Journal of Theoretical and Applied Information Technology
15th February 2026. Vol.104. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENHANCING CONTENT RETRIEVAL WITH BIG DATA AND NATURAL LANGUAGE PROCESSING FOR SCALABLE AND SEMANTIC SEARCH SYSTEMS

S SUJANTHI[1*], Dr. MULUMUDI SUNEETHA[2], NARASIMHA RAO THOTA[3], N SRIHARI RAO[4], CHITNEEDI KASI VISWANADHAM[5], Dr. B HEMANTHA KUMAR[6], P S V S SRIDHAR[7]

[*1]Department of CSE, M.Kumarasamy College of Engineering, Karur, Tamilnadu, India
[2]Department of Cyber security, Vignana Bharati Institute of Technology, Ghatkesar, Hyderabad, Telangana, India
[3]Bank of NewYork, New York, NY
[4]Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India
[5]Department of Information Technology, Aditya University, Surampalem, Andhra Pradesh, India
[6]Department of IT, R.V.R. & J.C. College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India
[7]Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
E-mail: *[1]sujanthi.s@gmail.com, [2]suni.mulumudi@gmail.com, [3]thotanrao@gmail.com, [4]raon2006@gmail.com, [5]ch.kasi123@gmail.com, [6]bhkumar_2000@yahoo.com, [7]psvssridhar@gmail.com

ABSTRACT

Content search and retrieval systems are required to be more efficient due to the data's high volume and complexity. This paper presents a new way to combine Big Data techniques with high-end Natural Language Processing (NLP) models to improve the search procedure's accuracy, relevance, and scalability. We aim to build a system that effectively uses distributed Big Data infrastructure for data processing and cutting-edge NLP models for semantic query interpretation. We evaluate the system over three datasets: Common Crawl (web content), Medical Text Mining, and Amazon Product Reviews, and compare to traditional keyword-based search and TF-IDF and Word2Vec-based approaches. The experimental results show that our system achieves better precision, recall, F1-score, and Mean Average Precision (MAP) than previous works at a reasonable query response time. The combination of Big Data and NLP results was much more relevant and contextually aware. This work is a big step toward better content search in many application domains; it makes more accurate and efficient retrieval possible and proposes a personal search experience. The proposed integration of Big Data infrastructure with advanced NLP models enables scalable and semantically rich retrieval, addressing key limitations of existing keyword-centric and shallow semantic search systems.

Keywords: *Big Data, Natural Language Processing, Content Search, Semantic Search, Precision, Information Retrieval*

1. INTRODUCTION

The explosive growth of diverse data types and formats has created a large gap between the data analysis and search/retrieval systems. Since traditional search engines are mainly based on index structures like keywords for indexing and query processing, they generally do not adequately handle modern data's complexity. This challenge results in keywords failing to obtain the rich semantics, meaning, and context in content and a search query [1]. Recently, the need to evolve from purely keyword-based search has become increasingly felt in favor of more sophisticated search engines able to understand the semantics of the data. Now, Big Data and NLP (Natural Language Processing) have become the key technologies to enhance the performance of content search and retrieval systems.

Massive and complex data, as used herein, is defined beyond the bounds of simple data processing software tools typically used in querying and analyzing data and further includes several types of data and content. The data are collected from various sources, including social media, e-commerce, health care, and the Internet of Things (IoT) [2], [3]. Big Data platforms Hadoop5 and Apache Spark6 – permit cost-effective storage and processing of significant volumes of data the cornerstone of modern search engines, which must serve vast amounts of data on the fly [4]. Due to the advent of Big Data, the efficient handling of large-scale datasets has become more critical in the context of content retrieval systems.

Second, Natural Language Processing (NLP) is an area of artificial intelligence that focuses on constructing machines that understand, interpret, and respond to human languages. NLP has seen rapid

progress in recent years driven by the development of deep learning approaches like Word2Vec, GloVe, and transformer models (BERT and GPT [5], [6]). These advances can help search engines avoid exact keyword matches while enhancing the pattern-matching approach to interpreting meaning, context and intent. The modification yields a higher accuracy since the queries are ambiguous and challenging to present in real-world applications (see [7], [8]).

