International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Survival Study on Document Classification Methods for Efficient Semantic Analytics

Sharada C[1], T N Ravi[2], S Panneer Arokiaraj[3]

[1]Research Scholar, Department of Computer Science, Thanthai Periyar Government Arts and Science College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India
[2]Associate Professor, Department of Computer Science, Jamal Mohamed College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India
[3]Associate Professor, Department of Computer Science, Thanthai Periyar Government Arts and Science College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

ABSTRACT
Text mining is procedure of mining the unknown information present in data. Text classification is sub-domain of text mining that acts main part at labeling documents based on their semantic meaning and context. Semantic document classification is a technique that employed the semantic analysis to improve the accuracy level. Document classification is information filtering device employed to enhance retrieval outcomes from query procedure for taking good decisions. Different ML methods are employed to classify available text documents. Semantic analysis is carried out for efficient text classification through semantic keywords using independent features of keywords in documents. The information extraction from the resources and knowledge discovery is a significant area for research. Different document classification techniques are carried out for efficient semantic analysis. In order to address these issues, dissimilar ML and DL methods are discussed for document classification.

Keywords: Text mining, text classification, semantic document, machine learning, semantic keywords, knowledge discovery, deep learning

1. Introduction

Text mining is a developing area in computer science as well as engineering for NLP. Text Classification is branch of NLP. Text Classification is process of tagging raw information text through predefined labels as of websites, online chat messages, and so on. Text Classification comprised SA, CA, TC and QA for reasoning through binary classification, multi-classification as well as multi-label classification. The process of mining the helpful data as of unstructured source like HTML, WWW, news articles, digital libraries, online forums as well as other kinds of documents are hard. With large development on Internet, document classification has received large attention in recent years. The document comprises of texts, and so on. Every document possesses classification issues. Documents are categorized along with their subjects or attributes like document type, etc.

1.1 The main contribution of the article is given as:

- To enhance the accuracy and reduce the time, the keywords were identify for efficient document classification using different datasets namely, CACM data set, Cranfield dataset and (10) Dataset Te xt Document Classification.

- To reduce the error rate, by using different keyword extraction methods for document classification.
- To compare different conventional document classification approach performance such as, LSTM Neural Network, Causal Knowledge Extraction, Integrated Deep Learning Paradigm, Text Mining and Deep Learning Methods, GAE based Document Embedding method, Efficient Data Extraction from Unstructured Documents, Semantic-Based Summarization Approach, and Optimization Driven Cluster based Indexing and Matching. This approach is performed to efficient semantic analytics.

1.2 The contents of this paper are organized as follows:

- Section 1 presents the introduction to the document classification and the challenges faced by during document classification.
- Section 2 presents the review of various machine learning concepts for efficient document classification.
- Section 3 describes the document classification datasets used in this study.
- Section 4 explains the description of different methods for document classification.
- Section 5 explains different performance metrics employed for document classification.
- Section 6 discusses the results attained after experiments.
- Section 7 concludes survey work with future direction.

2. Literature Review

LSTM neural network was introduced in [1] for document classification. It is utilized to categorize files to several types. These are significant factors at allocation network planning. The file will have information about destination, type as well as cost of every utility line. However, classification accuracy was failed to sufficient the improved by LSTM neural network. Two-stage DL pipeline was designed in [2] to extort the cause-as well as-effect triples and built the causal graph dataset. But, extra annotators and inter-annotator agreements parameters failed to integrate the hierarchical depiction of failure modes. B-MLCNN was introduced in [3] with deep learning model. It obtains entire textual review as solitary document as well as categorizes obtainable sentiments. But, it failed to recognize explicit polarity depend on contextual location of text for SA for numerous databases. Latent Dirichlet allocation (LDA) system was introduced in [4] to extract the relevant keywords/key phrase of every chapter as of every mineral exploration reposts. It find to automated text analysis aid to rapid processing of enormous quantities of reports that recognize target mineral schemes as well as their connected geological position and rock mineral composition. But, LDA system was not used for performing topic extraction to extort mineral resource topics

A new document embedding method was introduced in [5] for clustering through graph auto encoder. Undirected and weighted sparse graph was built as of many documents. Graph construction method produces sparse undirected graph as of set of documents efficiently detains elevated similarities among documents through its weighted edges. But, keyword extraction process was not carried out by document embedding method. A semantic similarity measure was introduced in [6] to develop the metadata contained in the scientific articles and important n-grams. Accuracy of our similarity measure is estimate the database of articles is construct and their similarity values computed through human specialist. But, the similarity measure failed to automatically address the research problems.

