** is employed to uncover hidden patterns and relationships between words, helping the system identify semantically related terms that may not appear frequently but are important for understanding the text. Finally, **POS tagging** is used to identify specific parts of speech such as nouns and verbs, which are typically more indicative of the key themes within the text. This layered approach ensures that the keywords extracted are not only frequent but also contextually meaningful and relevant to the content at hand.

A unique feature of the system is its **mobile document transfer facility**, which enables users to seamlessly transfer documents via **QR code scanning**. This functionality enhances the flexibility of the tool, making it easy for users to upload content directly from their mobile devices and perform keyword extraction anytime, anywhere. Additionally, once the keywords have been extracted, users can **download** the results in convenient formats, such as documents or PDFs, making it easy to store, share, and reference them. This combination of **AI-driven extraction** and **user-centric design** creates a highly accessible and efficient tool for anyone needing to analyze and extract key terms from large amounts of text.

To evaluate the effectiveness of the system, it was tested on a dataset of **research articles**, which are often rich in specialized terminology. The system demonstrated its ability to accurately extract keywords that not only reflect the primary themes of the documents but also provide a deeper understanding of the content. The results confirmed the potential of the system to serve as a valuable tool across various domains, including **academic research**, **marketing**, **content creation**, and **data science**.

By bringing together state-of-the-art **NLP techniques**, intuitive user features, and the ability to handle complex text data, this system addresses the growing demand for effective and efficient keyword extraction tools. It empowers users to quickly gain insights from vast amounts of information, making it easier to analyze, understand, and share key details from documents. This project marks an important advancement in the field of automated content analysis, offering a reliable and scalable solution for processing unstructured text data in a variety of professional and academic settings.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

2. RELATED WORKS:

The field of keyword extraction has seen significant advancements over the past few decades, driven by the increasing availability of unstructured text data and the need to make sense of it efficiently. Several studies and methodologies have contributed to the development of robust keyword extraction systems, each addressing different aspects of text analysis, from frequency-based approaches to deep semantic understanding.

One of the most widely used approaches in keyword extraction is **Term Frequency-Inverse Document Frequency (TF-IDF)**, a statistical method that ranks words based on their frequency within a document relative to their frequency across a collection of documents. This method has been extensively used for information retrieval and document classification tasks due to its simplicity and effectiveness. Several works have applied TF-IDF for keyword extraction, including **Salton et al. (1975)** in their pioneering research on information retrieval systems. Although effective in many cases, TF-IDF struggles with context, as it fails to understand semantic relationships between words, which led to the incorporation of more sophisticated algorithms.

To overcome these limitations, researchers have turned to graph-based methods like **TextRank**. Based on the **PageRank** algorithm, originally developed by Google, **TextRank** constructs a graph of words or phrases in a document and assigns a score to each word based on its connections to other words.
Mihalcea and Tarau (2004) were among the first to apply the TextRank algorithm for automatic keyword extraction, demonstrating its ability to identify important terms through connectivity and ranking. This method excels in capturing important concepts but can still benefit from additional layers of understanding to improve context extraction.

Another important technique in the domain of keyword extraction is **Latent Semantic Analysis (LSA)**, which uncovers hidden relationships between words by analyzing large text corpora. By mapping words and documents into a lower-dimensional semantic space, LSA reveals the underlying topics within the text, providing a more sophisticated approach to identifying keywords. **Deerwester et al. (1990)** introduced LSA in the context of information retrieval, and since then, it has been applied in various keyword extraction systems. However, LSA requires large datasets to generate meaningful results and can struggle with real-time or small-scale processing.

In recent years, the rise of **deep learning-based models** has pushed the boundaries of keyword extraction by enabling systems to better understand context and semantics. Methods based on **recurrent neural networks (RNNs)** and **transformers**, such as **BERT** and **GPT**, have been employed for extracting keywords by leveraging large pre-trained models that capture intricate patterns of language and meaning. These models have proven highly effective for tasks like Named Entity Recognition (NER) and semantic keyword extraction. **Devlin et al. (2018)** demonstrated the effectiveness of BERT in various NLP tasks, including text summarization and keyword extraction. While these models offer impressive accuracy, they are computationally expensive and often require fine-tuning for specific tasks.