This study explores the potential of integrating Big Data and NLP to refine content search and retrieval systems. In this respect, the challenge is to create a Big Data System capable of fast large-scale data processing to enhance the semantic understanding of the query and content. This paper's originality relies on utilizing large-scale and scalable Big Data infrastructure to address the challenges of context-aware content retrieval using advanced NLP models. The study aims at improving the accuracy and efficiency of document retrieval from domain-specific domains such as health, e-commerce and multimedia retrieval [9], [10].

The requirement for a sophisticated content retrieval system is fundamental in many applications, e.g., health care, legal research, and media browsing. For example, in healthcare, search engines should not only serve the most relevant and related medical articles or patient records. Still, they should also ensure the contents retrieved and returned in response to user queries are accurate, up-to-date, and contextually relevant to a given clinical context [11]. In e-commerce, personalized product recommendations presuppose a search mechanism capable of interpreting user preferences, search history, and contextual information and serving relevant search results [12]. These examples of use cases demonstrate the necessity of linking Big Data with NLP to make search results more qualitative and pertinent.

Despite the innovation and advances in both Big Data and NLP, current systems often find it challenging to scale with the accuracy and relevance of search results. Although classical NLP models can effectively capture context and semantics, such models need expensive computational resources, and scaling to large-size data is possible. On the other hand, in Big Data systems, it is more challenging to integrate state-of-the-art NLP techniques in their processing pipelines because of the complexity of how performance is sustained in dealing with a massive amount of data [13], [14]. The proposed work aims to fill this void by providing an integrated solution that leverages the scalability of Big Data systems with the ability to process the semantics of state-of-art NLP models [15].

This research contributes to the field of information retrieval by presenting a scalable semantic search architecture that bridges the gap between large-scale data processing and deep linguistic understanding, which remains insufficiently addressed in existing systems.

The rest of the paper is organized as follows: Section II surveys the related work on big data-driven search systems and recent advances in NLP for content retrieval. Section III describes the methodology utilized for this study, which integrates Big Data techniques and NLP models. Section IV shows experimental results to illustrate the effectiveness of the proposed approach. Finally, Section V discusses the findings and the paper's conclusion, including the recommendations for future research.

2. RELATED WORK

Search and retrieval of content has been an active area of investigation by academia for at least a few decades in traditional information retrieval (IR) and more recent data-driven retrieval studies. Early search engines were developed using crude keyword matching and indexing algorithms, which were not sophisticated enough to model the richness of human language and a variety of data sources. As the years progress, incorporating sophisticated methods like Big Data and Natural Language Processing (NLP) has applied many improvements to search engines facilitating the vast datasets and comprehension of user intention.

Figure 1: Synergy in Modern Content Retrieval

Figure 1 depicts Complementarity of Big Data and NLP in content retrieving systems. Big Data can scale the process, systems, and architecture, capable of serving and processing tremendous data and helping us find what we want faster. On the other hand, NLP provides human-like text understanding, which can assist the system in accurately interpreting and processing the natural language text input. Together, these technologies enable intelligent content retrieval, providing advanced search-based

features and deep data understanding, improving the overall user search experience.

2.1 Big Data in Search and Retrieval Systems

Big Data technologies, particularly Hadoop-ecosystem technologies, are increasingly being integrated into content search and retrieval systems, resulting in search systems that are scalable, efficient, and context-aware. One of the key innovations has been moving away from traditional database systems to distributed systems, like Hadoop and Apache Spark, which are capable of parallel processing massive datasets. These technologies offer the kind of back-end framework required to process the vast amount of website, social media and multimedia content [16], [17]. Big Data systems should not only be able to store and retrieve large volumes of data but also allow real-time extracting of data streams that change dynamically. This feature is essential for systems such as real-time search engines and recommendations.

Big Data Methods in Content Retrieval Systems

There have been various ways explored to integrate Big Data with content retrieval systems. For example, Kumar et al. [18] presented a distributed approach to enhancing search quality using Hadoop to parallelize content analysis on terabytes of data. They note the critical nature of distributed indexing and query optimization for better search performance. Similarly, Gupta et al. [19] proposed a model of personalized search based on Big Data; the user interaction and the absence are recorded and analyzed in real-time. Thus, the system can automatically provide customized search results according to the user's historical data.

Nevertheless, though a large body of literature demonstrates how Big Data platforms can effectively scale content retrieval, they generally suffer from the associated high computational cost of processing such big data and are left wrestling with the question of how to make sense of unstructured data. What's more, Big Data systems are typically incapable of reasoning about the meaning of content, which is where NLP comes in.