A new 1D vertical region detection method was introduced in [7] for instance detection for automatic information extraction task from invoices. The designed method decreases this issue to 1D region recognition. Proposed network has less metrics and small inference times. But, the time complexity was not reduced by designed method. Summarization method was introduced in [8] for Knowledge Graphs (KGs) generation of semantic maps. Affinity Propagation and PAM were used for computing semantic distance among nodes at KG and form significant clusters. But, the designed method failed to visualize as well as examine the KGs through enhanced clarity , interpretability at the required level

A document retrieval mechanism was introduced in [9]. Keywords were identified as of pre-processed document. But, the keyword extraction accuracy was not improved by document retrieval mechanism. IAOCOOT method was introduced in [10] for query growth to retrieve semantic feature which match query term. Though error rate was minimized, the time complexity was not reduced by IAOCOOT algorithm.

An effective text categorization method was introduced in [11] with mixture of active and transfer learning. The learning increased the efficiency and capacity with less training data. Length Feature Weight (LFW) was introduced in [12] for vectorized depiction to divider documents. However, complexity study was not carried out by LFW. Probabilistic based Latent Dirichlet Allocation method was designed [13] to enhance accuracy of document clustering. However, it failed to extort important aspects for enhancing clustering result.

Document-relational Graph Convolutional Networks (GCN) was introduced on [14] to increase the accuracy level during text classification through combining cumulative TF-IDF mean for document-document relations. WDAB-LSTM was introduced in [15] to optimize the text features for enhancing the classification performance.

A text classification method was designed in [16]. Topic nodes were employed to construct the triple node sets like word and so on. A novel hybrid approach termed Tasneef was designed in [17] to address the computational challenges with minimum memory usage and runtime overhead for Arabic text classification (ATC).

A new filter feature selection approach termed Multivariate Feature Selector (MFS) was introduced in [18] for text classification. The designed approach computed the score for every feature to expose the hidden information at class, document and document-class levels. A deep learning model depending on BiLSTM, Attention and CNN was introduced in [19] to address the gradient disappearance problem with large number of network layers.

Multi-Label Guided Network (MLGN) was designed in [20] to perform document depiction. The correlation knowledge was used to increase forecast accuracy at down/stream tasks. PLAML was designed in [21] to improve the prompt-based learning for multi-label categorization. Chinese multi-label short text categorization method was designed in [22] depending on Generative Adversarial Network (GAN). BERT was augmented through pinyin embedding for text vector representation to improve text information.

U-PLM was introduced in [23] to join the user individuality as well as PLMs. U-PLM injected the user identity for pre-trained language method to give personalization as of diverse viewpoints. The dual-targeted textual adversarial method was introduced in [24] to create the dual-targeted adversarial examples. The designed method divided into different class through sustaining meaning as well as grammar of inventive sentence.

HGCN was designed in [25] for structured long document categorization. Graph network was employed to extort fine-grained aspects of document. The SSL was introduced in [26] with method learning procedure for label consumption as of one-hot method viewpoint.

SCAM was designed [27] for text categorization from semantic perspective between characters. SCAM increased method learning as well as text categorization tasks through combining the similar-radical characters. Graph Receptive Transformer Encoder (GRTE) was introduced in [28] to join graph neural networks (GNNs) for text classification in inductive fashion. GRTE employed the information from graph-structured relations.

Triple Alliance Prototype Orthotist Network (TAPON) was designed in [29] to construct the generic meta-mapping for tail classifier support. A dynamic chaotic model was introduced in [30] for transforming neurons states at network through neural dynamic features. The designed model increased the text classification accuracy for real-world applications.

3. Dataset description

There are three dataset used for performing document classification, namely CACM data set, Cranfield dataset and (10) Dataset Text Document Classification. The dataset name is CACM is deduces through URL of the dataset from https://ir.dcs.gla.ac.uk/resources/test_collections/cacm/. It consists of 429 words. CACM gathering is group of titles as well as abstracts as of journal CACM. If memory serve, number of journal reference every other. Bits are:

- cacm.tar.gz - Every bits put together
- README – Files short information
- cacm.all - Text of documents
- cite.info - Citation information key
- common_words - Stop words employed
- qrels.text - Relevance judgement lists
- query.text - Original query text

Among all words, document is indicated relevant to the user query.

The second dataset name is Cranfield. The URL of the dataset is given as https://github.com/oussbenk/cranfield-trec-dataset. Cranfield dataset includes two types of conclusion like 1400 collection and 200 collections. This dataset contains the (cran.all) documents, queries (cran.qry) and relevant assessments (i.e. cranqrel). The query relevance for Cranfield is 1400 (cranqrel.txt) with codes for relevancy scale. It comprised three columns. The first is query number. The second is relevant document number. The third one is the relevancy code. From that, the documents are classified based on the keywords.