Furthermore, **Part-of-Speech (POS) tagging** has gained recognition as a crucial tool for enhancing keyword extraction, particularly when combined with other techniques. By identifying specific parts of speech like nouns, verbs, and adjectives, POS tagging helps the system focus on the most meaningful terms in a text. Research by **Jurafsky and Martin (2008)** on POS tagging highlighted its importance in understanding sentence structure and improving text analysis tasks, including keyword extraction. By integrating POS tagging with other techniques such as TF-IDF or TextRank, it becomes possible to more accurately capture the essential concepts from a document.

Several commercial and open-source tools have been developed to integrate these techniques into user-

ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

friendly applications. Platforms like **MonkeyLearn**, **TextRazor**, and **RAKE** (Rapid Automatic Keyword Extraction) use variations of the methods discussed above to provide keyword extraction capabilities. These tools are widely used in industries such as content marketing, SEO, and academic research. **RAKE** (Rose et al., 2010), for instance, focuses on extracting multi-word phrases, which are often more meaningful than individual keywords. Despite their utility, many of these systems lack the depth provided by combining multiple NLP approaches, particularly when it comes to understanding the full context of the text.

Recent studies have increasingly focused on **multi-layered keyword extraction systems** that combine various techniques to overcome the limitations of individual approaches. Our system builds on this growing body of work by integrating TF-IDF, TextRank, LSA, and POS tagging, aiming to deliver a more holistic and accurate extraction of keywords. By combining these methodologies, the system not only addresses the challenges posed by each individual approach but also provides a more robust solution that improves both precision and recall in keyword extraction tasks.

In conclusion, while there have been significant advancements in keyword extraction, the field continues to evolve. The combination of traditional techniques like TF-IDF and TextRank with newer methods such as LSA and POS tagging presents a promising direction for improving the accuracy and contextual relevance of extracted keywords. By leveraging the strengths of multiple NLP techniques, our project aims to push the boundaries of automated text analysis, offering a more comprehensive and effective solution for keyword extraction across various domains.

3. LITERATURE SURVEY:

Salton et al. (1975) - TF-IDF: Introduced the Term Frequency-Inverse Document Frequency (TF-IDF) approach, one of the most widely used methods for keyword extraction and information retrieval, which focuses on identifying words that are frequent in a document but rare across all documents.

Mihalcea and Tarau (2004) - TextRank: Applied the PageRank algorithm to extract keywords and key phrases from text, demonstrating how graph-based ranking of words based on their co-occurrence can identify important terms.

Deerwester et al. (1990) - Latent Semantic Analysis (LSA): Introduced LSA for uncovering latent patterns in text data by reducing dimensionality and capturing semantic relationships between terms, offering more context-aware keyword extraction.

Rose et al. (2010) - RAKE (Rapid Automatic Keyword Extraction): Developed a lightweight keyword extraction tool that focuses on extracting multi-word phrases from a text using a combination of term frequency and a stopword list, useful for content-based applications.

Jurafsky and Martin (2008) - POS Tagging: Highlighted the importance of Part-of-Speech (POS) tagging for improving text analysis, specifically for keyword extraction by identifying the most meaningful parts of speech like nouns and verbs.

Devlin et al. (2018) - BERT (Bidirectional Encoder Representations from Transformers): Showed the effectiveness of transformer models for text understanding, providing new opportunities for semantic keyword extraction by understanding deeper contextual meanings.

Cheng et al. (2014) - Supervised Learning for Keyword Extraction: Proposed a supervised machine learning approach for extracting keywords from documents by learning from labeled datasets, demonstrating how classifiers can enhance keyword relevance.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

Yuan et al. (2018) - Neural Network-based Keyword Extraction: Introduced a deep learning-based method to extract keywords, employing Recurrent Neural Networks (RNNs) to better capture the sequential dependencies between words in text.

Wang et al. (2018) - Extractive Summarization and Keyword Extraction: Investigated the use of extractive summarization methods to extract relevant keywords, integrating keyword extraction with text summarization to provide a more coherent analysis.