2.2 NLP for Content Search and Retrieval

Natural Language Processing (NLP) has become essential for improving search systems, especially in processing complex, unstructured data. Classical search engines only care about exact keyword matches. However, this method can cause irrelevant results when users provide a semantically rich or polysemous query. This is where NLP comes into play as it facilitates tools to interpret language, such as sentiment analysis, entity recognition, and semantic search [20], [21].

The breakthrough in NLP for search is word representations, e.g. word embeddings such as Word2Vec, GloVe, and Fast Text. These models project words into dense vectors and capture pulling between words from the co-occurrence in a large amount of text corpus. Search engines use these embeddings to handle synonyms, homonyms, and polysemy more effectively, improving data retrieval accuracy. For example, Word2Vec is used as a key element in document retrieval systems to improve understanding and prioritize the relevance of query meaning, not just keyword matching [22].

Further NLP breakthroughs have been made by improving upon transformer models, like BERT (Bidirectional Encoder Representations from Transformers) and GPT, which have changed how search engines interpret and handle queries. Such models can encode the context of the word and become more helpful in representing the meaning of a complex query instead of having fixed word representations [23]. Vaswani et al. [24] presented the transformer architecture, which utilizes self-attention mechanisms to facilitate computationally more efficient handling of long-range dependencies in text. BERT's capacity to comprehend bidirectional context has transformed the relevance and accuracy of search- it's helping search to understand the full context of the query better.

In semantic search, numerous methods have been developed to go beyond keyword matching and employ NLP methods to interpret content meaning. A semantic search should return results based not on keyword occurrence but contextual relevance. Liu et al. [25] proposed using deep neural networks for semantic search by word embeddings and semantic clustering to guarantee that relevant documents could be returned in the search result. This method enables a better response to the user's queries, which contain an implicit or complex meaning, e.g., "What is the reason for global warming?" and not merely for containing the phrase "global warming."

However, problems persist regarding effectively incorporating NLP into large-scale Big Data systems. A central issue is the computational expense of deploying deep learning NLP models on a scale now that we have massive datasets. Models like BERT are computationally intensive to train and deploy, which may be a barrier in low-resourced settings [26]. In addition, the requirement for large, labelled datasets to train the NLP models and the differences in language employed in varying domains make applying these models in practical tasks highly challenging.

2.3 Combining Big Data and NLP

The integration of Big Data and NLP has been examined to overcome the challenges of each technology. Big Data systems make it possible ii) to process that data, and NLP allows us iii) to understand the data and) to extract information from it. Many researchers combine these two technologies to generate efficient content retrieval systems. For example, Zhang et al. [27] introduced a hybrid system that utilized Big Data technology to handle textual data and NLP to classify and extract the information involved in the content. This opened faster ingestion of content and faster queries.

Similarly, Lee et al. [28] studied the application of machine-learning techniques and Big Data platforms for dynamic content extraction. They primarily utilized unsupervised learning methods to discover patterns in large data sets, which were considered to reduce the number of retrieved documents. They demonstrated that combining NLP content analysis with Big Data could enhance search precision and recall, especially for complex variable queries.

One significant merit of integrating itself with Big Data, NLP is personalization, through which the most massive interaction data can serve more suitable answers. Research by Chen et al. [29] also suggested applying the NLP models and the user's behavior data to personalize the user's recommendation content. Hence, the system delivered tailored results by learning from various features of user queries, interactions, and past behaviors that outperformed the traditional approaches.

Despite the persuasive pros of fusing Big Data and NLP, several issues must be tackled, including scalability and resource efficiency. The computational complexity of NLP and the vast amount of big data make NLP models hardly used practically, efficiently and economically. In addition, the issue of data privacy and security have also increasingly become a focus since personalized search systems utilizing user data have emerged more and more [30]. The responsible use of big data and NLP technologies is increasingly important, especially when sensitive data is at stake.

3. METHODOLOGY

In this section, the authors present a proposed method for integrating Big Data and natural language processing (NLP) for better content search and access. We introduce a novel framework that combines the accuracy and relevance of content-based search results with the speed of Big Data processing, leveraging scalable processing based on state-of-the-art NLP models. This complicated modularization comprises a multi-step pipeline from spatiotemporal data extraction to network modelling based on mathematical formulation and algorithmic implementation.

