Cogent Engineering

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/oaen20

Advanced text documents information retrieval system for search services

Chiranjeevi H S & Manjula K. Shenoy |

To cite this article: Chiranjeevi H S & Manjula K. Shenoy | (2020) Advanced text documents information retrieval system for search services, Cogent Engineering, 7:1, 1856467, DOI: 10.1080/23311916.2020.1856467

To link to this article: https://doi.org/10.1080/23311916.2020.1856467

Published online: 28 Jan 2021.

H S & Shenoy, *Cogent Engineering* (2020), 7: 1856467
https://doi.org/10.1080/23311916.2020.1856467

cogent engineering

COMPUTER SCIENCE | RESEARCH ARTICLE

Advanced text documents information retrieval system for search services

Chiranjeevi H S[1] and Manjula K. Shenoy[1]*

Received: 21 July 2020
Accepted: 03 November 2020

*Corresponding author: Manjula K. Shenoy, Information and Communication Technology, Manipal Academy of Higher Education, India
E-mail: manju.shenoy@manipal.edu

Reviewing editor:
Lin Wang, School of Management, Huazhong University of Science and Technology, Wuhan, China

Additional information is available at the end of the article

Abstract: Information technology has explored the growth of text documents data in many organizations and the structural arrangement of voluminous data is a complex task. Handling the text document data is a challenging process involving not only the training of models but also numerous additional procedures, e.g., data pre-processing, transformation, and dimensionality reduction. In this paper, we describe the system's architecture, the technical challenges, and the novel solution we have built. We propose a Recurrent Convolutional Neural network (RCNN), based text information retrieval system which efficiently retrieves the text documents and information for the user query. Pre-processing using tokenization and stemming, retrieval using TF-IDF (Term Frequency-Inverse Document Frequency), and RCNN classifier which captures the contextual information is implemented. A real-time advanced search system is developed on a huge set of MAHE University dataset. The performance of the proposed text document retrieval system is compared with other existing algorithms and the efficacy of the method is discussed. The proposed RCNN-based text document information retrieval model performs better in terms of precision, recall, and F-measure. A high-quality and high-performance text document retrieval search system is presented.

Chiranjeevi H S

ABOUT THE AUTHOR

Chiranjeevi H S is an Experienced Director with a demonstrated history of working in the information technology and services industry for over 10 years. Skilled in Business development, Databases, Information engineering, and Management. Strong business and operational professional, He is a founder director for Mythinit Technology and DataSciii Solutions Private Limited, where he is currently on sabbatical. His industry-oriented work has led onto extended research as a Ph.D. candidate in MIT, Manipal Academy of Higher Education, Manipal.
Manjula Shenoy K is professor in Department of Information and Communication Technology, MIT, MAHE, Manipal. She has 25 years of teaching and industry mentoring experience. She has over 16 (20) publications in international journals, one (1) government funding and guiding four PhD students in the area Semantic Web Technologies, Data Mining, and Big-Data Analytics, Sentiment Analysis, and Cloud Computing.

PUBLIC INTEREST STATEMENT

The growth of text documents data is unstoppable and structural arrangements of this data's is a challenging task in many organizations. Some key process and procedures need to be applied to the text documents data for the better customer service in the organizations. With respect to this scenario, the new techniques, system architecture, and novel solutions are built. We have designed a Recurrent Convolutional Neural network (RCNN) based text information retrieval system which efficiently retrieves the text documents and information for the user query. A real-time advance search system is developed using large dataset from the MAHE university use case. The performance of the proposed text document retrieval system is compared with other existing algorithms and the efficacy of the method is discussed. The proposed RCNN-based text document information retrieval model performs better in terms of precision, recall, and F-measure. A high-quality and high-performance text document retrieval search system is presented.

Subjects: Technology; Computer Science; Databases; Database Management; Computer Engineering; Information & Communication Technology (ICT)

Keywords: information technology; text documents; search engine; information retrieval; tokenization; recurrent convolutional neural network; retrieval efficiency

1. Introduction

The nature and volume of data have hit the technology changes in recent years, which in turn a major problem on data management and retrieval techniques. Information communication has completely changed nearly every aspect of our lives. Once thought as an unrealistic dream, Data has finally come to fruition—enabling computers to understand and interact with us while processing their thinking. More than 70% of organizations are expected to invest or invested in big data and big data analytics. By 2020, digital information in the world is expected to reach 46 trillion gigabytes (Chiranjeevi & Shenoy Manjula, 2019). Today we can say information system is changed with complete digitization where clerical works like a clerk in the front-end office say—sir please give your identification and come after one day I will search your record and keep it ready. Today a large section of people is dependent on these kinds of systems in their daily personal and professional life. So, the retrieving of the information system is becoming or reaching the stage of one of the popular technologies for accessing the information, technology-enhanced towards having built-in search engines of web applications (Chiranjeevi et al., 2016).

Recent developments made in Information Processing Systems (IPS) have enabled the growth of data. The retrieval of required information from a large database is a challenging task due to the data stored in the text document form (Bijalwan et al., 2014). This can be handled by the "information retrieval process" which retrieves text documents based on a parallel search model. Any sort of industry is cautious about the technical documents of their products. Hence, industrial people submit their regulatory forms in pdf forms. There are several reasons for storing text documents in various formats and especially in pdf structure. Simultaneously, customers face some issues in retrieving the relevant text documents like variants format constraints. Text classification is widely examined by the machine learning community (Gonzalez et al., 2015) where most of the classification techniques predicted different levels of accuracy and effectiveness. In today's explosion of data, cognizance, and information in databases are increasing where the users also encounter issues for better search and retrieval rate.

