[1] Hemendra Shanker Sharma
[2]Ashish Sharma

Query Expansion Information Retrieval using Customized Ontology Technique

Journal of Electrical Systems

Abstract: - Information from online archives is now much more widely used and accessible than before. As a result, searching becomes more challenging and time-consuming. This vast data utilization is the focus of a significant area of research called information retrieval (IR) systems. The goal is to reduce retrieval time while also maintaining and improving answer relevancy. To solve the issues stated, it is necessary to provide an IR Model. The provided method takes care of indexing, similarity keyword extraction, semantic similarity, updating historical data, and updating. The effectiveness of the suggested strategy and the current methods are compared in terms of performance with distinguished parameters. A novel similarity estimate approach is used to group the texts and determine how similar works are based on the obtained score. It is compared to the existing methods to evaluate the findings and shows the usefulness of the suggested model on the scales of accuracy, mean absolute error, precision, recall, sensitivity, and specificity. The improvement of personalization performance, which is an update based on historical knowledge, is one of the main goals of this effort. The Impact Score Estimation technique is used to improve data extraction using semantic keyword extraction and indexing. To cluster documents based on computed scores, the algorithm is to evaluate similarity estimation which can improve searching by speeding up information retrieval and processing. Decision tree classifiers provide better results for class 3 which is 0.93. While the micro average ROC curve generates 0.87 accuracy.

Keywords: Information retrieval, Natural Language Processing, Query Expansion, Ontology, Similarity index.

I. Introduction

Huge data is generated in an integrated environment from multiple sources. The generated data is housed in a repository. The need to retrieve context-specific relevant data is a challenging need of time. Due to the rapid expansion of data resources on the internet in a variety of formats, information retrieval (IR) techniques are growing in popularity [1]. The primary goal of an IR system is to locate pertinent data and documents that satisfy user needs. As searching for documents through the internet have become an imperative part of people's life, there is a great demand for obtaining such documents from a huge source of information which was appropriate to the information referred [2]. Searching for information and documents from a Web repository through a search engine gives many a times irrelevant pages. This is because search engines solve queries based on text pattern matching. This may sometimes output relevant as well as matched but irrelevant information [3].

Semantic web technology plays a significant part in retrieving relevant information. Semantic search tries to understand the user intention and improves the search accuracy and in turn, appears in searchable data space.[4] Ontology plays a vital role in Semantic Web Technology. The component of the Semantic Web referred to as Ontology provides a common understanding of a particular term along with its relationship with the other terms in the query.[5] The search domain's ability to semi-automatically adapt to domain evolution is represented by the adaptive ontologies, which represent both the search domain and user domain. The data found in ontologies were used to enhance the semantics of submitted queries as well as online data. Accordingly, the system automatically uses the chosen ontologies to enrich the produced query by the query enrichment criteria. Based on web data found in the graphs that validate the enriched query, the domain ontology is modified [6].

The vision of the Semantic Web proposes a situation where the information and administrations on the Web can be semantically translated and handled by machines to encourage human utilization. The semantic Web depends vigorously on the formal ontologies that structure basic information with the end goal of far-reaching and transportable machine understanding [7]. Semantic Web innovation depends on metaphysics as an instrument for displaying a conceptual perspective of the genuine and relevant semantic examination of documents [8]. Accordingly, the accomplishment of the semantic Web will be subject to the expansion of ontologies, which

[1] *Corresponding author: Department of Computer Engineering & Applications, GLA University, Mathura, Uttar Pradesh, India, hss.agra@gmail.com

2 Department of Computer Engineering & Applications, GLA University, Mathura, Uttar Pradesh, India, ashish.sharma@gla.ac.in

requires quick and simple designing of philosophy and evasion of a learning procurement bottleneck. However, there are certain challenges to which knowledge can or should be formalized [9].

First, as a result of the gigantic measure of data now accessible to data frameworks worldwide as unstructured content and media archives, changing this aggregate of data into formal ontological learning at a reasonable rate is presently an unsolved issue.[10] Second, the Boolean model (Classical IR Model) does not scale for search which returns a huge result [11]. It does not afford the facility of clear ranking criteria. The fundamental objective of these theses is to achieve an Ontology-Based Information Retrieval Model that can support semantic similarity and retain the view of approximate search in a document repository [12]. Additionally, this research investigates how to offer an adaptive update model for information retrieval, a semantic search model for the user's given query, and how to develop rank-based concurrent history and knowledge update models. Increasing precision and recall is the goal of ontology-based semantic web information search.

To improve Precision, Query Expansion is a more promising technique [13]. The goal of query expansion is to increase the effectiveness of information retrieval operations by presenting additional documents that match the search criteria of the query. [14,15] Here, expansion terms depending on the initial query terms are inserted using the seed query reformulation technique [16]. The two categories of classic query expansion techniques are log-based QE and automatic relevance feedback, which includes pseudo-relevance feedback (PRF) [17]. There are two problems with many information retrieval systems: query growth and poor retrieval performance. There are just a few term relationships, some of which might be false expansion terms that create topic drift, as well as a few terms that are unique to the query set and do not appear in the document set for a particular quantity [18].

However, considering the technique used in the retrieval model, a Novel Information Retrieval Framework is projected to improve the effective access to online available data. Semantic keyword extraction and indexing are used to extract keywords and carry out later alterations [19]. The keyword is then given an Impact Score Estimation to determine its prominence in each. The similarity of the documents is then evaluated using a unique similarity estimate technique to cluster the documents based on the derived score.

The contribution of this article is as follows:

- To explore Semantic Web Information Retrieval for Enhancing Search Significance with distinguished parameters.
- Enhancement in Data extraction using Semantic Keyword Extraction and indexing is implemented with the Impact Score Estimation model.
- The Novel Similarity Estimation Algorithm is used to compute document similarity to cluster the documents based on computed scores, which can improve search by speeding up processing and information retrieval.
- Improving the aspect of personalization to protect the privacy of the users who are querying, a novel privacy preservation algorithm is developed.
- For user-specific, efficient, relevant document retrieval, the User History and Relevant Search are kept on file.

The rest of the work can be organized as follows; Section II is focused on the literature review and tries to identify the gaps in the work. Section III covered the proposed methodology and overall system architecture. This section also covered the proposed algorithm for optimized query expansion. Section IV describes the evaluation strategy that explains the dataset and estimated parameters. Section V focused on the results and discussed the information extraction accuracy. Finally, the conclusion and future scope are described in section VI.

II. Literature Review

The review of this article focused on the existing methodologies of the information retrieval process, Pre-processing, Indexing, Ranking, and ontology-based semantic mining. The focus of the distinct methods that were utilized by the authors for the information retrieval. Authors [20] studied the basic concepts of Information Retrieval through Semantic Web. That was essential to retrieve the appropriate information from the bulk web documents. The probabilistic model ranks the retrieved documents using probability-relevant information. Carpineto et al. [21] survey work briefly explained the indexed techniques namely Signature file and Inverted index. In the signature-indexed approach, the bit is appended to the document which can be called a signature. Singh et al. [22] proposed a concept-based paragraph vector formation for effective information retrieval. This semantic vector is formed by utilizing the recurrent neural networks (RNN) with Long Short-Term Memory (LSTM) cells. The meaning of the words in a sentence was extracted and entrenched in the semantic vector. Dahir et al. [23] introduced a conceptual similarity-based semantic algorithm to narrate attributes. This paper develops an algorithm based on similarity comparison in the domain of attribute retrieval.