The third dataset name is (10) Dataset Text Document Classification. The dataset comprised collection of ~1000 newsgroup documents from 10 different newsgroups. The 10 newsgroups gathering has popular database for experiments at ML methods as text classification as well as text clustering. There is a file (class.txt) with reference to document_id number as well as newsgroup. There are 10 files with all documents. In the database, duplicate messages are detached as well as original messages include “From” and “Subject” headers. Each novel message in packed file initiates through four headers. The Newsgroup and Document_id are referenced against class.txt. Every newsgroup file at bundle denotes the solitary newsgroup. Every message in file is text of newsgroup manuscript posted to newsgroup. Database comprised 10 newsgroups namely,

- Business
- Entertainment
- Food
- Graphics
- Historical
- Medical
- Politics
- Space
- Sport
- Technologies

4. Methodology

Text classification expertise denotes routine label generation for text along with exact content of text in assured classification scheme. Main objectives of the NN is to examine text content, to extort semantic data in, to complete text feature representation, to adjust network parameters, and to forecast label for document classification. The data pre-processing, feature extraction and classification are three essential processes for categorizing the documents. The data pre-processing and keyword extraction are carried out for reducing dimensionality of information.Overall process of document classification is given below,

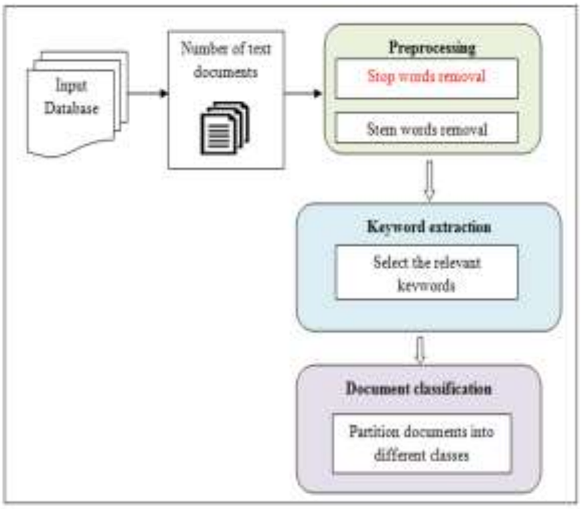


Figure 1 Architecture Diagram of Document Classification

Figure 1 illustrates an architecture diagram of text document classification. Initially, amount of documents is gathered as of input database. After document collection, data preprocessing is carried out. In second step, the keywords or features are extracted from input database. Finally, the document classification is carried out to categorize the documents with the extracted features.

4.1 LSTM Neural Network

Long Short Term Memory (LSTM) is type of RNN with sequence information as input. LSTM is employed at progress direction of series as well as nodes are connected at chain. LSTM is used to address gradient disappearance and gradient explosion through long sequence training. Long text information are preprocessed previous to giving as input to method for training. Words in text are converted to word vectors through embedding. Information carried by path with great correlation based on document category. Look of path and document be in kind are not self-governing. The allocation of path at documents group diverged as of Poisson allocation. Incidence of path is randomly independent event as well as forecasted through Poisson allocation.

4.2 Causal Knowledge Extraction

WRN records are generated through industry with business work flows. The digital records are essential one for asset management. Two-stage DL pipeline is introduced to extort the cause-as well as-effect triples for graph dataset formation. New sentence-level noise removal technique removed information irrelevant to contributory semantics. An annotated dataset was employed to perform efficient noise removal and causality extraction. F1-score are improved for recognition of Cause as well as Effect entities correspondingly. The pipeline is used in the real-world dataset to create the graph database. The designed framework used technical personnel to query causes of tools malfunction for allowing answers to questions.

4.3 Integrated Deep Learning Paradigm

Integrated DL paradigm is introduced for efficient document-basis of sentiments analysis. The sentiment analysis is employed at e-commerce and/or CC oriented businesses. DL paradigms are used for document-basis of sentiment study to classify polarity of contextual sentiments to positive, negative, neutral for making organization decisions. The text as well as disambiguation of natural languages are hard to give accurate recognition as well as extraction subjected to document-basis of information. B-MLCNN is introduced through integrated DL paradigm.It considered textual appraisal as solitary document as well as categorized into obtainable sentiments. BERT pre-trained language method handled feature vector depiction as well as gathered global aspects. MLCNN through kernel dimensions performed the feature extraction. Softmax function is used in B-MLCNN to attain the better classification results.