The conventional model makes use of the Bag of Words (BOW) process which preserves each word by their frequency (Kumar et al., 2014). The documents are related to each other by their Boolean or vector retrieval models. Some documents are decontextualized which depends on stop words removal, stemming, etc. These models offered richness and complexity of the bases of the information retrieval process. Document Image Retrieval (DIR) facilitates the significance of retrieving, index, and annotation of visual information. It works on two aspects: (a) Image and (b) Text using Optical Character Recognition (OCR) (Chao & Fan, 2004). Deploying OCT techniques on an unformatted structure is a complicated and error-prone process. These objects often denote information to degrade the complexity and vagueness of the retrieval rate. Complex problems exist in classifying the data with better accuracy of the representation.

Searching space grows exponentially in which optimization is much required. Regularization of the term ensures that the feature representation models give diffuse vector on given underlying models. Classification-based dictionary refinement process enhances the class corpus which in some cases, increased computational overhead in real-world applications (Jun et al., 2014). Coming to, relevancy model which states positions of query terms in feedback documents. Weighting, Ranking, and Association models are related to the estimation of correlation measure which degraded multiple query systems. Indexing and text tags are not the best practice for a retrieval system, an automatic classification of text documents and indexing the text corpus in the

documents is the new technology in the trend. The requirements of the organizations and the need for the text documents retrieval system is growing in the market due to large set of documentation. The most prominent applications of text classification include subject categorization of organizations department documents, education sector, and health care, etc. these requirements led us to survey many organizations and we found motivated to proceed with the collection of requirements, one of the motivation is described in the next section which tells the urge of developing an advanced text documents search system.

1.1. Motivation

A study is carried out in the Manipal Academy of Higher Education (MAHE), which is handling a large volume of text documents data. The case study involved with understanding the usage of the papers and the text documents generated in the 12 institutions under the Manipal Academy of Higher Education. Table 1 describes the number of institutions and the text documents generated in the last three years.

The Manipal University Administration Departments are associated with search and retrieval of text documents every day. The departments are listed (HR, Finance, Admission, Legal, Quality, Purchase, Alumni Centre, Warden Office, Student welfare, Director of Research, Registrar office, PRO, Statistics), the documents distribution across the departments is shown in Table 2. The described usage of text documents in the organization led to the proposed research and to develop an efficient search system.

In recent times, due to the widespread use of machine learning and deep neural networks, led to a massive research scope for various Natural Language Processing (NLP) research. Recurrent Neural Network (RNN) was utilized in several types of research dealing with text retrieval but it lacked effectiveness when considering the semantics from the entire documents. Convolutional Neural network (CNN) took over RNN which can deal with a greater number of documents as well as to retrieve exact relevant documents. Even though enhanced results are obtained using CNN, it

Table 1. Paper consumption from 01/04/2015 to 31/03/2018 at MAHE, Manipal

Sl. No	Institutions under Manipal Academy of Higher Education	Papers consumed
1	MAHE	44,02,000.00
2	MIT	32,30,000.00
3	KMC	7,24,000.00
4	MCOPS	4,25,000.00
5	SOAHS	3,50,000.00
6	MMME	2,71,500.00
7	SOCMA	2,80,000.00
8	SOLMA	2,35,000.00
9	SOMMA & MCONS	2,00,000.00 and 1,80,000.00
10	WGSMA	2,55,000.00
11	CODMA	4,50,000.00
TOTAL		1,10,02,500.00

Table 2. Document distribution (the type of documents)

10 million text documents as described in Table 1		
35% of students document 40% of Academic documents	80% of Administration documents	100% English

still had the drawback of considering the semantics of text more precisely (Siwei et al., 2015). The focus of the work is to propose a novel text document information search architecture and develop a performance-enhanced retrieval system using Recurrent Convolutional Neural network (RCNN) and retrieval techniques.

1.2. Literature review

This section presents the prior techniques involved in the information retrieval process. In (Mubashir et al., 2018), the authors presented a pattern-based comprehensive stemmer and short text classification for the Urdu language. A rule-based stemmer is suggested to categorize the text mentioned in the Urdu language. The condition-based text classification limited the suitability of stripping approaches. The author in (Sezer, Theo Gevers et al., 2017) detected the text for the fine-grained object using text recognition and encoding process. The system is degraded by limiting pixel-level annotations. The same author extended the study (Sezer, Van Gemert et al., 2017) using text retrieval from natural scenery images. Word recognition accuracy consumes higher time for data dictionaries. In (Florian et al., 2012), the researchers presented a visual classifier training for text document retrieval with accurate filters. The active learning model reduces the effort of labeling but the support of multi-labeling is not focused.

The author (Yousif et al., 2019) reviewed stemming techniques in Arabic text classification which reduced the dimensionality curse. At k-fold cross-validation, the feature extraction process reduced classification accuracy. In (Xiao et al., 2019), they presented a multi-domain model for neural networks using orthogonality constraints for private and shared features. The information retrieval rate is less due to inefficient feature representation. Detected text and caption in videos using language independency were suggested by (Xuzhao et al., 2011). Multimedia documents are collected and processed for detecting text and caption in videos. An inefficiently structured tree has increased computational complexity. The author in (Said et al., 2013) analyzed biomedical datasets using graph kernels and controlled vocabulary. A graph structure is developed from semantic information using kernel classifiers. Weight-based feature modeling on connected nodes demolishes the efficiency of classification predictions.

The author (Harald et al., 2013) studied user-guided filtering for real-time analysis of microblog messages. Supervised classification is employed for twitter sentiment analysis that reduced the complexity of the data warehouse which enhanced the management overhead. In (Li et al., 2014), studied the bag of frames for music information retrieval. Each audio is represented in code words and then formulated into a term-document structure that reduced the data dimensionality. Though the system achieved better accuracy, it does not ensure the generalizability of the data. In (Wei et al., 2018) they presented a transferred neural network for detecting text in videos. It aimed to eliminate the false-positive rate but the usage of c-mean clustering throws a higher collision rate. The author in Peter Whitehead et al. (2017) discussed the evidence of system thinking by analyzing the linguistic elements of the corpus of documents. Semantical information analysis degraded the long-term goal assessment and centroid placements of the reference data vector.