Kumar et al. [24] familiarized a method for pulling out the social network to be combined with existing tracking methods of social networks using URLs. The various relations such as strength relations and relations based on the online academic database are extracted from social networking. The strength of heterogeneous documents is analyzed based on their relations in the social network. Values are estimated for every individual or group related to their interest and implementation. Rashad et al. [25] recommended an idea to design NLI-GIBIR and developed a novel framework that could be appropriate for graph-based bibliographic information retrieval systems. The methods for comprehending and analyzing natural language bibliographic queries were integrated into this innovative framework. The system considered natural language queries as the input and output yielded accurate replies. By converting the input into a database query language, this was made possible.

ALMarwi et al. [26] suggested K-Means clustering depending on Particle Swarm Optimization (PSO). PSO extends the clustering approach in a global search manner and provides optimal centroids which helped in generating more compact clusters with improved accuracy. PSO optimization is adaptable to the changing database environment. The new implementation with PSO was the MapReduce methodology. The document terms were extracted after pre-processing of the data document. According to the Term weights Document Term Matrix (DTM) was formed based on Document vector representation. Padaki et al. [27] proposed several approaches in the literature to extract useful information from the data. Information management works with both structural and non-structural data, whereas in the information retrieval process, non-structural data is used to search for information. In IR, formal expressions known as queries are used to search the web for pertinent information. While retrieving multiple pieces of information that fit the query with varying degrees of relevancy, the query in this case does not specifically designate a single answer.

Jain et al. [28] proposed the use of the Latent Semantic Indexing (LSI) approach in 1990. This method is especially effective and is applied in the Probabilistic LSI and Singular Value Decomposition (SVD) algorithms. The terms are automatically arranged according to their meanings using LSI, which first generates a semantic space and then maps each phrase into it. The LSI method just makes it hard to control how much the queries expand, and the expanded searches contain a lot of extraneous terms. Raza et al. [29] proposed using the local co-occurrence method for expansion; Utilizing the document collection's word frequency, expansion was carried out. This technique has the potential to boost IR effectiveness by 6 to 13%. Simply said, this approach has not been able to show how the term is connected and what it means. Hameed et al. [30] research suggested query expansion in Urdu information retrieval. However, query expansion only marginally raises the Mean Average Precision (MAP) value by 22–24% when applied to the Kullback-Leibler model.

Meng et al. [32] used the ULMS Meta thesaurus (2015) and offered QE. Meta Map is used to convert user queries or words into UMLS CUIs. The MRCONSO Meta thesaurus table then lists the synonyms of the user words or queries as well as the terms that were used in expansion queries. Simply put, employing a Meta thesaurus on several user queries, and enlarged searches slows down IR. Additionally, Ojha et al. [31] used WordNet to locate synonyms for user-entered words. By determining the Part of Speech (POS) of each word before processing, the query expansion procedure is carried out. Then, using WorldNet, synonyms are found for each term to broaden the query. According to the study's findings, using expansion queries increased precision and recall by about 40% and 24%, respectively, over not using them. Authors proposed two stages for the QE strategy that is used, the first involves using the Hopfield word network's recursive structure, which is most connected to other words chosen, to significantly reduce overweighting by grouping terms on queries based on semantic correlations. WordNet is utilized by the word to aid in candidate extraction. MAP usage increased from 4% to 12% in the CACM and CERC collections used in the evaluation [33] [34]. It is simply that the drawn-out inquiry thwarts the IR's presentation and its ability to give word associations when WordNet/Meta thesaurus is used on specific client questions.

The primary research projects that have been undertaken include context-sensitive retrieval, query expansion to improve the effectiveness of information retrieval, semantic keyword extraction, and impact score estimation to achieve the keyword in the document to determine the importance of the keyword in each document. Therefore, a fuzzy semantic-based keyword searching method is required for retrieving the information effectively with increased accuracy. Also, the history-based and knowledge-updated retrieval processes are used to improve the system's performance. The current research has some benefits such as reduced processing time, high accuracy in retrieving the information, and less time taken for the retrieving process.

III. MATERIALS AND METHODS

The proposed method is used to retrieve the information and prevent the overwhelming restrictions in the prevailing model. Initially, load the set of datasets to the proposed retrieval techniques. Then, extract the keywords from the dataset. Based on the extracted keywords indexing method is performed. For every keyword, find the relevant documents and update the corresponding keywords with index terms in the knowledge database. Thus, the extracted keywords with corresponding documents are located. Then, the user gives a query as input and it is pre-processed for keyword extraction. Every user's query is maintained in a database. Then, the current query and similar queries in the history (remaining user's queries) are analyzed. If a query is exactly matched means the user gets the result immediately. Otherwise, semantic search is done with the help of a knowledge database [35].

Based on the similarity values documents are retrieved. The most recent user's searches are compared to the papers that have been retrieved, and the documents are then given rank values. After that, update the database's knowledge and history. It will improve the precision of information retrieval that is semantically based. The information retrieval system's proposed design for semantic keyword matching is shown in Figure 1. The suggested scheme considers phenomena like keyword matching, semantics, and query pattern evaluation. The suggested matching method combines the adaptability of keyword-based retrieval with the semantic keyword matching algorithm, which offers a means for semantic level matching. The ability of the proposed technique to be improved via this research is due to the query and Metadata typical dataset of the semantic search algorithm.

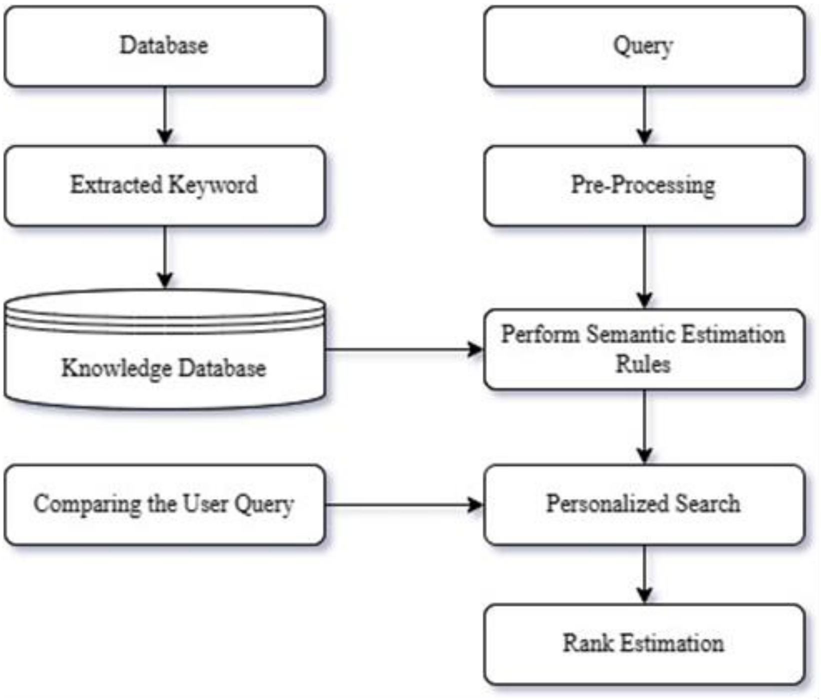


Fig.1. Overall Methodology of the work

The main intention of ontology-based semantic web information search is to maximize accuracy and reduce the search time. Figure 2 shows the Semantic keyword-matching architecture. The proposed model accuracy is improved by using various elements as follows.