4.4 Text Mining and DL Techniques

Huge-scale mineral resource reports provide information prosperity for geological knowledge mining as well as knowledge discovery. Geological situations at mineral deposits erudite large deal as of mineral exploration statements. Mineral exploration information is queried as well as aggregated to alleviate prospect examination danger and to minimize the costs. It becomes demanding one to extort valuable geological information with no scanning large number of reports in unstructured textual format. Latent Dirichlet Allocation (LDA) system is introduced to extract the relevant keywords/keyphrases from every mineral exploration reposts. Topic graph identified geological entity for knowledge graph construction. LDA employed the graph visualization to discover contents of report. Text mining and ML method incorporated semantic analysis for information extraction. An automated text analysis performed the large amount of reports to recognize the target mineral schemes as well as their connected geological position. The designed approach converted the mineral exploration data into a structured form in geological knowledge mining.

4.5 Graph Auto Encoder (GAE) based Document Embedding technique

Document embedding technique is used for clustering with DNN. Deep neural network-based document embedding technique depends on number of document clusters. Graph auto encoder document embedding method (GDEM) is used for generating document embeddings with higher similarity characteristics among documents to similar document cluster. GDEM is designed for clustering through graph auto encoder. Undirected as well as weighted sparse graph is built as of collection of documents. Every file is denoted through node. Every weighted edges generated in graph contain elevated cosine similarities among two end nodes. Graph auto encoder to graph determined the node embedding vectors. GDEM method performed outstanding as well as efficient document embeddings through medium- and large-sized files.

4.6 Effective Data Extraction from Unstructured Documents

Key data extraction as of unstructured documents is key issue at different industries. ML methods are used to address the keyword extraction to use textual, visual as well as 2D spatial layout data. Grid based approach is introduced to represent file as 2D grid as well as feed to fully convolutional encoder-decoder network. The designed approach is employed to address the semantic instance segmentation issue. Novel technique is introduced for instance recognition for automatic information extraction task from invoices. Line Item detection is minimized as of 2D bounding box recognition issue to 1D vertical region recognition. Ground truth of line items is symbolized as vertical areas rather than 2D boxes. The ground truth attained through eliminating width dimension of boxes. In network structural design, dimensionality reduction is developed through average-pooling of feature maps pass during decoder.

4.7 Semantic-Based Summarization Approach

KGs is considered as influential device for denoting the semantic structured data as well as for allowing intelligent system development. The semantic maps are used in the summarization method for KGs. The centroid-based clustering algorithm like Affinity Propagation and PAM is introduced to gather semantic distance among nodes at KG and to form the significant clusters. Affinity Propagation and PAM performed well in internal validation parameters. Computed centroids are used to employ informative depiction of knowledge graph (KG). Semantic distance, clustering algorithm and centroids basis of inference is used to perform KG. The semantic map efficiency is constructed in visualizing the KGs.

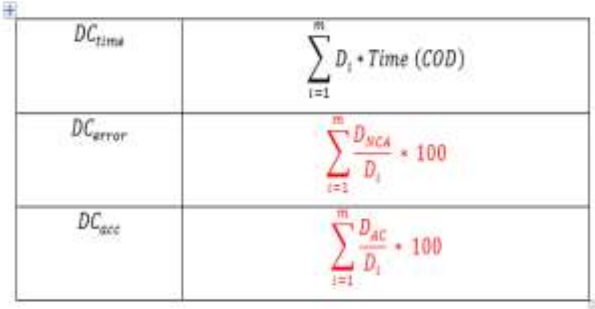
4.8 Optimization Driven Cluster based Indexing and Matching

Document retrieval method is used for reducing time consumed for guide to recall document for analyzing ideas, themes, and file content based on goals. Document retrieval method is introduced with MB-FF. Keywords as of files are recognized from pre-processed file. The data pre-processing is carried out through stemming as well as stop word elimination. TF-IDF is employed at keyword extraction. Holoentropy concept is employed in significant keyword selection process. Selected keywords are employed for relevant document retrieval. MB-FF is used through two-stage mod-Bhattacharya distance match for cluster based indexing. MB-FF method at document retrieval method is used to improve the performance of precision, recall, and f-measure.

5.Experimental setup and performance Metrics

The experimental analysis of existing document classification methods is carried out using three different parameters, namely document classification time, document classification error, document classification accuracy. The document classification time refers to the time consumed for classifying the documents. Document classification time 'DC_{time}' is measured by product of number of documents 'D_i' as well as actual time consumed for classifying single document '$Time\ (COD)$'. It computed in milliseconds (ms). Document classification error is the second metric used for efficient analysis. The document classification error 'DC_{error}' is measured by ratio of the number of documents that were not classified accurately 'D_{NCA}' to the total number of documents 'D_i'. It is measured in percentage (%).Document classification accuracy is measured as ratio of number of documents which accurately classified 'D_{AC}' to total number of documents. It is measured in terms of percentage (%).The formulation of the performance parameters are given in table 1.