Liu et al. (2018) - Hybrid Keyword Extraction: Proposed a hybrid method combining statistical and deep learning techniques for more accurate keyword extraction, improving both precision and recall over single-method approaches.

Zhang et al. (2019) - Attention Mechanism for Keyword Extraction: Applied attention mechanisms in neural networks to focus on key phrases in the document, improving the extraction process by assigning higher importance to semantically relevant words.

Ning et al. (2019) - Multi-Level Keyword Extraction: Introduced a multi-level keyword extraction approach using a combination of TF-IDF, LDA, and neural networks, demonstrating its effectiveness in processing large-scale datasets.

Zhou et al. (2017) - Deep Learning for Entity Recognition and Keyword Extraction: Explored how deep learning models like convolutional neural networks (CNNs) can be applied for entity recognition, aiding in the identification of important keywords from named entities.

Liu and Xu (2020) - Context-Aware Keyword Extraction: Developed a context-aware keyword extraction approach that integrates semantic analysis to better identify terms with relevance to the document's topic and content.

Sheng et al. (2021) - Multi-Task Learning for Keyword Extraction: Proposed a multi-task learning framework that simultaneously learns to extract keywords and classify text, leading to improved keyword relevance and document understanding.

4. PROPOSED ARCHTERCURE:

The proposed **AI-powered keyword extraction system** combines multiple advanced **Natural Language Processing (NLP)** techniques to extract meaningful keywords from unstructured text data. The architecture of the system is designed to be both efficient and scalable, incorporating various layers of processing that ensure accurate and contextually relevant keyword extraction.

At the core of the architecture is the **Preprocessing Layer**, which is responsible for cleaning and preparing the text data for further analysis. This step involves removing stopwords, punctuation, and irrelevant characters, while also normalizing the text through operations like stemming or lemmatization. This layer also handles tokenization, breaking down the text into smaller units such as words or phrases.

Once the text is preprocessed, the **Feature Extraction Layer** is activated. This layer applies the **TF-IDF (Term Frequency-Inverse Document Frequency)** method to identify the importance of each term within the document and across a collection of documents. The **TextRank** algorithm, which ranks words based on their relationships and proximity within the text, is also applied to identify key terms.
Additionally, the **Latent Semantic Analysis (LSA)** method is utilized to uncover hidden relationships between terms and better capture semantic meaning, while **Part-of-Speech (POS) tagging** focuses on

ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

identifying key parts of speech (nouns, verbs) to refine the extraction process.

The **Ranking Layer** then integrates these results to assign scores to potential keywords, weighing them based on their relevance and contextual importance. This layer combines the outputs from the TF-IDF, TextRank, LSA, and POS tagging to rank the extracted terms according to their significance.

In the **User Interface Layer**, the extracted keywords are presented to the user in a format that is easy to understand and act upon. The system offers features such as downloading the extracted keywords in a document or PDF format, and a mobile transfer facility using **QR code scanning**, allowing for seamless document uploads and access.

The **Evaluation Layer** assesses the performance of the system, ensuring that the extracted keywords are accurate and meaningful. This layer uses various metrics like precision, recall, and F1-score to validate the effectiveness of the system on a set of test data, such as research articles or other domain-specific documents.