A. Pre-processing

The pre-processing is the preliminary stage of the information retrieval technique. Unstructured data contains enormous information about the search and must be reduced to limited content. The unstructured dataset is the required source of the technique which is in the arrangement of text records. Then pre-process the dataset by removing the stop words and stemming.

i) Elimination of Stop-words

Articles, prepositions, and conjunctions are regular applicants for a gradient of stop-words. Removal of stop-words is a significant benefit. It decreases the extent of the indexing arrangement significantly. It is run off the elimination to get a pressure in the measure of the ordering structure of at least 40% exclusively with the end of stop-words.

ii) Stemming

The stemming used the part-of-speech (POS) tagger in tandem with a regular appearance for identifying the utmost noun phrases to produce the term list. Besides, stemming has the secondary effect of decreasing the size of the indexing arrangement as the sum of individual index terms is minimized. Inheritor variability stemming is established on the resolution of morpheme restrictions, customs information from organizational semantics, and further multifaceted than affix elimination stemming mechanism.

iii) Domain Analysis

The domain exploration of the search info system to retrieve the corresponding information. The specific domain is selected and then perform pre-processing stage to retrieve domain-related information. The specification of a representative term for a mutual area of consultation such as functions, relations, classes, and other objects is the ontology representation [36].

iv) Indexing Method

The method of storing data for effective recovery in response to a search query is known as indexing. A search engine keeps track of all the topics discovered during the swarming process and puts them in a directory for easy content retrieval. There are three major types of indexing methods are as follows,

- Forwarded Index Method - The word list for the individual file will be stored.
- Inverted Index Method - The document list for the respective term will be retained in this index.
- Graph Indexing Method - Provided a query graph that shows the index term besides extracting the set of answers. It contains the query graph and returns the query results which validate the graph.

In this work, each index word is analyzed and stored in the knowledge database for retrieving the relevant document based on the search model.

B. Query Pattern Evaluation

The user gives a query as the required input of the retrieval system. In this flow, the user query is pre-processed by eliminating the stop words and stemming and formerly analyzing the current user query with the remaining user queries. If the user query matches with the existing queries, then immediately retrieve the relevant documents. Or else, a semantic search process is implemented [37].

Making queries is important to obtain the required accuracy and recall. Because there are so many relevant documents to consider, accuracy is frequently more important for web search engines than recollection. Due to the high quantity of records in the web environment, many documents that are frequently relevant to the inquiry cannot be further graded based solely on the interior topographies of the document.

C. Semantic Keyword Search Model

The goal of the proposed approach is to take care of the semantics and descriptions of the web service, which are not handled by the existing approaches. The web service ontology framework is expanded in this suggested method to create an external database and implement the matching algorithm to produce accurate results for web services. The keyword matching, semantic, and ontology algorithms are phenomena that are included in this study's matching engine. With the restriction that it cannot comprehend the user's objectives, keyword matching is used to swiftly deliver the results.

In this method, a search engine first creates a preliminary set of ranks, from which users choose the pertinent documents. The technique used to define the ranking of search query results is called grading. Matching and ranking are used in the search, and when matching is used, a subset of the components is chosen to be tallied. By using some important judgments, the ranking establishes the level of matching. When syntactic or semantic charting is used, the ranking is obtained. The score of the websites will determine how the rank is determined. The following are more thorough explanations of the suggested ontology-based semantic web information retrieval approaches.

- Syntactic Ranking Method: The search relies on term matching between the engine database and the query.
- Semantic Ranking Model: This is dependent on the outcome significance that attained the linking gap among the syntax and semantics that provides a concentrated outcome and improved gratification to the user.

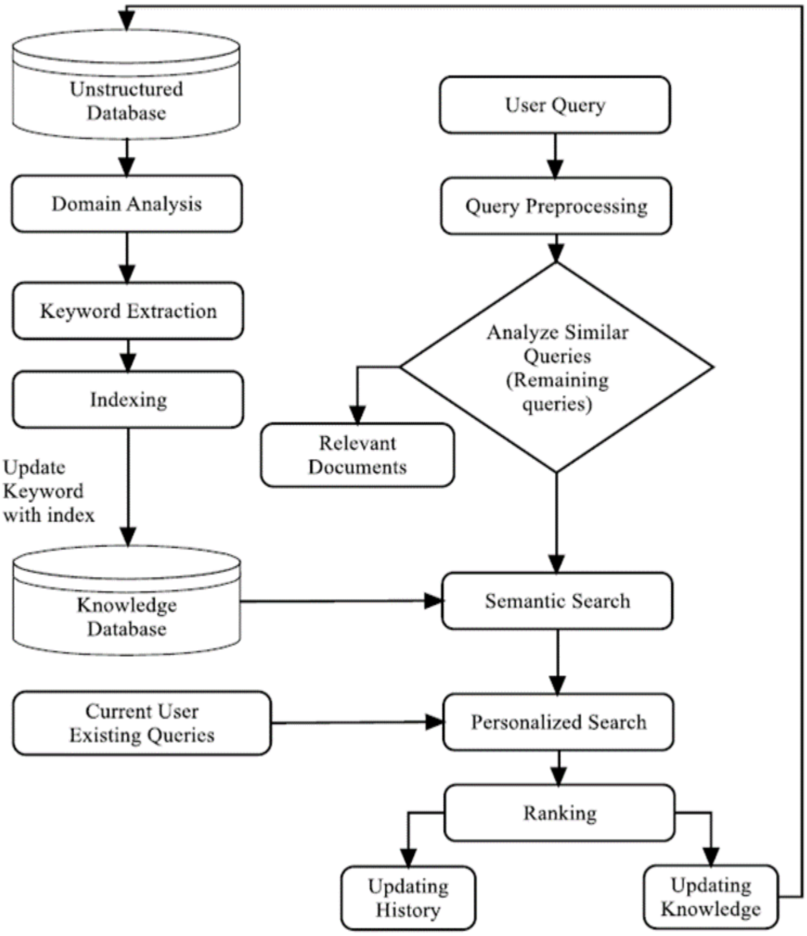


Fig.2. Proposed Flow diagram for updating history and knowledge

D. Proposed Keyword Matching and Query Evaluation Algorithm

```
QueryExpansionAlgorithm (D, K, F, Index)
D=Dataset upload
PD= Processed Dataset
Start loop I 0 to D:
    Start loop J 0 to D:
        PD=Stemming and Stop words
    End loop
End loop
Initialize the array of Distinguished Keywords
Start loop I 0 to D:
    Start loop J 0 to D:
        Update the array of Distinguished Keywords
    End loop
End loop
F [] = File IDs
Start loop I 0 to D:
    Start loop J 0 to D:
        F [] = update File IDs
    End loop
End loop
Index [] = 0
Start loop I 0 to K:
    Start loop J 0 to PD:
          Start loop K 0 to D:
                 Index [] = Update Indexes
        F [] = update File IDs
         End loop
    End loop
End loop
Start loop I 0 to Query_list:
       Q = Stop words and Stemming
      H = History
     Sk = Semantic Keywords
    Mi = Matched Value
End loop
Cosine = Similarity for Cosine
Start loop I 0 to K:
    Start loop J 0 to Mi:
          Cosine = $\sum_{l=1}^{n} P_l Q_l / \sqrt{\sum_{l=1}^{n} P_l^2} \sqrt{\sum_{l=1}^{n} Q_l^2}$
    End loop
End loop
PS = Initiate Personalized Search
Start loop I 0 to K:
    Start loop J 0 to Mi:
          PS = Cosine = $\sum_{l=1}^{n} P_l Q_l / \sqrt{\sum_{l=1}^{n} P_l^2} \sqrt{\sum_{l=1}^{n} Q_l^2}$
    End loop
End loop
Start loop I 0 to K:
    Start loop J 0 to Mi:
          PS = Cosine = $\sum_{l=1}^{n} P_l Q_l / \sqrt{\sum_{l=1}^{n} P_l^2} \sqrt{\sum_{l=1}^{n} Q_l^2}$
    End loop
End loop
Start loop I 0 to PS:
    Start loop J I+1 to PS:
          If PSj > Psi:
            Exchange Psi, Psj;
          End if
    End loop
End loop
Updation of Semantic Knowledge
```