The study in (Bo et al., 2019) presented an extraction model for emotions prediction using ranking methods. Initially, the emotion features are categorized into emotion dependent and independent features which help to find the relevancy for each emotion. The relevancy measure determines the sorts of emotions. Feature normalization affects the performance of extraction models. In (Francisco et al., 2016), the authors presented a generic summarization for music information retrieval. Set construction using maximal marginal relevance which throws higher redundancy rate in multi-class rate. Text categorization using a genetic algorithm was studied by (Alian Diaz et al., 2018). It resolved classification problems via an optimized solution and examined it in terms of accuracy, precision, and recall. The index maintenance of classification is not properly assigned for testing documents.

The author in Yong et al. (2019) analyzed underground forums for data breaches. The topic of forums was modeled using Latent Dirichlet Allocation (LDA) which takes a higher data tree structure. A similar study did by (A. Aljamel et al., 2019) presented smart information retrieval using a centric optimization approach. The feature selection model is optimized using genetic algorithms compared to basic classifiers. Though the results have proven better accuracy, the settings of the threshold level applicable to the small-scale application. The author (Junjie et al., 2015) presented an attribute-based re-ranking model for web image search. Initially, the images are represented in a graphical model based on pre-defined attributes. Then, hypergraph ranking is then used for ordering the images. Further, the visual joint process determines the image classification and the Image retrieval rate during hyperedges imposes a higher hierarchical tree structure. In (Eugene et al., 2017), they developed an e-discovery algorithm for analyzing the effectiveness of the text classification. The analysis of small documents significantly varies the performance of the system.

In Deepak et al. (2017), the authors have discussed mutual information for text feature selection. The authors dictated the significance of mutual information using all classifiers on four standard datasets. The unlabeled documents degraded the classifier's performance with a higher false-positive rate. In (Jonathan et al., 2017), the authors presented an emotion detection model for ensemble classification using word embedding concepts. Each data is tagged with pre-trained word vectors degraded sentiment analysis classification. The authors in (Swapnil et al., 2013) developed document classification using topic labeling. It worked based on the closeness value of aggregated topics. But the reassigning of class labels incurs a higher number of features. In (Frinken et al., 2012; Rajendra et al., 2017), they discussed the feature selection model using commonality rarity measure computation. The features are mapped and aligned for document classification. The dependency rate of document classification extremely avoids local minimization search.

1.3. Research gaps

Text classification is to classify the documents concerning their predefined labels. Several applications like document filtering, meta-data generation, word sense disambiguation, and text document organization were related to this domain. Depending on constraints, the applications of text classification differ. Here, multi-label text classification is employed where each document is related to different labels with a common cause. Image-based text classification is a subcase of the multi-label domain which explores the issues like variant user's opinion, word ambiguity, and context-sensitivity. Since text mining engages significant textual components, it is essential to enhance text refining algorithms, in specific to text document formats. The above-mentioned issues are resolved regarding the university use case described in the motivation section which uses a huge set of text documents.

The structure of the paper is organized as follows: Section 2 presents the overall architecture of the proposed search system and its various components of the research methodology; Section 3 describes how the system gathers and processes a large set of text documents, development results and analysis and Section 4 presents our conclusions and future work efforts.

2. Methodology: System model and architecture

This section presents the research problem, objectives, and working process of the research study. The proposed text document search system is designed with entities of front end and back end feature. Figure 1 presents the workflow of the research model.

Figure 1. Workflow of the proposed model.

The architecture describes each process of a text document information retrieval [IR] system as shown in Figure 2.

The information retrieval major components are the indexing process and the querying process. The indexing process involves some of the pre-processing techniques which explained. The RCNN technique is implemented in the proposed architecture, the neural network approach is very effective to compose the semantic representation of the texts, and also it can capture more contextual text information of features compared to other traditional methods.

a) Data collection: It is the foremost step that determines the effectiveness of the research study. Several sets of text documents are collected from the MAHE university repository. The collected text documents data are then converted into corpus text and also for scanned text documents data is converted to corpus text using Optical Character Recognition (OCR).

b) Data pre-processing: It is the second step that assists to eliminate irrelevant data such as, noise, incomplete, and sensitivity. The pre-processing techniques involved here are described in Figure 3.

c) Document data store: A relational database system is used to store text documents and metadata.

2.1. Tokenization

A contextual detail is obtained using text. It is represented as tokens such as words, phrases, and symbols. The possible tokens are further used as input for database processing. Most of the languages do not have clear boundaries between words. It uses whitespaces between the words.

2.2. Stop-word filtering

Common words like “and, are and this” are frequently used which is not used for knowledge source. It is vital to eliminate from textual data.

2.3. Parts of speech tagging/Information extraction

It improves the word and its context with detailed information about itself and its neighbours. The input to a tagging algorithm is a sequence of words and a tag set, and the output is a sequence of tags and a single best tag for each word.