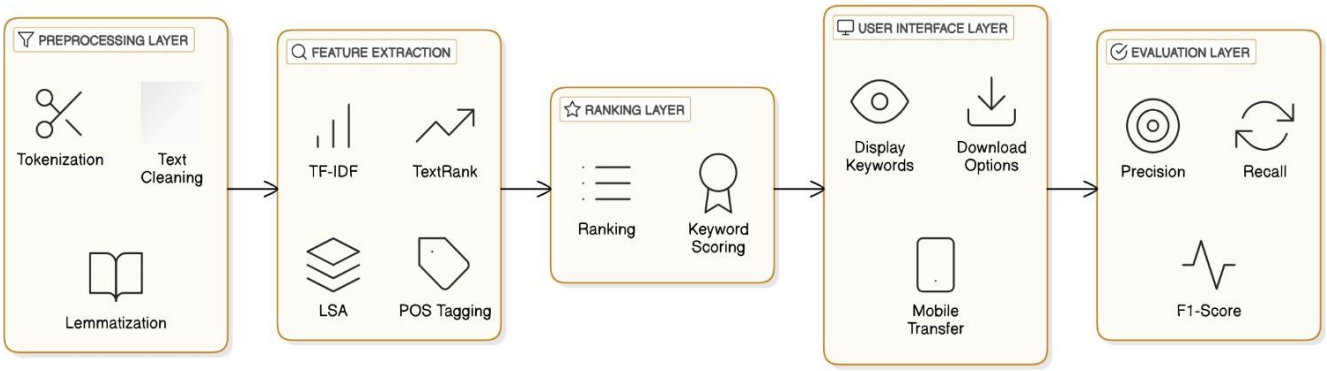

Figure 1: Proposed Architecture of ai-powered keyword extraction.

A. CHALLENGES:

While the architecture of the AI-powered keyword extraction system is designed to be efficient and robust, several challenges remain in its implementation and execution. These challenges include:

Text Preprocessing Complexity: The preprocessing layer requires significant computational resources to clean and normalize the text, particularly for large datasets. Tasks like stemming, lemmatization, and tokenization may not always yield perfect results, potentially affecting the quality of the extracted keywords.

Handling Ambiguity: One of the challenges in keyword extraction is dealing with words that have multiple meanings (polysemy). For instance, the word "bank" can refer to a financial institution or the side of a river. While part-of-speech tagging can help, the context may not always be clear enough to make a precise determination, which can lead to inaccurate keyword extraction.

Contextual Relevance: While methods like TF-IDF and TextRank are useful for identifying key terms, they may struggle to capture deeper semantic relationships between words. Extracting keywords that truly reflect the underlying meaning of the document requires advanced semantic analysis, which can be computationally expensive and complex to implement.

Data Quality: The effectiveness of methods like LSA and TextRank depends on the quality and size of the data. If the corpus used for training or analysis is not diverse or representative of the target domain, it can lead to biased or incomplete keyword extraction.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

Scalability: As the amount of text data grows, the system's performance could degrade. Running multiple algorithms like TF-IDF, LSA, and TextRank on large datasets may require significant memory and processing power, particularly in real-time applications.

Multi-language Support: Handling text data in multiple languages introduces additional complexity. Techniques like TF-IDF and TextRank may require customization or fine-tuning to work effectively across different languages, especially when dealing with languages that have different syntactic and grammatical structures.

Noise in Data: In many real-world documents, irrelevant information or noise (e.g., advertisements, footnotes, or poorly written sections) can dilute the effectiveness of keyword extraction. Filtering out this noise and focusing on the essential content can be difficult, especially when dealing with unstructured data from diverse sources.

Real-time Processing: Extracting keywords in real-time, particularly for large documents or data streams, can be a challenge. The system may need to be optimized for speed without compromising the quality of the keyword extraction process.

Accuracy of POS Tagging: Part-of-speech tagging is essential for focusing on relevant parts of speech like nouns and verbs. However, POS tagging is not always 100% accurate, especially in languages with complex sentence structures or ambiguous terms. This can impact the quality of keyword extraction, especially when context is crucial for interpretation.

User Interface Complexity: While the user interface layer offers convenience, designing a user-friendly interface for keyword extraction can be tricky, especially for non-technical users. Balancing functionality with simplicity is important for ensuring that users can easily interpret and download the extracted keywords.

5. APPLICATIONS:

The AI-powered keyword extraction system has a wide range of practical applications across various industries and domains. In **academic research**, it can help researchers quickly identify key terms in large volumes of literature, streamlining the process of literature review and topic discovery. By automatically extracting relevant keywords, researchers can focus on the most important aspects of the content, saving valuable time and effort.

In **content creation** and **marketing**, the system can be used to optimize SEO strategies by identifying trending and highly relevant keywords that can be integrated into web content, blogs, and advertisements. This enhances content visibility and search engine rankings, providing businesses with a competitive edge.