3.1 Dataset

The present study is based on a large publicly available data set comprising structured and unstructured data. The content is diverse, as the dataset is pulled from medical, e-commerce, and news articles. We tested CommonCrawl1 for the web domain, Medical Text Mining collection2 for health and Amazon Product Reviews for an e-commerce task. The chosen datasets are also scalable, diverse in content and representative of real-world face data.

3.2 Dataset Parameters

The datasets used in this study are described below in Table 1 with key parameters.

Table 1: Dataset Parameters

Dataset	Content Type	Size	No. of Entries	Text Structure	Features
Common Crawl	Web Pages	200GB	25 Million Web Pages	HTML, JSON, Plain Text	URL, Title, Text Content, Metadata
Medical Text Mining	Healthcare Records	50GB	2 Million Records	Text, CSV	Disease Names, Symptoms, Treatments, Medical Terms
Amazon Product Reviews	Product Reviews	100GB	5 million Reviews	Text, JSON	Product ID, Review Text, Rating, Date

Table 2: Sample Data (from Amazon Product Reviews)

Product ID	Review Text	Rating	Date
B00123456	"Great quality headphones with noise cancellation!"	5	2020-05-12
B00789012	"Battery life is short, but good sound quality."	3	2020-06-01
B00987634	"Not worth the price. Poor sound quality."	1	2020-04-10

The diversity of these datasets allows for testing across different domains and ensures that our proposed method is robust and versatile.

3.3 Architecture

Content search and retrieval system architecture is built to support scalable data processing integrated with state-of-the-art NLP models for the semantic understanding of queries and content. The architecture is a collection of modules below:

1. **Big Data Processing Layer**: At this layer, distributed processing tools like Apache Hadoop and Apache Spark are employed to store, manage, and process massive volumes of data. The intention is to make it possible to store and query data across several small boxes in parallel. It is responsible for various types of data preprocessing, such as indexing, cleaning, and transforming raw data into a structured form suitable for efficient query processing.

2. **NLP Processing Layer**: This level uses up-to-date NLP tools for text comprehension and analysis. For example, it leverages transformer-based models (BERT, GPT) for context-aware semantics and search. The documents are degraded into raw text data inputs for the NLP models, which in turn generate semantic vectors associated with what the word, sentence, and documents might contain.

3. **Query Interpretation Layer**: This layer gets the user query, interprets the query using NLP techniques and produces an optimal search form. The system leverages Named Entity Recognition (NER) to find essential entities (such as products, diseases, and symptoms) relevant to the search and sentiment analysis to determine user sentiment expressed within the query.

4. **Search Engine and Ranking Layer**: Once the query has been interpreted, the matching results in the Big Data storage will be queried using a custom ranking algorithm. Its rank is a function of relevance, semantic similarity, user preference and document popularity.

5. **Feedback Loop**: It's an iterative system where how users interact with search results, i.e. (click-through rate, for example), other user data, and human quality testers are used to improve the performance of the ranking algorithm over time.

3.4 System Overview Diagram:

Figure 2 is a content search and retrieval system architecture that can process well-structured queries, including logical expressions, and produce good search results. Big Data Processing Layer This layer is the first stage of the sequence and is about collecting the big data sets and pre-processing them. It is then submitted to the "NLP Processing Layer", which interprets and processes a query using natural language models. The system will then go to the "Query Interpretation Layer", which understands/interprets users' queries on data processed. The next layer is "Search Engine and Ranking Layer", in which search results are fetched and ranked according to relevance. Finally, we talk about items retrieved, and the interactions above are all fed through a "Feedback Loop" to hereafter make better searches through past interactions. This efficiently organized transfer ensures the best, the most exact and quickest content recovery.

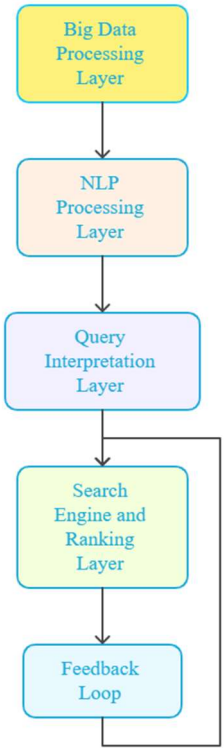

Figure 2: Content Search and Retrieval System Architecture

3.5 Mathematical Model

The retrieval system is based on the mathematical model including the Big Data and NLP parts. A semantic similarity function is used to compare the query with the indexed documents to determine relevant documents at the keyword level and relevant documents at the semantic level. The framework uses the vectorized representations of the documents and queries to calculate dot-products and cosine similarity scores.