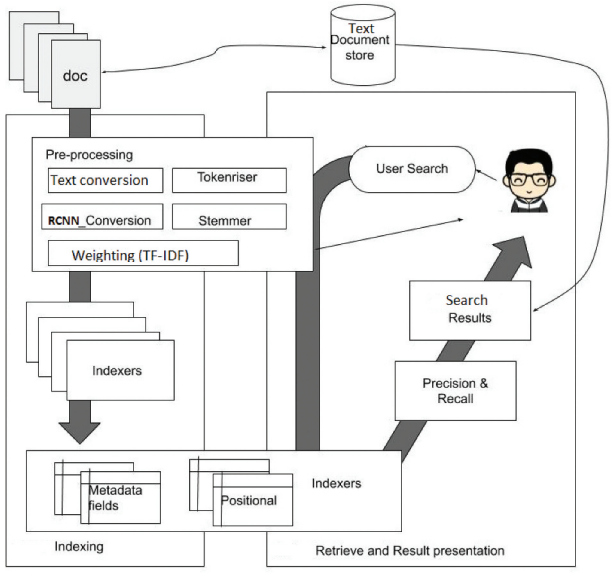
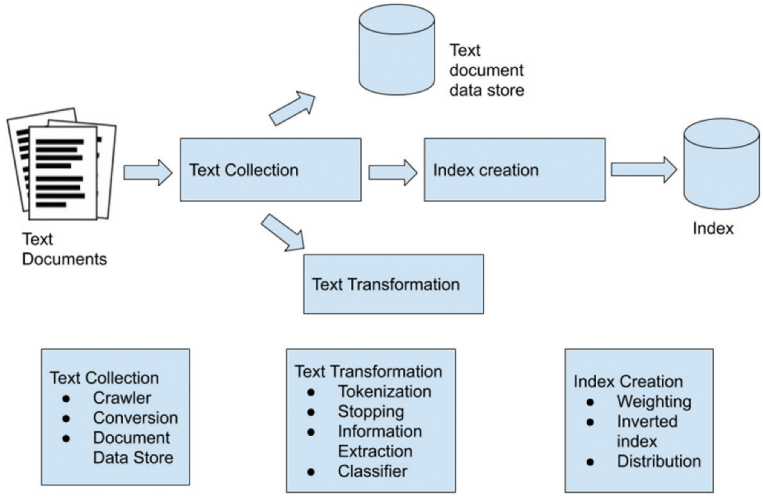


Figure 2. Text document information retrieval system architecture.

Figure 3. Indexing process.



2.4. Classifier

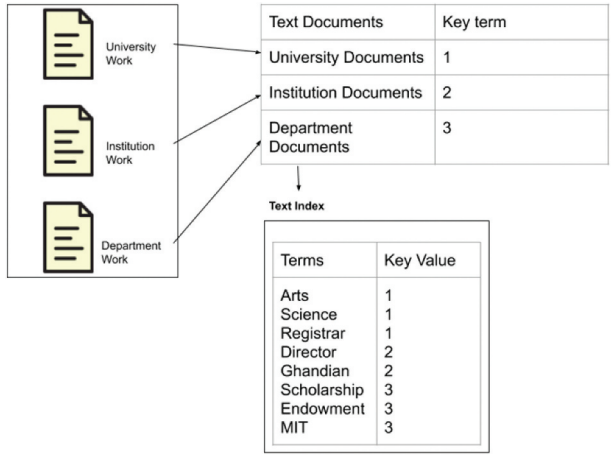
To identify the class-related metadata information in the text documents. The classification process assigns predefined key labels to the text documents.

d) Feature selection/Weighting: It is a vital step for achieving better-classified results. Here, Term Frequency-Inverse Document Frequency (TF-IDF) is employed for selecting the required features. The term frequency of each word is counted and then normalized using inverse document frequency. The frequently occurred words are normalized and mapped by its weight. By doing so, the common words are matched. Then, the weight of each word is estimated for deriving a pre-specified threshold. After these weights of each term in each document have been calculated, those which have weights all higher than pre-specified thresholds are retained. Subsequently, these retained terms form a set of key terms for the document set D.

e) Inverted index: Inverted index is the core task in the indexing process which converts the document term information into term document corpus and creates the inverted index as shown in Figure 2.1.

Text documents of different formats (PDF, Word, image file, scanned, and image formats) can be indexed. Initially, the text in the documents is indexed and stored in a repository which is retrievable through the search system. The process of indexing is preparing a second, separate

Figure 2.1 Text documents indexing based on Terms and Key value.



representation of the text documents that are optimized for retrieval in the text document server. The user search with the terms that are in the text documents, all the texts are indexed based on the tokenization after removing the stop words. This list of terms and the key value is called an inverted index, a bidirectional mapping between terms and documents is formed to get a term-document matrix. For metadata result lists, the information that is needed for the weighting of terms and text documents is also included in the index.

2.5. Learning classifiers

The pre-processing and the index construction tasks showed in Figure 3 is explained with the RCNN algorithm. The Pseudocode describes the explanation of the text documents index building process. Algorithm 1 describes the working of TF-IDF and RCNN for the text documents which are indexed.

2.6. Algorithm 1

Input: Set of documents $D = \{d_1, d_2 \ldots .d_n\}$; Minimum threshold values β, γ, and θ.

Output: Key terms D.

Steps: (a) Extracting the term $T = (t_1, t_2 \ldots t_3)$ from documents.

(b) Eliminating the stop words.

(c) Porter stemmer algorithm is used for deriving the words.

(d) Derived words connected with wordnet senses disambiguation for developing the database.

(e) Global words are obtained and generate the keywords.

(f) Validate all its weights.

(g) A set of key terms D is achieved.

(f) Classification.

It is the final step that operates based on key terms D. Recurrent Convolutional Neural Networks (RCNN) is employed as classifiers in text analysis. The whole text is considered a region to build convolutional neural networks. The semantics of the text is given as input to RCNN. It composes of three layers, namely, the convolutional layer, the recurrent layer, and the transcription layer. Figure 4 represents the architecture of the RCNN.

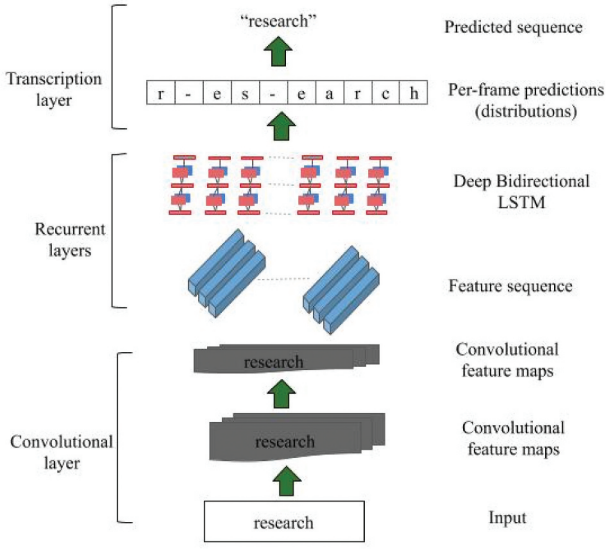
i) Transcription layers:

It denotes how the keywords are taken for building the layers. The input text is converted into frames. Here, the irrelevant spaces are eliminated.

ii) Recurrent layers:

The relevant features are extracted from patches of data using non-linearity of the affine function. Initially, the data patch is fed into recurrent layers and then features are obtained. For each frame, bidirectional LSTM is created in capturing the temporal dependencies using their historical information. Input document D consists of a set of words w_1, w_2 ....w_nWith their required connection parameters θ. Context helps to define the accurate meaning of the word. Here, bidirectional RNN is suggested for capturing the contents.