Table 1 Performance Metrics

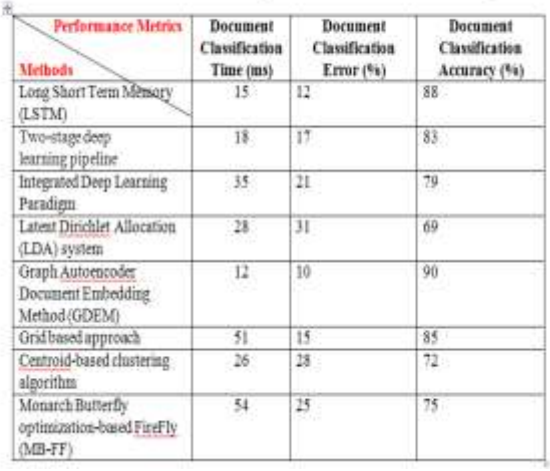


DC_{time}	$\sum_{i=1}^{m} D_i * Time\ (COD)$
DC_{error}	$\sum_{i=1}^{m} \frac{D_{NCA}}{D_i} * 100$
DC_{acc}	$\sum_{i=1}^{m} \frac{D_{AC}}{D_i} * 100$

6. Result and discussion

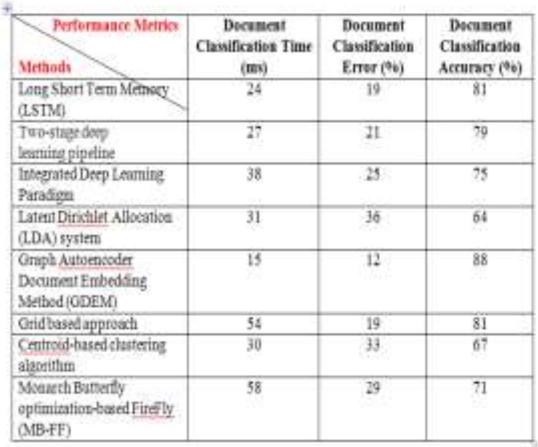
The result is based on accurate document classification model. The performance result is based on the number of keywords in the document. Performance evaluation metrics are employed to determine the effectiveness and performance of document classification. The performance of existing document classification methods is discussed with help of table and graphs.

Table 2 Tabulation for Experimental Results of CACM Dataset



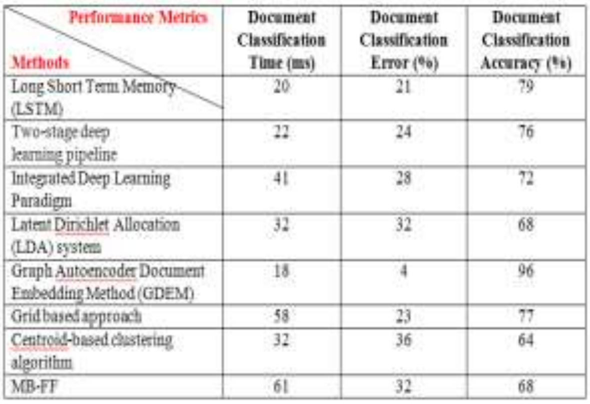
Performance Metrics Methods	Document Classification Time (ms)	Document Classification Error (%)	Document Classification Accuracy (%)
Long Short Term Memory (LSTM)	15	12	88
Two-stage deep learning pipeline	18	17	83
Integrated Deep Learning Paradigm	35	21	79
Latent Dirichlet Allocation (LDA) system	28	31	69
Graph Autoencoder Document Embedding Method (GDEM)	12	10	90
Grid based approach	51	15	85
Centroid-based clustering algorithm	26	28	72
Monarch Butterfly optimization-based FireFly (MB-FF)	54	25	75

Table 3 Tabulation for Experimental Results of Cranfield Dataset



Performance Metrics Methods	Document Classification Time (ms)	Document Classification Error (%)	Document Classification Accuracy (%)
Long Short Term Memory (LSTM)	24	19	81
Two-stage deep learning pipeline	27	21	79
Integrated Deep Learning Paradigm	38	25	75
Latent Dirichlet Allocation (LDA) system	31	36	64
Graph Autoencoder Document Embedding Method (GDEM)	15	12	88
Grid based approach	54	19	81
Centroid-based clustering algorithm	30	33	67
Monarch Butterfly optimization-based FireFly (MB-FF)	58	29	71

Table 4 Tabulation for Experimental Results of (10) Dataset Text Document Classification