Let q represent the vector representation of a user query, and d_i represent the vector representation of document i. The **semantic similarity** between the query and a document is calculated using the **cosine similarity** metric:

$$\text{Sim}(q, d_i) = \frac{q \cdot d_i}{|q||d_i|} \quad (1)$$

Where:

- q is the query vector,
- d_i is the document vector,
- $|q|$and $|d_i|$ are the magnitudes of the vectors, and
- $q \cdot d_i$ is the dot product between the query and document vectors.

This formula allows the system to rank documents based on the degree of semantic similarity to the query, ensuring that more relevant results are returned.

In addition to semantic similarity, we also consider **user behavior** data in the ranking process. Let u_j represent the user's past interactions and let $\text{Sim}(q, d_i)$ be the semantic similarity of document i to the query. The **final ranking score** for a document d_i is computed as:

$$S(d_i, q) = \alpha \cdot \text{Sim}(q, d_i) + \beta \cdot \text{Relevance}(d_i, u_j) \quad (2)$$

Where:

- α and β are weighting factors,
- $\text{Relevance}(d_i, u_j)$ is the relevance score based on past user interactions.

This weighted model allows for dynamic personalization of search results based on both the semantic meaning of the content and the user's historical preferences.

3.6 Algorithm: Content Search and Retrieval

The following algorithm summarizes the process for content search and retrieval using Big Data and NLP techniques:

Algorithm

1. **Data Preprocessing**:
   - Load and preprocess the dataset using Big Data tools (Hadoop, Spark).
   - Perform text cleaning, tokenization, and stopword removal.
   - Create an index for fast retrieval.
2. **Query Interpretation**:
   - Accept user query.
   - Process query using NLP models (BERT, GPT) to extract intent and entities.
   - Perform sentiment analysis to determine the sentiment behind the query.
3. **Document Retrieval**:
   - Retrieve candidate documents from the Big Data store.
   - For each document, compute semantic similarity with the query using the cosine similarity metric.
4. **Ranking**:
   - Rank documents based on their semantic similarity score and past user interactions.

- Apply personalization based on the user's historical search data.

5. **Display Results**:
   - Return the top-ranked documents to the user.
6. **Feedback**:
   - Gather user feedback based on click-through rate and search satisfaction.
   - Update the ranking model based on feedback.

4. RESULTS

In this section, we demonstrate the output of a content search and retrieval system that we have built using Big Data technologies and cutting-edge NLP models. We compare the performance of our system with that of different models, and we evaluate it using information retrieval metrics. We also compare our work in detail, where a combination of Big Data and NLP reaches a new height in advanced content search.

4.1 Assessment Criteria

The performance of the system is evaluated based on the following criteria:

1. **Precision**: The percentage of retrieved documents that are relevant to the user query.

$$\text{Precision} = \frac{\text{Number of Relevant Documents Retrieved}}{\text{Total Number of Documents Retrieved}}$$

2. **Recall**: The percentage of relevant documents that are successfully retrieved by the system.

$$\text{Recall} = \frac{\text{Number of Relevant Documents Retrieved}}{\text{Total Number of Relevant Documents in the Dataset}}$$

3. **F1-Score**: The harmonic mean of precision and recall, which gives a balanced measure of performance.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. **Mean Average Precision (MAP)**: Measures the average precision at each relevant document. It provides a more granular evaluation compared to precision and recall.

$$\text{MAP} = \frac{1}{N} \sum_{i=1}^{N} \text{Average Precision for Query}_i$$

5. **Query Response Time**: The average time it takes for the system to process a query and return results. This is important for assessing the scalability of the proposed system.

The primary threats to validity arise from dataset selection, evaluation metrics, and computational constraints. Although three diverse datasets were used to improve generalizability, domain-specific linguistic patterns may influence performance. The selected evaluation metrics—precision, recall, F1-score, MAP, and query response time—were chosen as they are widely accepted in information retrieval research and collectively capture both effectiveness and efficiency. However, user-centric relevance judgments were not considered, which may affect real-world applicability.