Figure 4. Architecture of the RCNN.



Let $C_1(W_i)$ be the context of the left words and $C_r(W_i)$ be the context of the right words. The below equation.1 represents the derivation of the context.

$$c_1(w_i) = f\left(W^{(1)}c_1(w_{i-1}) + W^{(sl)}e(w_{i-1})\right) \tag{1}$$

The context (Hidden) layer is denoted by$W^{(i)}$which derives right and left word of context. Similarly, the right context of the word is computed by the equation. 2.

$$c_r(w_i) = f\left(W^{(r)}c_r(w_{i+1}) + W^{(sr)}e(w_{i+1})\right) \tag{2}$$

Then, the word is represented in vectors of left, right and word embedding $e(wi)$ is given in the equation. 3:

$$x_i = [c_1(w); e(w_i); c_r(w_i)] \tag{3}$$

Hence, the significance of disambiguating words was applied with CNN in the fixed window. Then, the linear transformations with activation of *tanh* to x_i, is computed for giving input to its successive layer, given in the equation. 4.

$$y_i^{(2)} = \tanh(W^{(2)}x_i + b^{(2)}) \tag{4}$$

iii) Convolutional layers

The representations of words to its convolutional layer are computed as given in the equation. 5:

$$y^{(3)} = \max_{i=1}^{n} y_i^{(2)} \tag{5}$$

The text of variant lengths is converted into a vector in convolutional layers. The time complexity is of $O(n)$. then, the convolutional neural networks are given as equation.6:

$$y^{(4)} = W^{(4)}y^{(3)} + b^{(4)} \tag{6}$$

Finally, the output layers are then transformed into probabilities and given in the equation. 7.

$$p_i = \frac{\exp\left(y_i^{(4)}\right)}{\sum_{k=1}^{n} \exp\left(y_k^{(4)}\right)} \tag{7}$$

3. Implementation results and discussion

This section presents the implementation of the text document information retrieval system for the set of text documents database with different formats of documents and the analysis of the proposed study. Table 1 provides detailed information about each dataset used. The proposed model is implemented using Django and Python 3.6. Using HTML5, the text search engine front-end design is developed and the implemented code is hosted on the server www.textdocuments.in. In the UI the text documents of any format can be added on click of the Add-Document button, the preprocessing takes place concerning our proposed technique in the backend. The keywords are used for text documents searching as shown in Figure 5.

Initially, texts are extracted from the text documents and text corpus is built. Using the python tesseract function the scanned image text documents are converted. SQLite is used for storing the text corpus from text documents and images and an inverted indexing table is built. For different user keywords, the performance of the proposed IR system is evaluated based on performance metrics like Precision, Recall, and F-measure. After evaluating the proposed RCNN-based information retrieval, we employ K-Nearest neighbor Classifier (KNN) for comparing the performance. The metrics are defined as shown below.

3.1. Precision

The precision is used to measure the retrieved text documents know to be relevant to the given query. Precision metrics of a classifier is generally defined as the ability of the classifier not to label a sample as true (T) that is False (F). It is expressed as below,

$$P = \frac{T_p}{T_p + F_p} \tag{8}$$

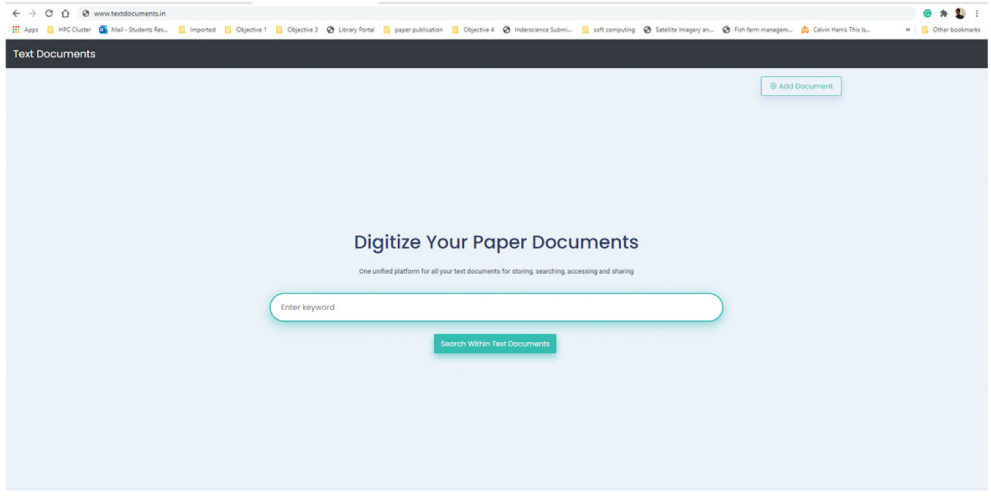


Figure 5. Text documents information retrieval system hosted on www.textdocuments.in.