Performance Metrics Methods	Document Classification Time (ms)	Document Classification Error (%)	Document Classification Accuracy (%)
Long Short Term Memory (LSTM)	20	21	79
Two-stage deep learning pipeline	22	24	76
Integrated Deep Learning Paradigm	41	28	72
Latent Dirichlet Allocation (LDA) system	32	32	68
Graph Autoencoder Document Embedding Method (GDEM)	18	4	96
Grid based approach	58	23	77
Centroid-based clustering algorithm	32	36	64
MB-FF	61	32	68

Table 2, Table 3 and Table 4 describe the performance results for eight different methods with three performance metrics for considered three datasets respectively. From the table, it is observed that Graph Auto encoder Document Embedding Method (GDEM) attained better results than any other existing methods for three datasets. For CACM Dataset, GDEM attained 90% of document classification accuracy, 12ms of document classification time and 10% of document classification error rate. For Cranfield Dataset, GDEM attained 88% of document classification accuracy, 15ms of document classification time and 12% of document classification error. For Dataset Text Document Classification, GDEM attained 96% of document classification accuracy, 18ms of document classification time and 4% of document classification error. Figure 2, 3 ,4, 5, 6, 7 ,8, 9 and 10 exemplify pictorial representation of CACM Dataset, Cranfield Dataset and (10) Dataset Text Document Classification.

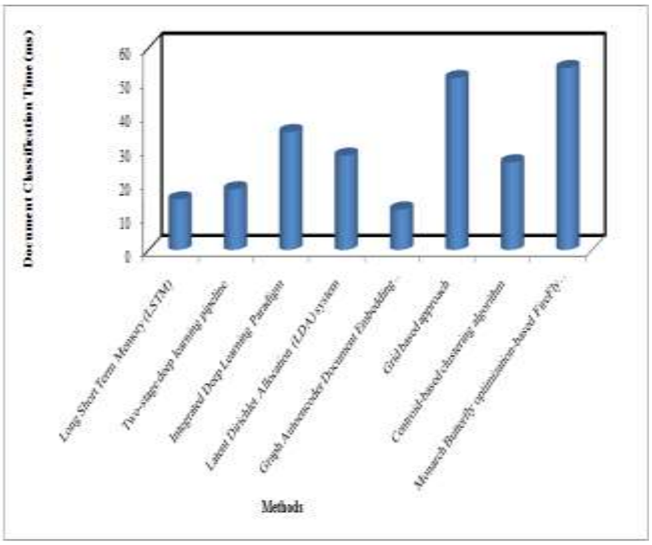


Figure 2 Pictorial representation of CACM Dataset using Document Classification Time

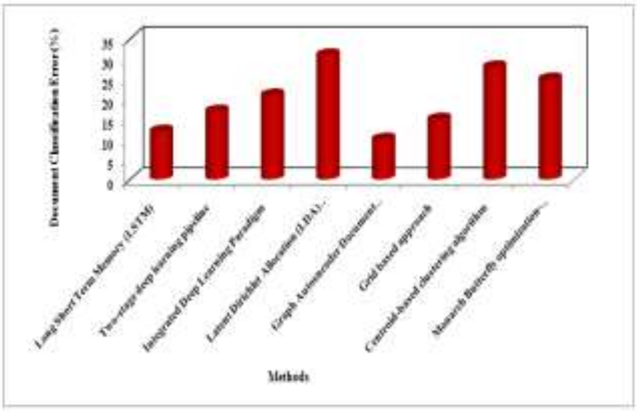


Figure 3 Pictorial representation of CACM Dataset using Document Classification Error

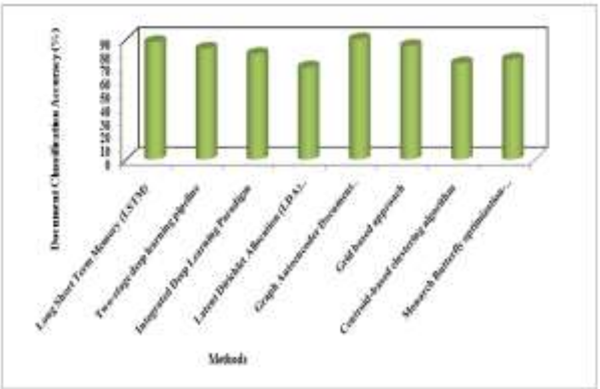


Figure 4 Pictorial representation of CACM Dataset using Document Classification Accuracy