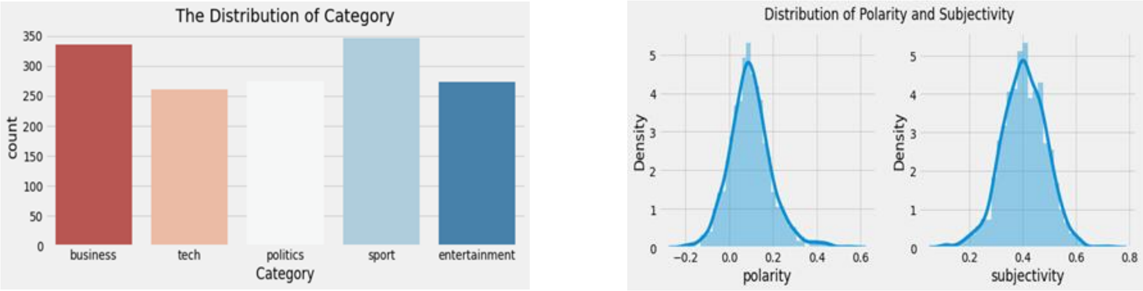
Making keywords into semantic keywords is the initial stage in the acquisition of the response's semantics. The suggested method uses a collection of ontologies in this approach to abstract the potential meanings of each keyword while combining similar meanings to avoid duplication. The structure then connects various disambiguation techniques to determine each keyword's relevance based on its context and the potential meanings of the remaining keywords. Additionally, the knowledge database indexes and stores the documents that were analyzed based on the given keywords. The query is treated as input as the user interpolates it and is pre-processed by removing all stop words and stemming.

The current user query is compared to past queries; if there are any matches, the relevant document is immediately obtained; otherwise, the current query is used as the input for semantic search. Additionally, the Wordnet library is used to perform the semantic search to identify the synonyms for each pre-processed query. Then, the user query is matched with the documents depending on the cosine similarity rule. After computing, the retrieved forms are not matched with the query to adopt the personalized search and update the results in the history. If the user is not fulfilled with the recovered documents, authors can introduce the answer as input to the database, and this is designated as knowledge updating.

IV. Prepare Your Paper Before Styling

A. Dataset Description

The dataset used for the current work is the BBC news dataset in this article for analysis purposes which is comprised of 2225 articles. The dataset is divided into five distinct categories such as politics, entertainment, tech, sports, and business as shown in figure 3a. The total dataset is segregated into train, test, and sample solution CSV files. There are three attributes present in the dataset the information article field, text or news, and category. It checked the polarity (strengthen the opinion) and subjectivity (aims to remove factual content) shown in figure 3b.



(a) Category Distribution (b) Polarity and Subjectivity

Fig.3. Dataset Distribution

The scatter plot among polarity and subjectivity has been depicted in figure 4. It shows the most polarity in the center while subjectivity is spread out among the graphs. That reveals that the dataset behavior is belonging to the different factual content with a limited number of opinions. The polarity lies between -0.2 to +0.5 and subjectivity lies between 0.1 to 0.7 which indicates maximum news belongs to the neutral behavior. The line plot for subjectivity and polarity shown in Figure 5 represents the frequency of each point on a line in case of subjectivity and polarity.

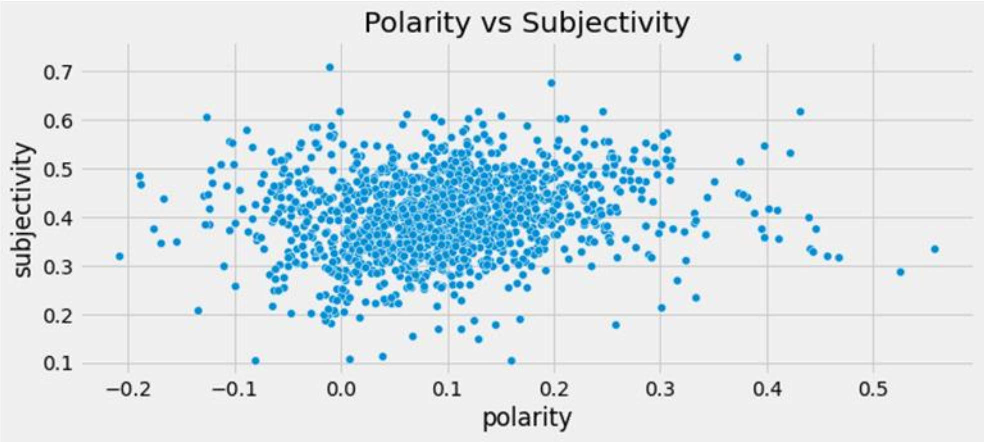


Fig.4. Scatterplot for polarity and subjectivity

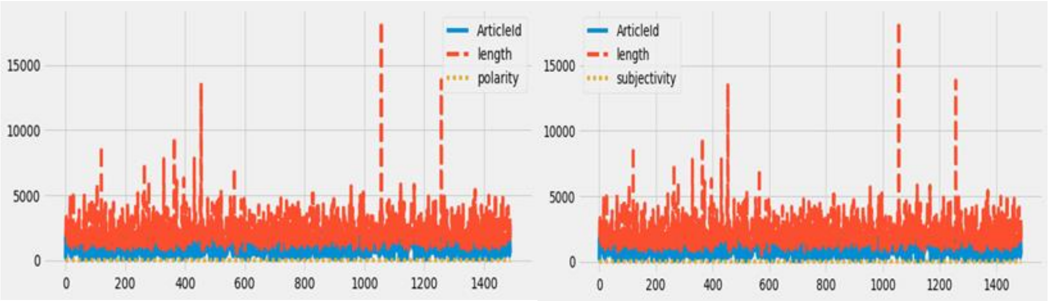


Fig.5. Lineplot for subjectivity and polarity

Figure 6 depicts the word cloud retrieved from the processed dataset. It is a pictorial representation of commonly used words that play an important role in the classification or categorization of the datasets. Some words that are present in the center indicate the most frequently used in the dataset

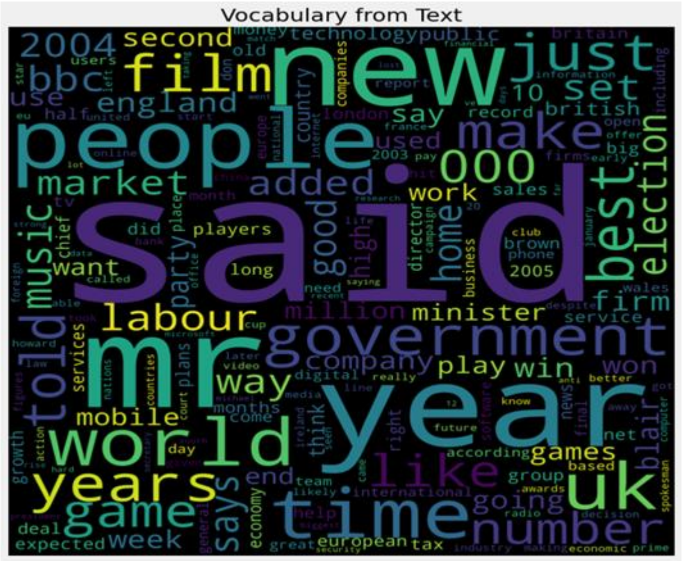


Fig.6. Word cloud retrieved from dataset

B. *Estimation of Parameters*

To evaluate the model performance, it utilized distinct parameters such as accuracy, validation losses, receiver operating characteristics curve (Roc) curve, precision, F1-score, R2-score, recall, time delay, and mean absolute error (MAE). These parameters are evaluated by using sub-parameters of the confusion matrix. The confusion matrix is achieved to define the performance of the algorithms such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These performance parameters are assigned in terms of matrix and the ratio of the distinct combination of the parameters (Equations 1 to 6).