3.2. Recall

The recall is used to measure the relevant text documents that are effectively retrieved for the given query. The recall is defined as the ability of the classifier to find all the False (F) samples. The recall is expressed as given below,

$$R = \frac{T_p}{T_p + F_n} \tag{9}$$

3.3. F-measure

The ***F-measure*** can be interpreted as a weighted harmonic mean of precision and recall. F-measure is expressed as given below,

$$F_m = 2\left(\frac{(P * R)}{(P + R)}\right) \tag{10}$$

Based on the above metrics the performance of the proposed RCNN-based text information retrieval system is evaluated. The inferred values are explained in section 4.1

3.4. Performance evaluation and discussion

Table 3 describes the performance metrics evaluated using RCNN classifier. The values of precision, recall, and F-measure are calculated concerning the query keyword and the final retrieved results from the set of text documents. The time complexity of the developed RCNN model performs $O(n)$ where n is the length of the text.

Figure 6 explains the graphical representation of the respective Table 3 with RCNN performance metrics. The data inferred from the graph, the average precision, recall, and F-measure values of the proposed approach are 87.5%, 68.75%, and 75.4 %, respectively.

To evaluate the effectiveness of our proposed approach, we have compared the performance and implemented the same approach using K-Nearest Neighbor (KNN) classifier. The performance

Table 3. Performance analysis of RCNN classifier

No	Precision (%)	Recall (%)	*F-measure* (%)	*Accuracy* (%)
1	100	75	85.714	98
2	50	50	50	
3	100	100	100	
4	100	50	66	

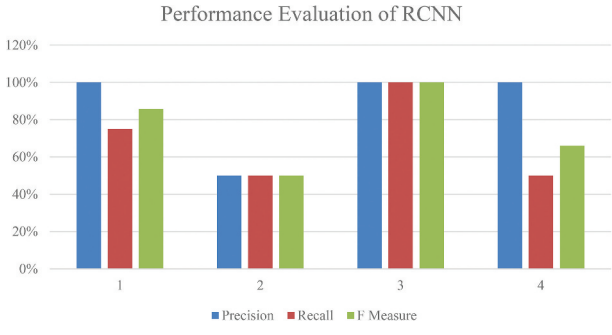


Figure 6. Performance Evaluation of RCNN.

Table 4. Performance analysis of KNN classifier			
No	Precision (%)	Recall (%)	F-measure (%)
1	97.65	73.27	84.184
2	47.657	48.27	48.47
3	47.657	98.27	65.14
4	97.657	98.27	98.47

of the same is evaluated using the metrics; precision, recall, and F-measure and the recorded values as shown in below Table 4.

Figure 7 shows the graphical representation of the performance metrics evaluated using KNN. As inferred from the below graph, the average precision, recall, and F-measure values of KNN classifier are 72.66%, 79.52%, and 74.07 %, respectively.

To understand the classifiers performance, the comparison of proposed RCNN and existing KNN is performed in addition to other existing algorithms from the literature. Initially, the Average precision value of each classifier is evaluated and is tabulated as described in Table 5.

Figure 8 shows the comparative graph for proposed and existing technique. Here RCNN is compared with developed KNN technique and other algorithms like LSTM and CNN (Semberecki & Maciejewski, 2017) where a similar approach for text information retrieval is carried out. As inferred from the below graph, it is evident that our proposed RCNN technique has performed better as the average precision value is 87.5% when compared with the next least value of 86.21% using LSTM. Here KNN registered the lowest precision value with 72.66% which is less than 82.07% of the CNN algorithm.

We have also compared our approach with two other techniques from relevant literature like Text-Block FCN (Zheng et al., 2016) and Logistic regression (Duy Duc et al., 2016) to make our statement of better performance of the proposed approach efficient. Table 6 shows the average performance metrics obtained for each technique concerning precision, recall, and F-measure.

Figure 9 shows the comparative graph of each technique following precision, recall, and F-measure. From the graph, it is inferred that the F-measure of RCNN (75.4%) has outperformed all other existing classifiers with an enhanced precision value of (87.5%) with the next best being

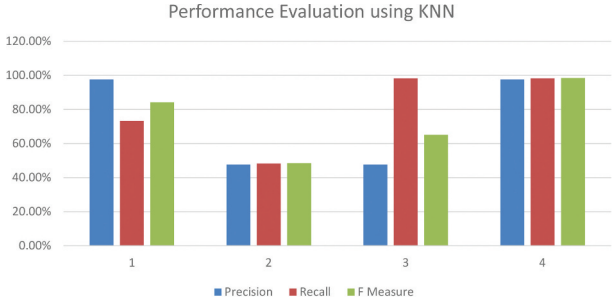


Figure 7. Performance Evaluation using KNN.

Table 5. Comparison of the precision value of proposed and existing classifiers				
Metric	**RCNN (%)**	**KNN (%)**	**LSTM (%) [31]**	**CNN (%) [31]**
Average Precision	87.5	72.66	86.21	82.07

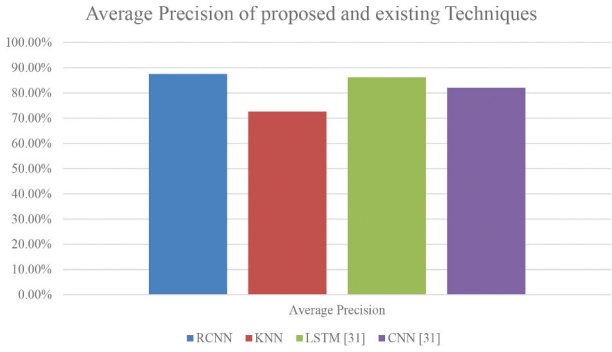


Figure 8. Average Precision of proposed and existing Techniques.

Table 6. Comparison of performance of proposed RCNN and KNN classifier

Methods	Precision (%)	Recall (%)	F-measure (%)
RCNN	87.5	68.75	75.4
KNN	72.66	79.52	74.07
Text-Block FCN [32]	83	67	74
Logistic regression [33]	70.32	66.74	63.18

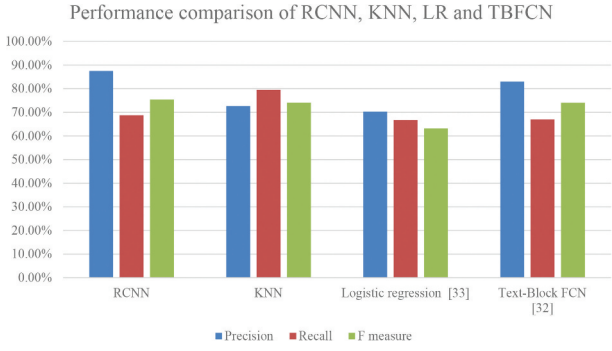


Figure 9. Performance comparison of RCNN, KNN, LR and TBFCN.