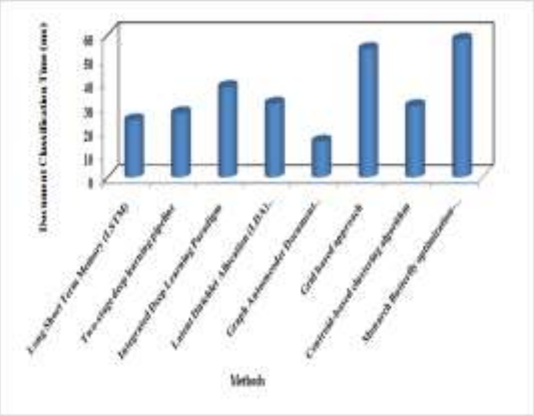


Figure 5 Pictorial representation of Cranfield Dataset using Document Classification Time

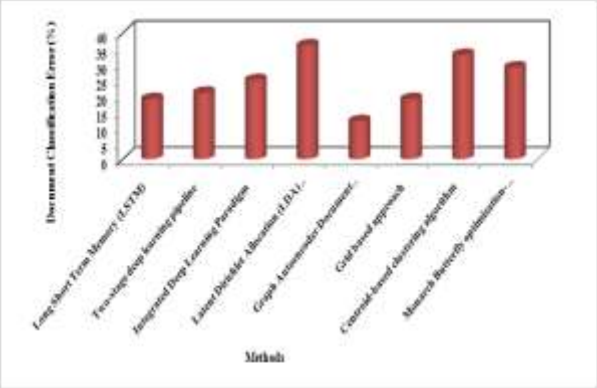


Figure 6 Pictorial representation of Cranfield Dataset using Classification Error

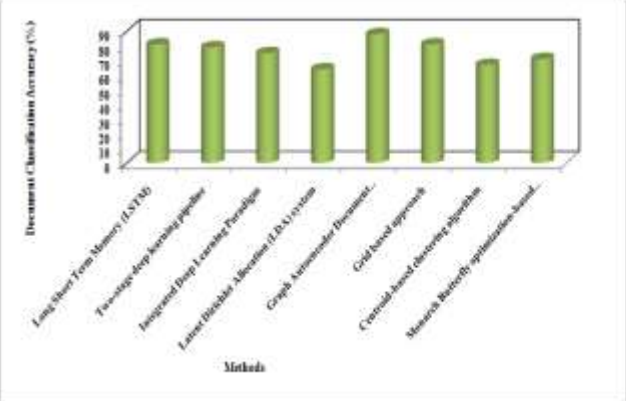


Figure 7 Pictorial representation of Cranfield Dataset using Document Classification Accuracy

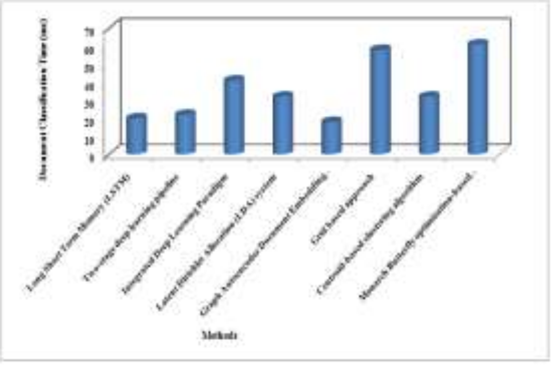


Figure 8 Pictorial representation of 10 Dataset Text Document Classification using Document Classification Time

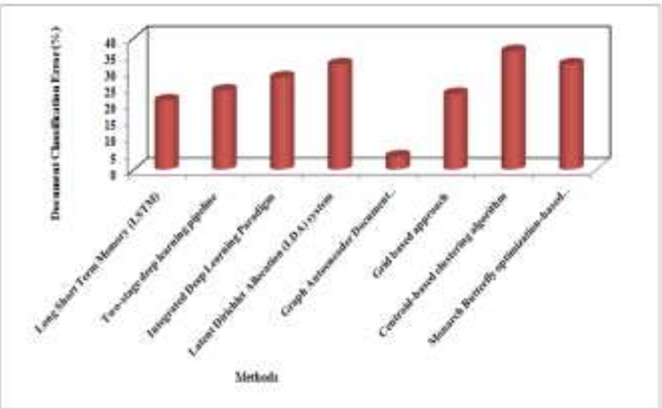


Figure 9 Pictorial representation of 10 Dataset Text Document Classification using Classification Error