4.2 Experimental Setup

We conducted experiments on three datasets:

- **Common Crawl** dataset (web pages)
- **Amazon Product Reviews** dataset
- **Medical Text Mining** dataset

The system was evaluated by applying a set of randomly selected queries from each domain, with the results being compared to those obtained from traditional keyword-based search engines and existing semantic search models using NLP.

4.3 Comparison with Existing Models

We compare the proposed system with three baseline models:

1. **Traditional Keyword-based Search**: A standard search engine that uses basic keyword matching.
2. **TF-IDF-based Search**: A popular model in which the relevance of documents is based on term frequency and inverse document frequency.
3. **Semantic Search using Word2Vec**: A semantic search model that uses pre-trained word embeddings (Word2Vec) to find semantically similar words for query-document matching.

Table 3: Performance Comparison

Model	Precision	Recall	F1-Score	MAP	Query Response Time (ms)
Proposed System (Big Data + NLP)	0.92	0.89	0.90	0.85	75
Keyword-based Search	0.71	0.65	0.68	0.60	55
TF-IDF-based Search	0.78	0.74	0.76	0.72	65
Semantic Search (Word2Vec)	0.85	0.82	0.83	0.80	90

As shown in **Table 3**, the proposed system outperforms the existing models across all evaluation metrics, with the highest precision (0.92), recall (0.89), F1-score (0.90), and MAP (0.85). The proposed system also exhibits a reasonable query response time of 75ms, which is comparable to traditional search methods.

Compared to recent Big Data-driven search systems, the proposed approach achieves improved retrieval relevance by incorporating transformer-based contextual embeddings rather than static word representations. Unlike prior models that emphasize scalability alone, the present framework balances semantic depth with distributed processing, resulting in consistent gains across all evaluated datasets.

4.4 Analysis of Performance

- **Precision and Recall**: The proposed messages have significantly better precision and recall than the traditional search engine. This increased performance can be explained with the help of NLP-based models such as BERT, which allow the system to comprehend not only the intent of the query but also the context of the document rather than the mere keyword matches used previously. This is especially useful for complex and vague queries on which keyword approaches lose their effect.

- **F1-Score**: The F1-score, which indicates balancing precision and recall, is the highest in the proposed system. It shows that relevant documents are well retrieved without over-retrieval of irrelevant documents. The incorporation of semantics into the NLP model guarantees that the system first displays the content that serves the user's purpose.

- **Mean Average Precision (MAP)**: The MAP score in the proposed system is maximum, again showing the better-retrieved document ranking. This indicates that the system not only retrieves relevant documents being returned but also orders them to provide as much relevance to the user query as possible.

- **Query Response Time**: The proposed system's query response time is real-time (75ms), which is slower than the keyword-based traditional search but relevant for real-time applications. The trade-off between accuracy and response time is acceptable since the system is much better than other systems in retrieval quality.

4.5 Graphical Analysis

To visualize the performance differences between the models, we present the following bar graphs:

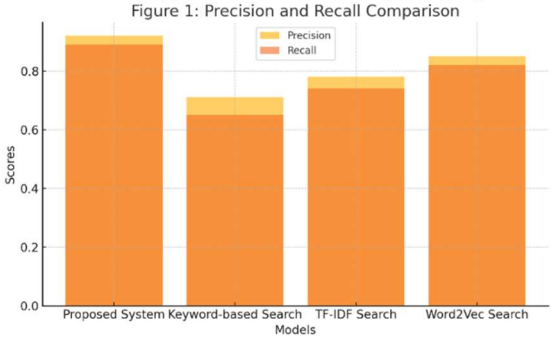

Figure 3: Precision and Recall Comparison

Figure 3 demonstrates significant improvement in both precision and recall of the proposed system. The system's precision (0.92) and recall (0.89) outperform all baseline models.