KNN in terms of F-measure with (74.07%). In terms of precision metrics, Text-Block FCN has shown the second-best of (83%).

4. Conclusion

In the proposed work, we have developed a text document information retrieval system using Recurrent Convolutional Neural Networks (RCNN) and a search architecture is implemented. Initially, the use case text document database of MAHE University is collected and the classification of text documents is applied. The case study describes the need for the proposed retrieval system and also tells how to minimize the usage of papers. In the pre-processing, the data are filtered using tokenization and steaming process, the white spaces are eliminated and the keywords are predefined. Term Frequency Inverse Document Frequency (TF-IDF) process is applied to compute the frequency of the words. Depends on frequency, the weight of each word was estimated and the weights are taken as input to the classifiers. In RCNN, each layer contributes a framework for the keywords. Based on the trained data, RCNN retrieves the text documents concerning the user query keyword. The developed system is compared with other existing techniques and the performance of each technique was evaluated. On evaluation, it was evident that our proposed RCNN has outperformed other classifiers with average precision, recall, and F-measure value of 87.5% and 75.4%. The accuracy of the designed system architecture is 98% and the developed application is hosted on www.textdocuments.in server.

The collection of text documents is from over 12 different knowledge sources, we have also discussed the techniques used for improving the search accuracy, such as domain-knowledge-based query expansion, and log analytics. The proposed methodology and system architecture are not limited to a domain and text document retrieval system performance results are extremely promising for organizations that want to extend the customer service effectively.

Future scope incorporates the capability to automatically infer structured knowledge from the vast variety of text documents without needing to rely on subject matter experts.

Acknowledgements
This work is supported by the Vision group of science and technology (VGST), Government of Karnataka, India [grant number 629, under RFTT Scheme, 25/08/2017, and submitted on 27/06/2019]. We thank our industry mentors who have supported the research work to carry out; Dr Syam Sundar, IBM India.

Funding
This work was supported by the Vission Group of Science and technology, Government of Karnataka, India [629]

Author details
Chiranjeevi H S[1]
E-mail: chiranjeevi.hs@learner.manipal.edu
Manjula K. Shenoy[1]
E-mail: manju.shenoy@manipal.edu
[1] Manipal Academy of Higher Education, ICT, MIT, Manipal 576104, India.

Citation information
Cite this article as: Advanced text documents information retrieval system for search services, Chiranjeevi H S & Manjula K. Shenoy, *Cogent Engineering* (2020), 7: 1856467.

References

Alian Diaz, M., Bertha Rio- Alvarado, A., Hugo Barron Zambrano, J., & YukaryGuerer, T. (2018). An automatic document classifier system based on genetic algorithm and taxonomy. *IEEE Access*, *6*(1), 21552–21559. https://doi.org/10.1109/ACCESS.2018.2815992

Aljamel, A., Osman, T., Acampora, G., Vitiello, A., & Zhang, Z. (2019). Smart information retrieval: domain knowledge centric optimization approach. *IEEE Access*, *7*(1), 4167–4183. https://doi.org/10.1109/ACCESS.2018.2885640

Bijalwan, V., Kumari, P., Pascual, J., & Semwal, V. B. (2014). Machine learning approach for text and document mining. *arXiv Preprint*, *arXiv*, (1), 1406.1580. https://arxiv.org/abs/1406.1580

Bo, X., Hongfei, L., Lin., Y., Diao, Y., Yan, L., & Xu, K. (2019). Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, *7*(1), 15573–15583. https://doi.org/10.1109/ACCESS.2019.2894701

Chao, H., & Fan, J. (2004, September). Layout and content extraction for pdf documents. In *International Workshop on Document Analysis Systems* (pp. 213–224). Springer, Berlin, Heidelberg.

Chiranjeevi, H. S., ManjulaShenoy, K., Prabhu, S., & Sundhar, S. (2016). DSSM with text hashing technique for text document retrieval in next-generation search engine for big data and data analytics. *2016 IEEE International Conference on Engineering and Technology (ICETECH)* (pp. 395-399). doi: 10.1109/ICETECH.2016.7569283

Chiranjeevi, H. S., & Shenoy Manjula, K. (2019). A Text Document Retrieval System for University Support Service on a High Performance Distributed Information System. *2019 IEEE 4thInternational Conference on Cloud Computing and Big Data Analysis (ICCCBDA)* Chengdu, China (pp. 50-54). doi: 10.1109/ICCCBDA.2019.8725768

Deepak, A., Kesari, V., & Tripathi, P. (2017). Mutual information using sample variance for text feature selection. *ACM Journals on Communication and Information Processing*, (1), 39–44. https://doi.org/10.1145/3162957.3163054

Duy Duc, A. B., Del Fiol, G., & Jonnalagadda, S. (2016). PDF text classification to leverage information extraction from publication reports. *Journal of Biomedical Informatics*, *61*(1), 141–148. https://doi.org/10.1016/j.jbi.2016.03.026

Eugene, Y., Grossman, D., & Frieder, O. (2017). Effectiveness results for popular e-discovery algorithms. *ACM Journals on Artificial Intelligence and Law*, (1), 261–264. https://doi.org/10.1145/3086512.3086540

Florian, H., Koch., S., Bosch., H., & Ertl.,, T. (2012). Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2839–2848. https://doi.org/10.1109/TVCG.2012.277