$$Accuracy = TP + TN/(Total\ predictions) \tag{1}$$

$$MAE = \sum_{i=0}^{n} |\hat{y}_i - y_i| \tag{2}$$

$$Precision = {}^{TP}\!/_{(TP + FP)} \tag{3}$$

$$Recall = TP/(TP + FN) \tag{4}$$

$$F1 - score = {}^{TP}\!/_{(TP + 0.5(FP + FN))} \tag{5}$$

$$R2 - score = 1 - \frac{SS_{res}}{SS_{tot}} \tag{6}$$

Where SS_{res} defines the residual sum of squares and SS_{tot} indicates the total sum of squares

V. Discussion

This section is focused on the result analysis measured from the distinct machine or deep learning algorithms. The proposed model is integrated with the proposed query expansion algorithms and existing mechanisms. Figure 7 shows the accuracy generated by the BiLSTM algorithm with validation accuracy, loss, and validation loss. It can be observed from the figure that accuracy is about constant after 15 epochs while the number of losses is more when epochs are less. The LSTM model produced better accuracy for twenty-five or more than twenty-five epochs with fewer training losses as shown in figure 8.

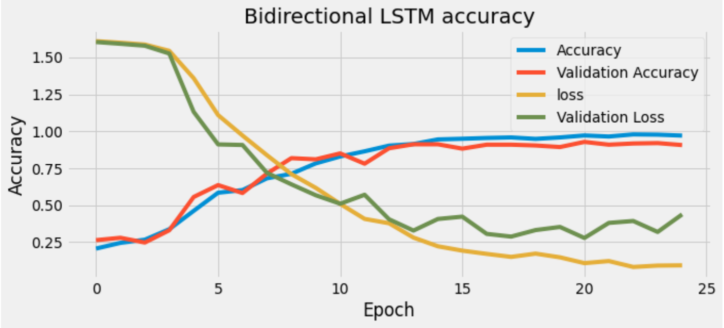


Fig.7. Accuracy and other parameter evaluation for Bidirectional LSTM

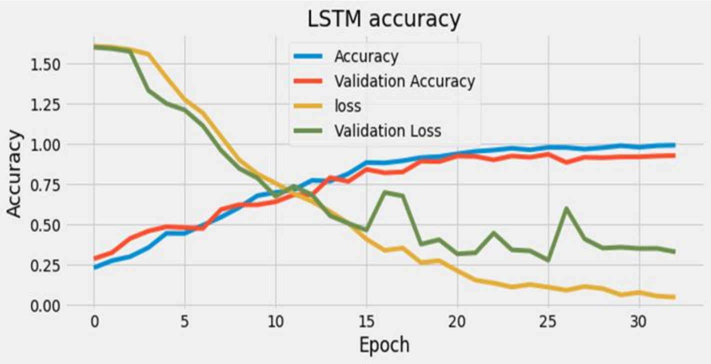
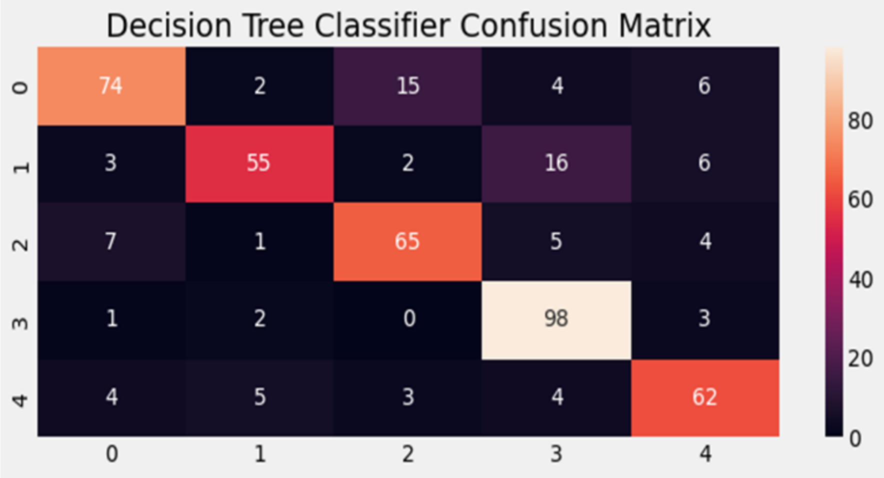


Fig.8. Accuracy and other parameter evaluation for LSTM

The model has evaluated the confusion matrix for the decision tree classifier for five distinct categories that are identified in terms of class 0 to class 4. The confusion matrix shows the correlation among the classes as shown in figure 9. The ROC curve is used to depict the performance of the classification models. Decision tree classifiers are providing better results for class 3 which is 0.93. While the micro average ROC curve generates the 0.87 accuracies shown in figure 10.


Decision Tree Classifier Confusion Matrix

	0	1	2	3	4
0	74	2	15	4	6
1	3	55	2	16	6
2	7	1	65	5	4
3	1	2	0	98	3
4	4	5	3	4	62

Fig.9. Confusion matrix for Decision tree classifier

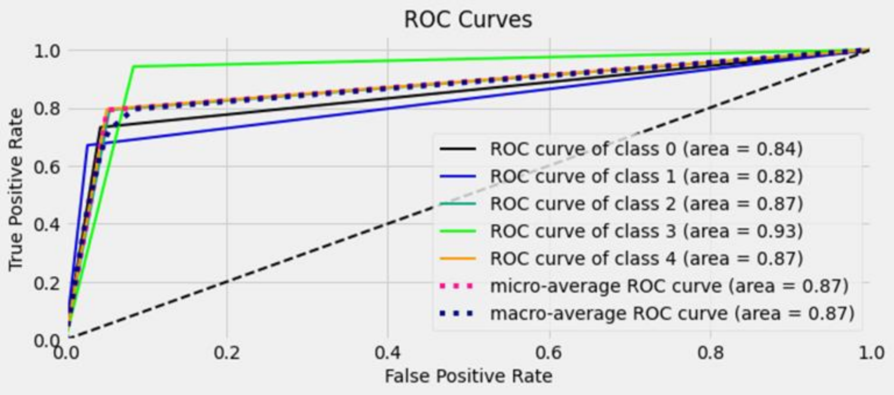


Fig.10. ROC curve for distinct classes with Decision tree classifier

The confusion matrix for logistic regression is shown in the figure 11 with distinct attributes. That depicts the high correlation between class 0. The ROC curve shows better results for class 3 as well which is 1.00. The micro average ROC curve generates the 0.98 accuracies shown in the figure 12.