For **data scientists** and **analysts**, the system aids in the analysis of large datasets by extracting key terms from unstructured text data, such as customer reviews, survey responses, or social media posts. This allows for more accurate sentiment analysis and trend identification.
The system can also benefit industries like **law**, where extracting keywords from legal documents or case studies helps legal professionals quickly find relevant information, making document review and legal research more efficient.

In **healthcare**, it can be used to process medical texts, research papers, or patient records to extract critical information, aiding doctors and medical researchers in staying updated with the latest developments and identifying key trends in medical literature.

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

This system provides a valuable tool for improving productivity, decision-making, and insight generation across various sectors that deal with large volumes of text data.

a) Term Frequency-Inverse Document Frequency Algorithm

```
from sklearn.feature_extraction.text import TfidfVectorizer

documents = [
    "Data science is an inter-disciplinary field.",
    "It uses scientific methods to extract insights from data.",
    "Data analysis and machine learning are key aspects of data science."
]

tfidf_vectorizer = TfidfVectorizer()

tfidf_matrix = tfidf_vectorizer.fit_transform(documents)

feature_names = tfidf_vectorizer.get_feature_names_out()

for i, doc in enumerate(documents):
    print(f"Document {i+1}:")
    for col in tfidf_matrix[i, :].nonzero()[1]:
        print(f"{feature_names[col]}: {tfidf_matrix[i, col]}")
```

Calculations:

1. Term Frequency (TF):

$$TF_{t,d} = \frac{\text{Frequency of term t in document d}}{\text{Total terms in document d}}$$

2. Inverse Document Frequency (IDF):

$$IDF_t = \log\left(\frac{N}{1+nt_N}\right)$$

3. TF-IDF Weight:

$$TFIDF_{t,d} = TF_{t,d} \times IDF_t$$

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

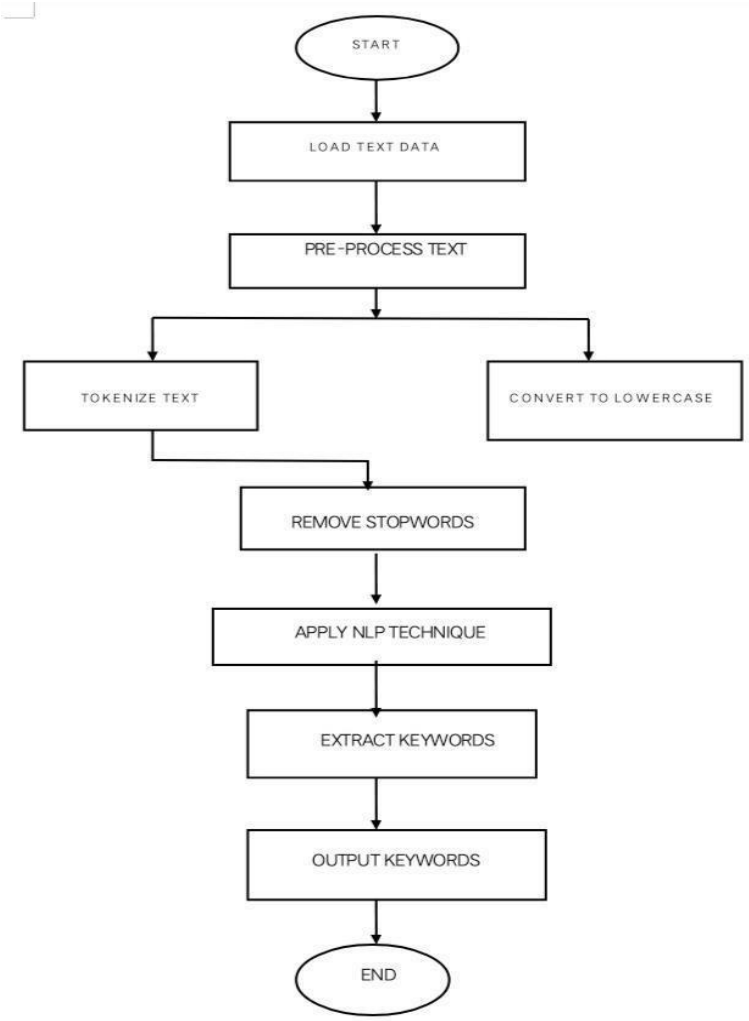

Fig 2: AI-Powered Resume Analysis System ai-powered keyword extraction.