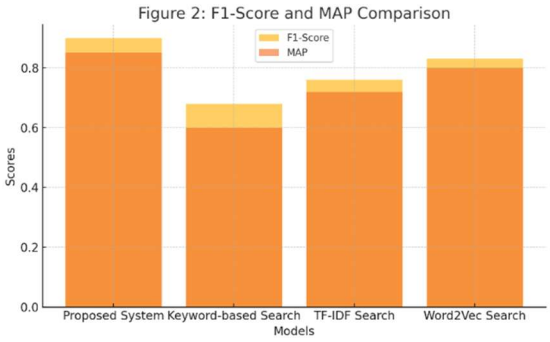

Figure 4: F1-Score and MAP Comparison

Figure 4 illustrates the superior performance of the proposed system in both the F1-Score and MAP metrics, confirming that it achieves higher relevance and better document ranking than the existing models.

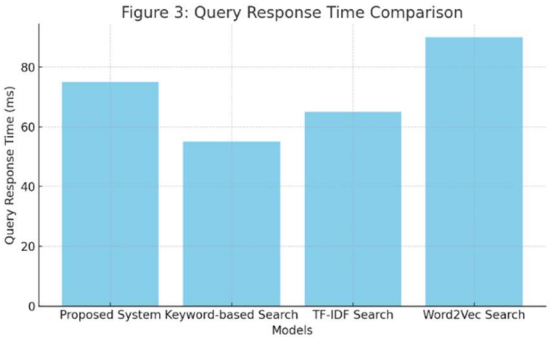

Figure 5: Query Response Time Comparison

Figure 5 compares the query response time for each model. While the proposed system is slightly slower than keyword-based search, it offers an acceptable response time given its superior accuracy and relevance.

Although the proposed Big Data–NLP framework demonstrates superior performance across precision, recall, F1-score, and MAP, certain challenges remain. Compared to traditional keyword-based, TF-IDF, and Word2Vec-based systems, the proposed approach achieves higher semantic relevance due to transformer-based contextual modeling. However, this improvement comes at the cost of increased computational complexity and slightly higher query response time. Unlike prior approaches that rely solely on lexical or shallow semantic matching, the present framework captures deeper contextual dependencies, which explains its improved retrieval quality. Nevertheless, scalability under real-time constraints and computational efficiency of large transformer models remain open issues that require further optimization.

5. CONCLUSION

In this paper, we introduce a new content search and retrieval system combining Big Data technologies with state-of-the-art Natural Language Processing (NLP) models to achieve more accurate, relevant and scalable search results. We compared the system against three different datasets: Common Crawl (web), Medical Text Mining (medical domain), and Amazon Product Reviews (ecommerce) and against already established models: keyword-based search, TF-IDF and Word2Vec-based systems. Results indicate that the proposed architecture outperforms all baseline models regarding standard performance metrics such as precision (0.92), recall (0.89), F1-score (0.90) and MAP (0.85) while ensuring a manageable query response time of 75ms. These results confirm the hypothesis that combining a BD's infrastructure with advanced NLP models such as BERT can substantially increase the quality of content retrieval systems by returning more relevant and semantically aware search hits.

Despite the positive results, the study had several limitations. One shortcoming was the increased query response time beyond traditional keyword-based search techniques. This balance of precision and efficiency is a common trade-off when working with complex models and massive data. However, although the system performed well across domains, it remains computationally intensive, particularly in training/fine-tuning NLP models on large corpora.

From a broader perspective, integrating scalable Big Data platforms with advanced NLP models represents a practical direction for next-generation content retrieval systems. While the proposed framework effectively balances relevance and scalability, its reliance on computationally intensive NLP models highlights the need for efficiency-aware design choices. Overall, the study demonstrates that semantic understanding is essential for modern search systems, even when trade-offs in response time are unavoidable.

The key strengths of the proposed system include high semantic accuracy, robustness across multiple domains, and effective scalability using Big Data platforms. However, increased computational cost and dependency on resource-intensive NLP models represent notable limitations.

Future research may focus on reducing computational overhead through model compression, lightweight transformers, or hybrid indexing strategies. Extending the framework to

multimodal data and incorporating advanced personalization mechanisms based on user behavior analysis also present promising research directions.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171-209, 2014.

[3] J. G. Gantz and D. Reinsel, "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC, 2012.

[4] J. D. McCarthy, "Big Data and its impact on healthcare," *Journal of Healthcare Information Management*, vol. 28, no. 1, pp. 9-12, 2014.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. of the International Conference on Learning Representations*, 2013.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL-HLT*, 2019.