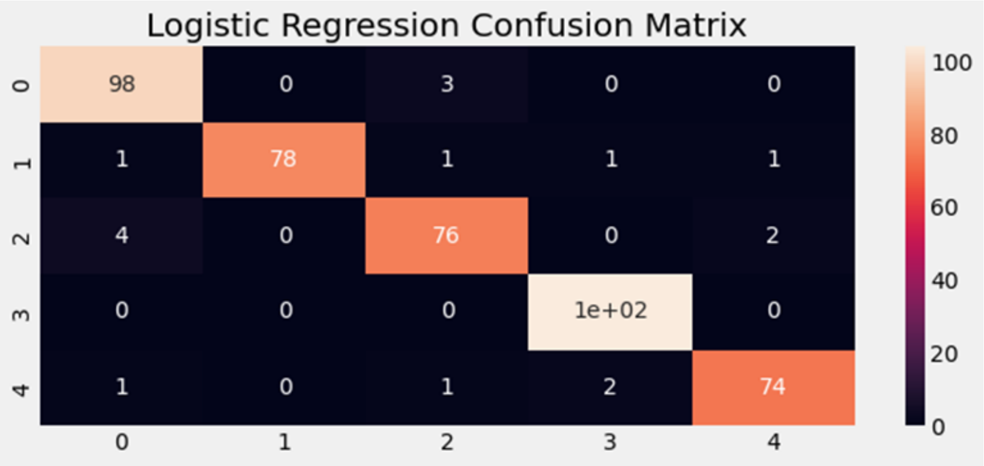


Fig.11. Confusion matrix for Logistic Regression

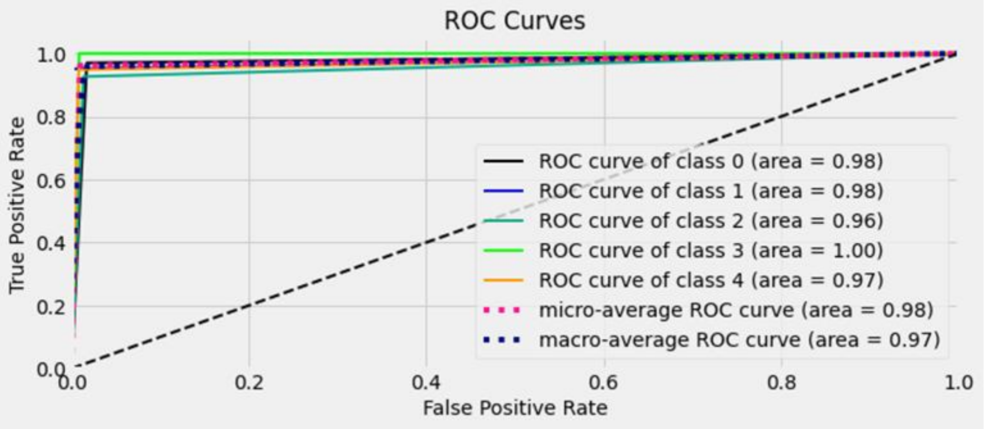


Fig.12. ROC curve for distinct classes with Decision tree classifier

The confusion matrix for support vector machine (SVM) is shown in the figure 13 with distinct attributes. That depicts the high correlation between class 0. The ROC curve shows better results for class 3 as well which is 1.00. The micro average ROC curve generates the 0.97 accuracies shown in the figure 14.

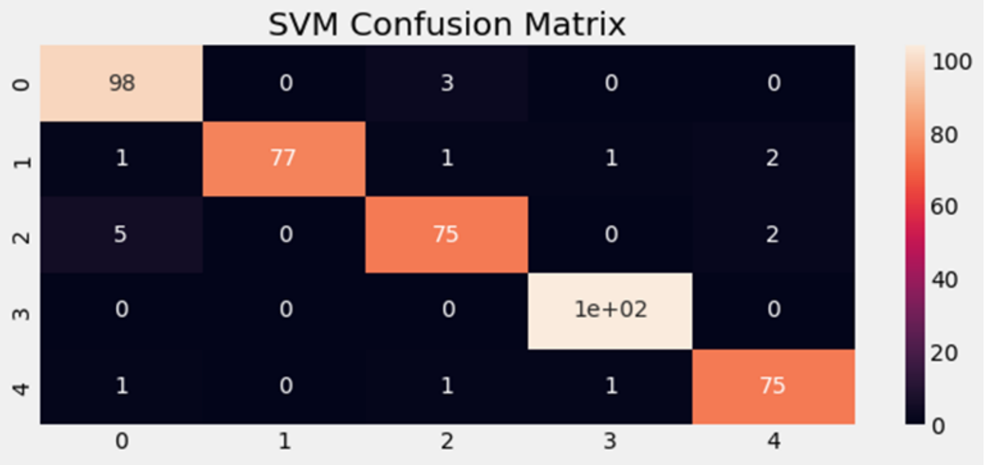


Fig.13. Confusion matrix for SVM classifier

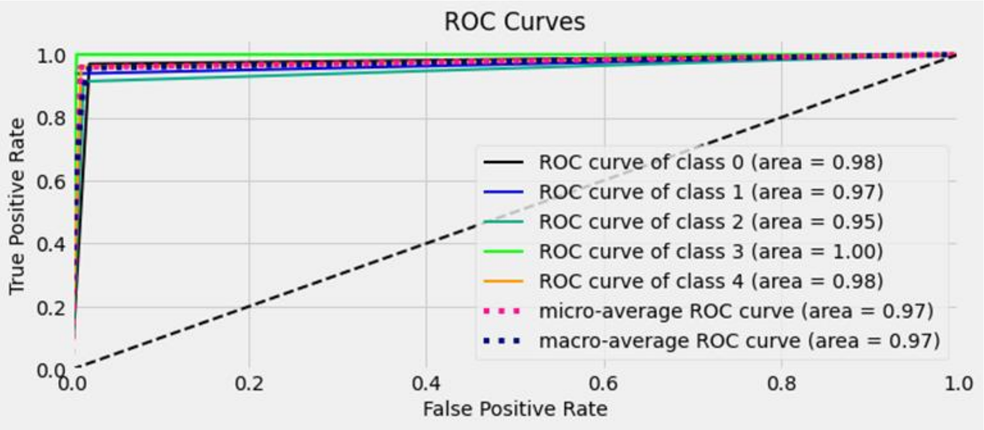


Fig.14. ROC curve for distinct classes with SVM classifier

The confusion matrix for the Random Forest classifier is shown fifteen in the figure with distinct attributes. That depicts the high correlation between class 0. The ROC curve shows better results for class 3 as well which is 0.99. The micro average ROC curve generates the 0.97 accuracies shown in the figure 16.