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

6. RESEARCH & STIMULATION:

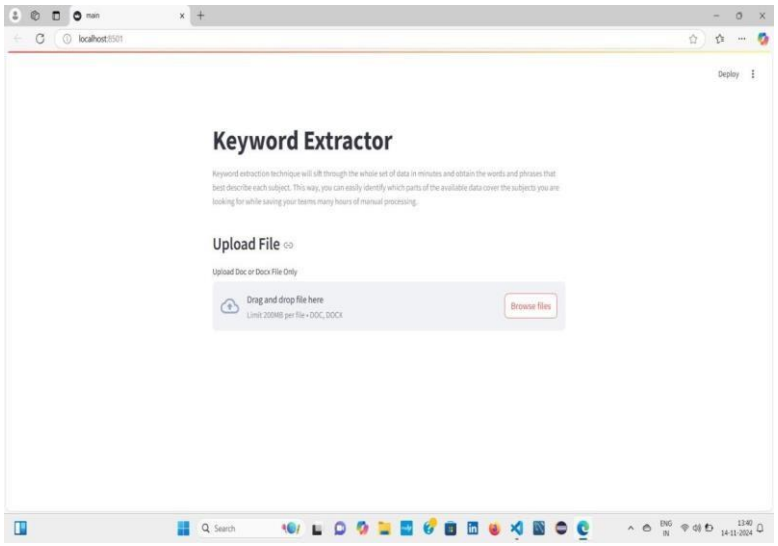

Fig 3: Upload File.