Francisco, R., Ribeiro., R., & Martins de Matos, D. (2016). Using generic summarization to improve music information retrieval tasks. *IEEE Transactions on Audio, Speech, and Language Processing*, *24*(6), 1119–1128. https://doi.org/10.1109/TASLP.2016.2541299

Frinken, V., Fischer, A., Manmatha, R., & Bunke, H. (2012). A novel word spotting method based on recurrent neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(2), 211–224. https://doi.org/10.1109/TPAMI.2011.113

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., & Greene, C. S. (2015). Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics*, *17*(1), 33–42. https://doi.org/10.1093/bib/bbv087

Harald, B., Thom., D., Edwin Puttmann, F. H., Koch, S., & Kruger, R. (2013). ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2022–2031. https://doi.org/10.1109/TVCG.2013.186

Jonathan, H., shmuelischeuer, M., & Konopnicki, D. (2017). Emotion detection from text via ensemble classification using word embeddings. *ACM Journals of Information Retrieval*, (1), 269–272. https://doi.org/10.1145/3121050.3121093

Jun, S., Park, S. S., & Jang, D. S. (2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. *Expert Systems with Applications*, *41*(7), 3204–3212. https://doi.org/10.1016/j.eswa.2013.11.018

Junjie, C., Junzha, Z., Wang, M., zhang, S., & Tian, Q. (2015). An attribute assisted reranking model for web image search. *IEEE Transactions on Image*

Processing, *24*(1), 261–272. https://doi.org/10.1109/TIP.2014.2372616

Kumar, S. C., Murthy, K. S., & Varaprasad Raju, G. S. (2014). Text Mining For Retrieving the Vital Information. *International Journal of Research in Computer and Communication Technology*, *3*(1), 99–103.

Li, S., Michael Yeh, C. C., Yu Liu, J., wang, J. C., & Hsuan Yang, Y. (2014). A systematic evaluation of the Bag of Frames representation for Music Information Retrieval. *IEEE Transactions on Multimedia*, *16*(5), 1188–1200. https://doi.org/10.1109/TMM.2014.2311016

Mubashir, A., Khalid., S., & Aslam, M. H. (2018). Pattern-based comprehensive Urdu stemmer and short text classification. *IEEE Access*, *6*(1), 7374–7389. https://doi.org/10.1109/ACCESS.2017.2787798

Peter Whitehead., N., Scherer, W. T., & Smith, M. C. (2017). Use of Natural Language Processing to Discover Evidence of systems thinking. *IEEE Systems Journal*, *11*(4), 2140–2149. https://doi.org/10.1109/JSYST.2015.2426651

Rajendra, K. R., Bhalla., A., & Srivastava, A. (2017). Commonality rarity measures computation. *ACM Journals on Information Retrieval Evaluation*, (1), 37–41. https://doi.org/10.1145/3015157.3015165

Said, B., Mishra, M., Huan., J., & Hong, M. (2013). Text categorization of biomedical datasets using graph kernels and controlled vocabulary. *IEEE Transactions on Computational Biology and Bioinformatics*, *10*(5), 1211–1217. https://doi.org/10.1109/TCBB.2013.16

Semberecki, P., & Maciejewski, H. (2017). Deep learning methods for Subject Text Classification of Articles. In *Proc.of the Federated Conference on Computer Science and Information Systems*, Prague (pp. 357–360). doi: 10.15439/2017F414

Sezer, K., Theo Gevers, R. T., & Smeulders, A. W. M. (2017). Words matter: Scene text for image classification and retrieval. *IEEE Transactions on Multimedia*, *19*(5), 1063–1076. https://doi.org/10.1109/TMM.2016.2638622

Sezer, K., Van Gemert, R. T. J. C., & Gevers, T. (2017). Con-Text: Text detection for fine-grained object classification. *IEEE Transactions on Image Processing*, *26*(8), 3965–3980. https://doi.org/10.1109/TIP.2017.2707805

Siwei, L., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural network for text classification. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, (pp. 2267–2273). Austin, Texas, USA.

Swapnil, H., Sandeep, C., Palshikar., G. K., & Chakraborti, S. (2013). Document classification by topic labeling. *ACM SIGIR Journals on Research and Development in Information Retrieval, 1*(1), 877–880. https://doi.org/10.1145/2484028.2484140

Wei, L., Sun, H., Jinghuichu, Huang, X., & Jiexiaoyu. (2018). A novel approach for video text detection and recognition based on a corner response feature map and. *Transferred Deep Convolutional Neural Networks*, *6*(1), 40198–40211. doi: 10.1109/ACCESS.2018.2851942

Xiao, D., Qiankunshi, Cai, B., Liu, T., Zhao, Y., & Ye, Q. (2019). Learning multi-domain adversarial neural networks for text classification. *IEEE Access*, *7*(1), 40323–40332. https://doi.org/10.1109/ACCESS.2019.2904858

Xuzhao, Hsiang Lin, K., Fu, Y., Hu, Y., Liu, Y., & Huang, T. S. (2011). Text from corners: A novel approach to detect text and caption in video. *IEEE Transactions on Image Processing*, *20*(3), 790–799. https://doi.org/10.1109/TIP.2010.2068553

Yong, F., Guo, Y., Huang, C., & Liu, L. (2019). Analyzing and identifying data breaches in underground forums. *IEEE Access*, *7*(1), 48770–48777. https://doi.org/10.1109/ACCESS.2019.2910229

Yousif, A. A., Xiang, J., Zhao, D., & Al. Qanses, M. (2019). A study of the effects of stemming strategies on Arabic document classification. *IEEE Access*, *7*(1), 32664–32671. https://doi.org/10.1109/ACCESS.2019.2903331

Zheng, Z., Zhang, C., Shen, W., Yao, C., Liu, W., & Bai, X. (2016). Multi-oriented text detection with fully convolutional networks. *IEEE conference on computer vision and pattern recognition, 2016*.

Cogent Engineering **(ISSN:) is published by Cogent OA, part of Taylor & Francis Group.**

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com