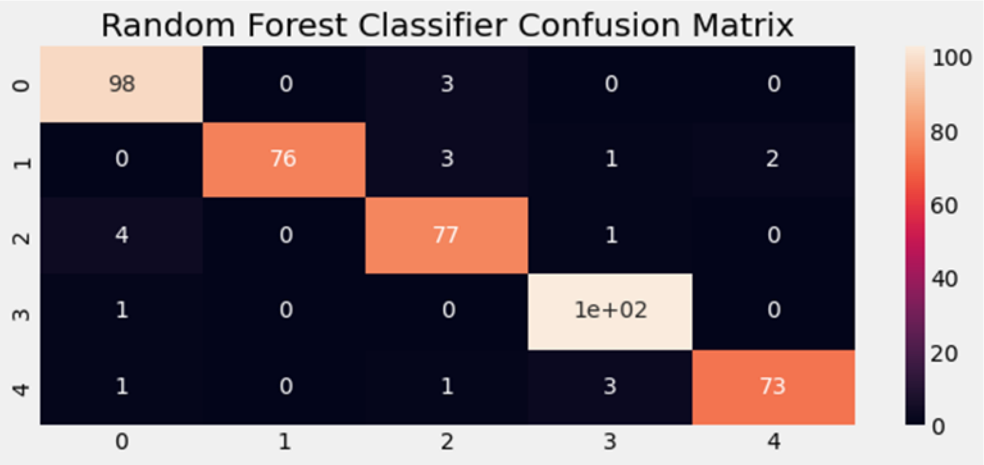


Fig.15. Confusion matrix for Random Forest classifier

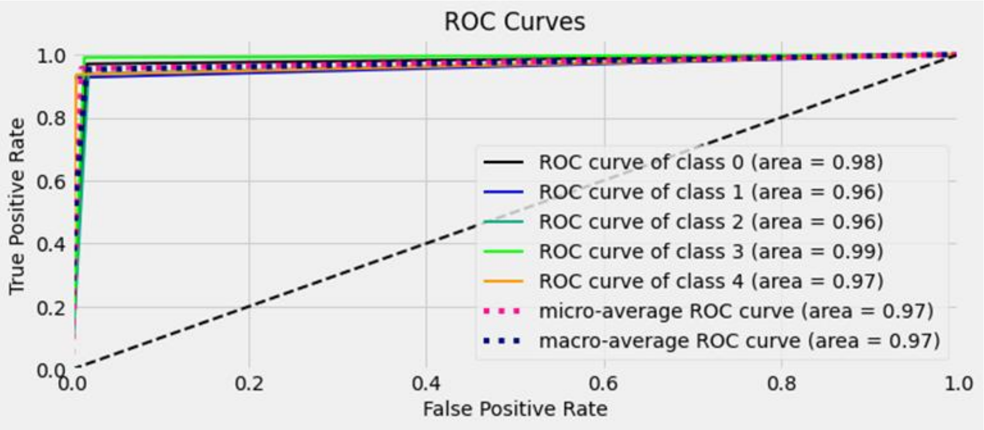


Fig.16. ROC curve for distinct classes with Random Forest classifier

The table I shows the distinct parameter values obtained with different models. It shows the best accuracy for logistic regression 96.1% and minimum time delay with the decision tree classifier. KFold accuracy is achieved by using the SVM model with an MAE factor of 0.082.

Table I: Model Evaluation with distinct parameters

SN	Model	Accuracy	F1-Score	R2-Score	MAE	Recall	Precision	Time Delay	KFold Accuracy
0	LSTM	0.927	0.927	0.844	0.128	0.927	0.928	504.168	NaN
1	BiLSTM	0.906	0.906	0.869	0.128	0.906	0.910	1107.58	NaN
2	Decision Tree	0.791	0.794	0.409	0.451	0.791	0.805	1.15	0.899
3	SVM	0.959	0.959	0.907	0.082	0.959	0.961	4.428	0.983
4	Random Forest	0.955	0.955	0.904	0.082	0.959	0.960	3.68	0.975
5	Logistic Regression	0.961	0.962	0.919	0.073	0.961	0.962	1.51	0.982

As the figure 17 shows the pictorial representation of accuracy, precision, and recall for different models and quantifiers such as BiLSTM, Logistic regression, SVM, Random-forest, LSTM, and decision tree classifier. The proposed model obtained the best accuracy for logistic regression with good precision and recall.

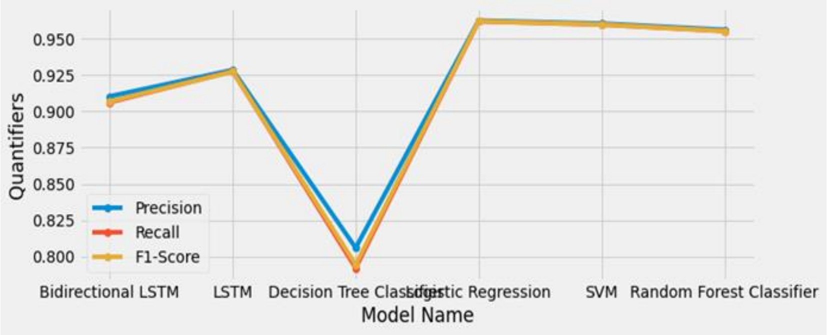


Fig.17. Comparison among distinct model's parameters

The figure 18 shows the comparison between the time delay and document size. As the diagram shows, with the increased size of documents time will also be increased. Although, the decision tree classifier algorithm is taken the minimum time delay while SVM takes the wore time for evaluation.

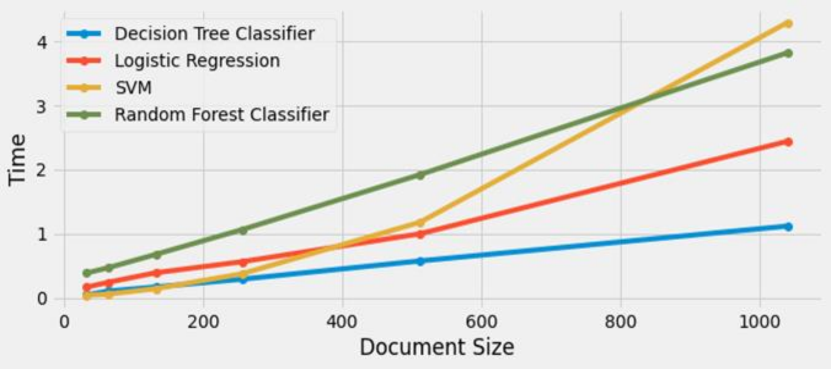


Fig.18. ROC curve for distinct classes with Decision tree classifier

The comparative analysis of the proposed work is done with the existing research conducted by several researchers with distinguished machine learning algorithms.

VI. Conclusion

Retrieving the information from the queries is a great challenge for online users. It becomes a more challenging and time-consuming process. In this article, a model is built for the extraction of information by using query expansion. To achieve these results several machine or deep learning approaches are integrated with the proposed algorithm. The proposed approach is also utilized based on the similarity index, semantic similarity, and historical data. This work's major objective is to develop an ontology-based information retrieval model that makes use of comprehensive domain ontologies and knowledge bases to support semantic retrieval capabilities while maintaining the idea of approximate search in document repositories. The improvement of personalization performance, which is an update based on historical knowledge, is one of the main goals of this effort. The Impact Score Estimation technique is used to improve data extraction using semantic keyword extraction and indexing. To cluster documents based on computed scores, the algorithm is to evaluate similarity estimation which can improve searching by speeding up information retrieval and processing.

VII. Future Scope

There is a huge future scope of research in the realm of information retrieval, query expansion methods, and semantic analysis, as information retrieval research will transform the world. As technology advances, search results can be improved, and information retrieval can be revolutionized. These findings point the way to a future where information retrieval is fast, accurate, and relevant. Using advanced algorithms and semantic comprehension to understand user intent will help search engines return more relevant and meaningful results. These methods may simplify searches across every field. Search engines can evolve by using advanced machine learning and natural language processing algorithms. In conclusion, information retrieval will improve in efficiency, precision, and usability. A new era of abundant and easily accessible information will enable advancements in various fields and industries.

References

[1] R. Kumar and S. C. Sharma, “Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval,” J. Supercomput. Available at: springer.com, vol. 79, no. 2, 2251-2280, 2023. doi:10.1007/s11227-022-04708-9 Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval.

[2] Machine Learning and Ontology-Based Novel Semantic Document Indexing for Information Retrieval – ScienceDirect Machine Learning and Ontology-Based Novel Semantic Document Indexing for Information Retrieval.

[3] B. Dobrucalı Yelkenci et al., “Online complaint handling: A text analytics-based classification framework,” Mark. Intell. Plan. Emerald, vol. 41, no. 5, 557-573, Apr. 28, 2023. doi:10.1108/MIP-05-2022-0188.

[4] Intelligent Ontology Based Semantic Information Retrieval Using Feature Selection and Classification | Cluster Computing. Available at: springer.com.

[5] Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6d82dcfa1b92b225acb106fdc2ddfa63fd94eebe, “Role of ontology in information retrieval.”

[6] D. J. Juliet Thessalonica et al., “Intelligent mining of association rules based on nanopatterns for code smells detection,” Sci. Program. Hindawi Limited, vol. 2023, pp. 1-18, Apr. 13, 2023. doi:10.1155/2023/2973250.