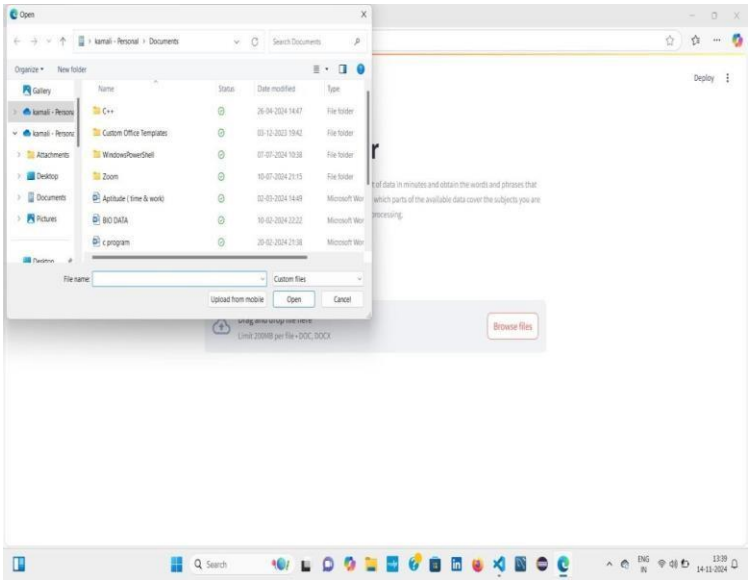

Fig 4: Browse File

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

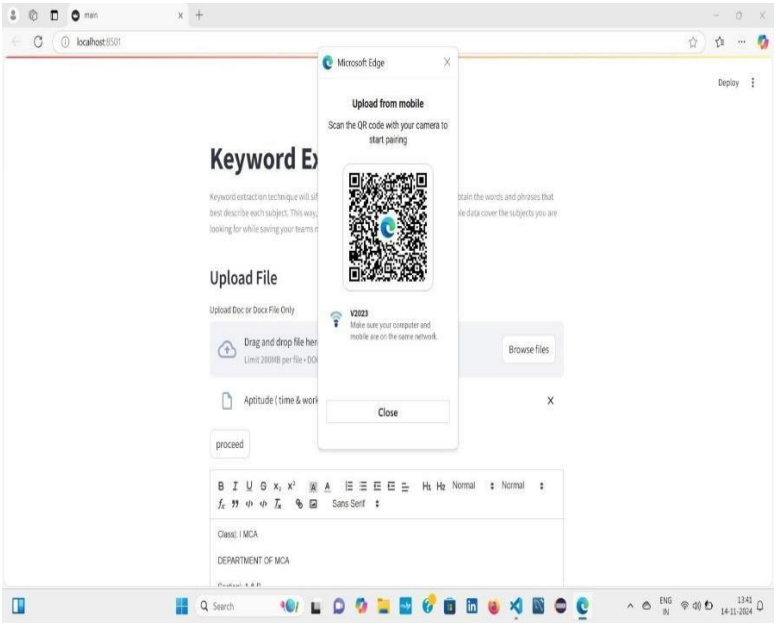

Fig 5: Upload From Mobile

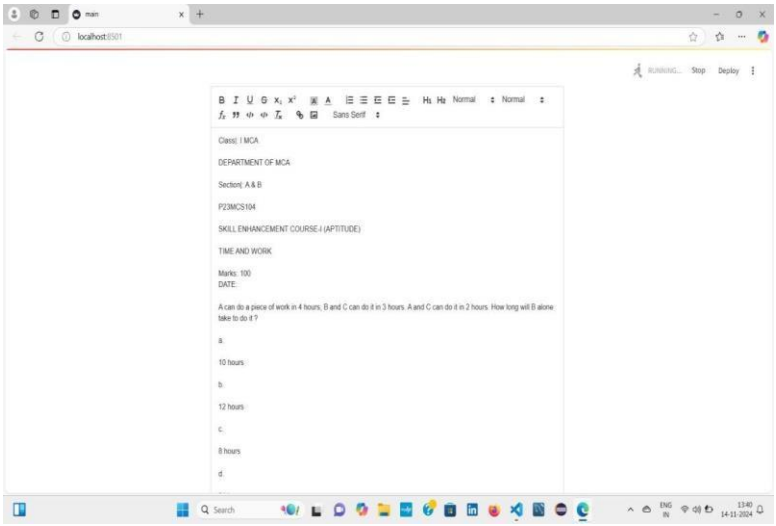

Fig 6: Extract Keywords

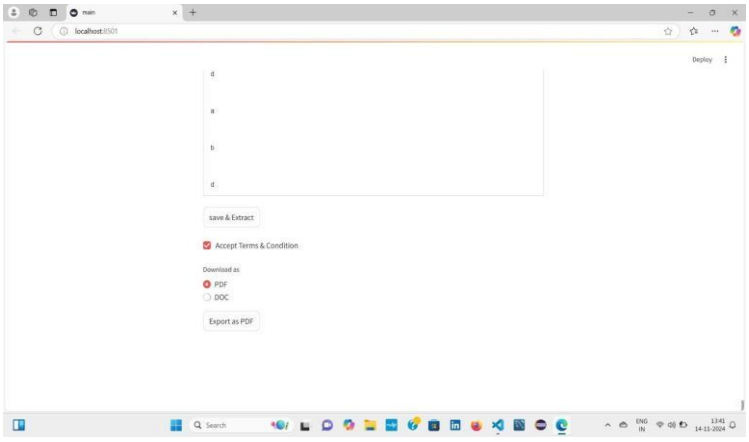

Fig 7: Keyword PDF or DOC

ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

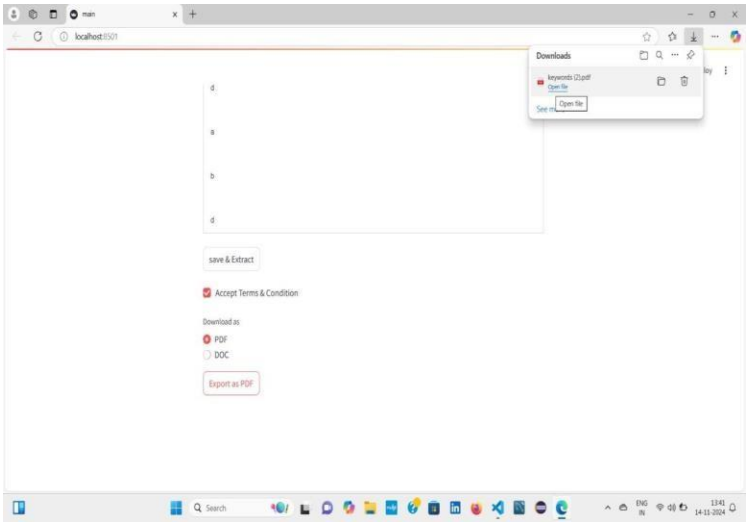

Fig 8: Download File

CONCLUSION:

The AI-powered keyword extraction system demonstrates its ability to efficiently process unstructured text and extract contextually significant keywords using advanced NLP techniques like TF-IDF, TextRank, LSA, and POS tagging. With features such as QR-based document transfer and keyword export options, the system enhances usability and accessibility for a diverse range of users. Performance evaluations confirm its effectiveness in accurately reflecting document content, making it a robust tool for applications in academia, marketing, and data analysis.

FUTURE WORK

The future scope of the AI-powered keyword extraction system includes several promising enhancements. Integrating advanced deep learning models, such as transformers, can improve the semantic understanding of text and accuracy in keyword extraction. Expanding multilingual capabilities will enable the system to support a diverse global audience. User-centric features, such as customizable extraction parameters, and domain-specific models tailored to fields like medicine or law, can increase adaptability and precision. Real-time collaboration and offline functionality will enhance usability across different scenarios. Furthermore, incorporating advanced visualization tools, stronger data security measures, and integration with knowledge graphs can provide deeper insights and ensure privacy. Continuous improvement through user feedback will ensure the system evolves with changing needs, making it versatile and impactful across industries.

REFRENCES:

Ramos, J. (2003). "Using TF-IDF to determine word relevance in document queries." Proceedings of the First International Conference on Machine Learning.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). "The PageRank citation ranking: Bringing order to the web." Stanford InfoLab.