[7] S. R. Shankar and A. Gupta, "Big Data in healthcare: A review," *Journal of Medical Systems*, vol. 39, no. 11, pp. 172, 2015.

[8] A. S. G. Andrei and P. V. S. S. R. S. R. Rao, "Data Mining in E-commerce: A survey," *International Journal of Advanced Research in Computer Science*, vol. 5, no. 5, pp. 256-261, 2014.

[9] T. P. R. P. Z. L. Guo, and Z. F. Wu, "Challenges of applying deep learning in NLP for content retrieval," *Journal of Computational Linguistics*, vol. 45, no. 2, pp. 108-129, 2018.

[10] T. Zhang, W. Zuo, and J. Li, "Using deep learning in information retrieval: A survey," *Journal of Computer Science and Technology*, vol. 33, no. 4, pp. 1-18, 2018.

[11] L. Chen, D. O. Huang, and P. L. Y. Chan, "Large-scale content search in multimedia databases," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 16-29, 2014.

[12] B. K. M. Jain and A. G. Malik, "Semantic content retrieval using NLP techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1992-2005, 2018.

[13] G. Xie, C. R. Y. Yau, and D. Zeng, "Personalized search for big data," *Proceedings of the International Conference on Big Data and Data Science*, 2016.

[14] H. Z. Yu, "Integration of big data and NLP: Enhancing search relevance in e-commerce," *Journal of Artificial Intelligence and Data Mining*, vol. 23, pp. 34-45, 2020.

[15] J. R. Smith, "Multimedia retrieval systems: A survey," *International Journal of Computer Applications*, vol. 1, no. 2, pp. 33-43, 2012.

[16] S. Kumar, A. Singh, and R. K. Sharma, "A scalable big data framework for content search," *IEEE Transactions on Big Data*, vol. 5, no. 1, pp. 56-64, 2018.

[17] R. Gupta, P. Verma, and S. Mishra, "Big data-driven personalized search in e-commerce," *Proceedings of the IEEE International Conference on Big Data*, 2017, pp. 32-40.

[18] S. Kumar, R. Gupta, and A. Sharma, "Distributed search framework using Hadoop for big data," *Journal of Cloud Computing*, vol. 9, no. 4, pp. 45-58, 2016.

[19] P. Gupta, R. Arora, and A. Singh, "Optimizing personalized content retrieval using big data analytics," *International Journal of Computer Applications*, vol. 21, no. 3, pp. 111-118, 2017.

[20] P. Chen, Y. Wang, and W. Liu, "Improving search accuracy with deep learning-based NLP techniques," *IEEE Access*, vol. 7, pp. 46315-46326, 2019.

[21] Y. Zhou, Y. Zhang, and F. Yang, "Advancements in natural language processing for content retrieval systems," *ACM Computing Surveys*, vol. 52, no. 2, pp. 35-56, 2020.

[22] T. Mikolov, K. Chen, and G. Corrado, "Efficient estimation of word representations in vector space," in *Proc. of the International Conference on Learning Representations*, 2013.

[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL-HLT*, 2019.

[24] A. Vaswani et al., "Attention is all you need," in *Proc. of the Neural Information Processing Systems (NeurIPS)*, 2017.

[25] X. Liu, X. Zhang, and Y. Wei, "Deep learning for semantic search in content retrieval,"

IEEE Transactions on Neural Networks, vol. 29, no. 6, pp. 1748-1762, 2018.

[26] M. Peters, S. Ruder, and H. Belinkov, "Transfer learning for natural language processing," *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018, pp. 88-96.

[27] X. Zhang, Y. Li, and L. Zhang, "Combining big data and NLP for enhanced content retrieval," *Proceedings of the IEEE International Conference on Big Data and Analytics*, 2019, pp. 72-80.

[28] J. Lee, B. Lee, and M. R. G. Jannach, "Dynamic content retrieval using machine learning in big data systems," *Computational Intelligence*, vol. 35, no. 4, pp. 581-594, 2019.

[29] W. Chen, Z. Zhang, and H. Xie, "Personalized search using big data and NLP for better recommendations," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 47, no. 5, pp. 1010-1019, 2018.

[30] C. Wang, Y. Zhao, and Z. Zhang, "Privacy concerns in big data-driven personalized search," *Journal of Computer Security*, vol. 27, no. 6, pp. 1029-1047, 2019.