[7] H. K. Azad et al., “Improving query expansion using pseudo-relevant web knowledge for information retrieval,” Pattern Recognit. Lett. Elsevier BV, vol. 158, pp. 148-156, Jun. 2022. doi:10.1016/j.patrec.2022.04.013.

[8] “A-Practical-Framework-for-the-Semantic-Web-Ontology-Learning.pdf,” A Practical Framework for the Semantic Web Ontology Learning. Available at: researchgate.net.

[9] R. Kumar et al., “Optimal query expansion based on hybrid group mean enhanced chimp optimization using iterative deep learning,” Electronics. MDPI, vol. 11, no. 10, p. 1556, May 12, 2022. doi:10.3390/electronics11101556.

[10] Framework for Context-Based Intelligent Search Engine for Structured and Unstructured Data | SpringerLink Framework for Context-Based Intelligent Search Engine for Structured and Unstructured Data.

[11] [11] Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6d82dcfa1b92b225acb106fdc2ddfa63fd94eebe, “Role of ontology in information retrieval.”

[12] O. Bakhteev et al., ‘Cross-Language Plagiarism Detection: A Case Study of European Languages Academic Works,’ Academic Integrity: Broadening Practices, Technologies, and the Role of Students. Springer International Publishing, 2022, pp. 143-161. doi:10.1007/978-3-031-16976-2_9.

[13] Available at: https://arxiv.org/pdf/2305.05754.pdf, “When and what to ask through world states and text instructions.”

[14] Improving Query Expansion Using Pseudo-Relevant Web Knowledge for Information Retrieval – ScienceDirect Improving Query Expansion Using Pseudo-Relevant Web Knowledge for Information Retrieval.

[15] “Electronics | Free full-text | Optimal query expansion based on hybrid group mean enhanced chimp optimization using iterative deep learning,” Optimal Query Expansion Based on Hybrid Group Mean Enhanced Chimp Optimization Using Iterative Deep Learning. Available at: mdpi.com.

[16] “Enhancing e-commerce product Search through reinforcement learning-powered query reformulation,” Enhancing E-commerce Product Search through Reinforcement Learning-Powered Query Reformulation. Available at: acm.org.

[17] “Electronics | Free full-text | Optimal query expansion based on hybrid group mean enhanced chimp optimization using iterative deep learning,” Optimal Query Expansion Based on Hybrid Group Mean Enhanced Chimp Optimization Using Iterative Deep Learning. Available at: mdpi.com.

[18] M. Jain et al., “An evolutionary game theory-based approach for query expansion,” Multimedia Tool. Appl. Springer Science+Business Media LLC, vol. 81, no. 2, pp. 1971-1995, 2022. doi:10.1007/s11042-021-11297-x.

[19] D. Kleyko et al., “A Survey on Hyperdimensional Computing aka Vector Symbolic Architectures, Part II: Applications, Cognitive Models, and Challenges,” ACM Comput. Surv. Association for Computing Machinery (ACM), vol. 55, no. 9, pp. 1-52, Jan. 16, 2023. doi:10.1145/3558000.

[20] M. R. Keyvanpour et al., “HQEBSKG: hybrid query expansion based on semantic KnowledgeBase and grouping,” IETE J. Res. Informa Uk Limited, vol. 68, no. 5, pp. 3750-3765, 2022. doi:10.1080/03772063.2020.1779618.

[21] C. Carpineto and G. Romano, “A survey of automatic query expansion in information retrieval,” ACM Comput. Surv. Association for Computing Machinery (ACM), vol. 44, no. 1, pp. 1-50, Jan. 2012. doi:10.1145/2071389.2071390.

[22] J. Singh and A. Sharan, “A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach,” Neural Comput. Appl. Springer Science+Business Media LLC, vol. 28, no. 9, pp. 2557-2580, 2017. doi:10.1007/s00521-016-2207-x.

[23] S. Dahir and A. El Qadi, “Query expansion based on modified Concept2vec model using resource description framework knowledge graphs,” IAES Int. Institute of Advanced Engineering and Science, vol. 12, no. 2, p. 755, Jun. 01, 2023. doi:10.11591/ijai.v12.i2.pp755-764.

[24] R. Kumar and S. C. Sharma, “Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval,” J. Supercomput. Springer Science+Business Media LLC, vol. 79, no. 2, no. 2, pp. 2251-2280, 2023. doi:10.1007/s11227-022-04708-9.

[25] M. Rashad et al., “RbQE: An efficient method for content-based medical image retrieval based on query expansion,” J. Digit. Imaging, vol. 36, no. 3, 1248-1261, Jan. 26, 2023. doi:10.1007/s10278-022-00769-7.

[26] M. Ghurab et al., “A hybrid semantic query expansion approach for Arabic information retrieval,” J. Big Data, vol. 7, no. 1. Springer Science and Business Media LLC, Jun. 29, 2020. doi:10.1186/s40537-020-00310-z.

[27] R. Padaki et al., “Rethinking query expansion for BERT reranking,” Lect. Notes Comput. Sci. Springer International Publishing, pp. 297-304, 2020. doi:10.1007/978-3-030-45442-5_37.

[28] S. Jain et al., “A fuzzy ontology framework in information retrieval using semantic query expansion,” Int. J. Inf. Manag. Data Insights. Elsevier BV, vol. 1, no. 1, p. 100009, Apr. 2021. doi:10.1016/j.jjimei.2021.100009.

[29] M. A. Raza et al., “A taxonomy and survey of semantic approaches for query expansion,” IEEE Access. Institute of Electrical and Electronics Engineers (IEEE), vol. 7, pp. 17823-17833, 2019. doi:10.1109/ACCESS.2019.2894679.

[30] A. Hameed, “Personalized query expansion,” Int. J. Inf. Syst. Comput. Technol., vol. 2, no. 1. Center for Research and Innovative Technologies, Jan. 30, 2023. doi:10.58325/ijisct.002.01.0043.

[31] R. Ojha and G. Deepak, ‘Metadata Driven Semantically Aware Medical Query Expansion,’ Knowledge Graphs and Semantic Web. Springer International Publishing, 2021, pp. 223-233. doi:10.1007/978-3-030-91305-2_17.

[32] X. Meng et al., “Personalized Semantic Retrieval System based on Statistical Language Model,” 20th International Fall Conference on Computer and Information Science (ICIS Fall), Oct. 13, 2021. IEEE/ACIS, 2021. doi:10.1109/ICISFall51598.2021.9627486.

[33] D. Mavaluru et al., ‘Ensemble Approach for Cross Language Information Retrieval,’ Computational Linguistics and Intelligent Text Processing. Berlin Heidelberg: Springer, 2012, pp. 274-285. doi:10.1007/978-3-642-28601-8_23.

[34] S. Haribabu et al., “A Novel Approach for Ontology Focused inter- Domain Personalized Search based on Semantic Set Expansion,” Fifteenth International Conference on Information Processing (ICINPRO). IEEE, Dec. 2019, 2019. doi:10.1109/ICInPro47689.2019.9092155.

[35] T. Xu et al., “Study of data retrieval optimization techniques based on user interest ontology,” Acad. Med., Apr. 14, 2022 3rd Asia and the Pacific Conference on Image Processing, Electronics and Computers, 2022. doi:10.1145/3544109.3544189.

[36] Q. Xu et al., “Clustering-based fusion for medical information retrieval,” J. Biomed. Inform. Elsevier BV, vol. 135, p. 104213, Nov. 2022. doi:10.1016/j.jbi.2022.104213.

[37] A. Abu Salimeh et al., “Natural language processing and parallel computing for information retrieval from electronic health records,” ITM Web Conf. EDP Sciences, vol. 42, p. 01013, 2022. doi:10.1051/itmconf/20224201013.