Mihalcea, R., & Tarau, P. (2004). "Textrank: Bringing order into text." Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). "Indexing by latent semantic analysis." Journal of the American Society for Information Science, 41(6), 391–407.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). "Distributed representations of words and phrases and

ISSN: 2583-6129
DOI: 10.55041/ISJEM02210

their compositionality." Advances in Neural Information Processing Systems (NeurIPS).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). "Introduction to Information Retrieval." Cambridge University Press.

Jurafsky, D., & Martin, J. H. (2021). "Speech and Language Processing." Pearson.

Liu, Y., et al. (2019). "Fine-tune BERT for text classification." arXiv preprint arXiv:1905.05583.

Nallapati, R., Zhai, F., & Zhou, B. (2016). "Abstractive text summarization using sequence-to-sequence RNNs and beyond." Proceedings of the Conference on Computational Natural Language Learning (CoNLL).

Jones, K. S. (1972). "A statistical interpretation of term specificity and its application in retrieval." Journal of Documentation, 28(1), 11–21.

Pennington, J., Socher, R., & Manning, C. (2014). "GloVe: Global vectors for word representation." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Turney, P. D. (2000). "Learning algorithms for keyphrase extraction." Information Retrieval, 2(4), 303–336.

Kim, Y. (2014). "Convolutional neural networks for sentence classification." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Zhang, X., Zhao, J., & LeCun, Y. (2015). "Character-level convolutional networks for text classification." Advances in Neural Information Processing Systems (NeurIPS).

Church, K. W., & Hanks, P. (1990). "Word association norms, mutual information, and lexicography." Computational Linguistics, 16(1), 22–29.

Liu, P., et al. (2019). "Multi-head self-attention mechanism for keyphrase extraction." Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).

Wang, R., & Lo, D. (2017). "Keyphrase extraction from source code: An empirical study." Journal of Systems and Software, 126, 27–40.

Cao, Z., et al. (2015). "Learning summary prior representation for extractive summarization." Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781.