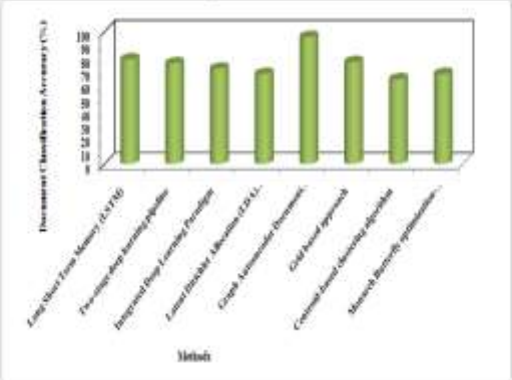


Figure 10 Pictorial representation of 10 Dataset Text Document Classification using Document Classification Accuracy

Figure 2, 3 , 4, 5, 6, 7 ,8, 9 and 10 exemplify the performance metric comparison of eight different existing methods for three different datasets respectively. In above figure, describe the clear explanation of GDEM is better than any other techniques for three different datasets. This is due to the application of undirected and weighted sparse graph. All weighted edges in graph include elevated similarities among two end nodes. Graph auto encoder determined node embedding vectors. This in turn helps to improve the document classification performance. For CACM Dataset, GDEM increased the document classification accuracy by 15% and reduced document classification time by 55%, and document classification error by 48% when compared to other existing methods For Cranfield Dataset, GDEM increased the document classification accuracy by 20% and reduced document classification time by 56% and document classification error by 51% when compared to other existing methods. For 10 Dataset Text Document Classification, GDEM increased the document classification accuracy by 34% and reduced document classification time by 44% and document classification error by 85% when compared to other existing methods.

7. Conclusion

In the survey, document classification is carried out with number of document using dissimilar ML and DL methods. Different performance metrics are employed to examine document classification results with help of extracted keywords. When the result is discussed with another study, extracted keywords with machine learning and deep learning method has attained improved document classification results. In the study, it is evident that GDEM has attained higher accuracy for three different datasets. Graph Auto encoder deep learning has achieved highest accuracy result than all the other existing techniques. These results strongly suggest that DL with technical extracted keywords can be implemented in future for efficient document classification instead of the other conventional machine learning classifier discussed in existing studies. The limitation of our proposed method failed to consider the domain of the query and minimum packet ratio. In order to, focus on excellence of this database through agreements metrics, investigating advanced deep learning and to classify explicit polarity depend on contextual location of text for sentiment categorization when subjected to numerous databases.

REFERENCES

1. Zhu Yishun, Wu Guoyue, Liang Yi, Ma Yige and Wu Jiangwei, "Classification of Distribution Network Planning Documents Based on LSTM Neural Network", Procedia Computer Science, Elsevier, Volume 228, 2023, Pages 914-919
2. Brad Hershowitz, Melinda Hodkiewicz, Tyler Bikaun, Michael Stewart and Wei Liu, "Causal knowledge extraction from long text maintenance documents", Computers in Industry, Elsevier, Volume 161, October 2024, Pages 1-15
3. Peter Atandoh, Fengli Zhang, Daniel Adu-Gyamfi, Paul H. Atandoh and Raphael Elimeli Nuhoho, "Integrated deep learning paradigm for document-based sentiment analysis", Journal of King Saud University - Computer and Information Sciences, Elsevier, Volume 35, Issue 7, July 2023, Pages 1-15
4. Qinjun Qiu, Miao Tian, Liufeng Tao, Zhong Xie and Kai Ma, "Semantic information extraction and search of mineral exploration data using text mining and deep learning methods", Ore Geology Reviews, Elsevier, Volume 165, February 2024, Pages 1-16
5. Sungwon Jung and Sangmin Ka, "GAE-Based Document Embedding Method for Clustering", IEEE Access, Volume 10, December 2022, Pages 130089 - 130096
6. Yannick-Ulrich Tchantchou Samen, "A Semantic Similarity Measure for Scholarly Document Based on the Study of n-gram", Journal of Web Engineering, Volume 21, Issue 7, October 2022, Pages 2095 – 2113
7. Arsen Yeghiazaryan, Khachatur Khechoyan, Grigor Nalbandyan and Sipan Muradyan, "Tokengrid: Toward more Efficient Data Extraction from Unstructured Documents", IEEE Access, Volume 10, April 2022, Pages 39261 – 39268

8. Pablo Camarillo-Ramirez, Francisco Cervantes-Alvarez and Luis Fernando Gutiérrez-Preciado, "Semantic Maps for Knowledge Graphs: A Semantic-Based Summarization Approach", IEEE Access, Volume 12, January 2024, Pages 6729 - 6744
9. Mamta Kayest and Sanjay Kumar Jain, "Optimization driven cluster based indexing and matching for the document retrieval", Journal of King Saud University - Computer and Information Sciences, Elsevier, Volume 34, Issue 3, March 2022, Pages 851-861
10. Ram Kumar, S. C. Sharma, "Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval", The Journal of Supercomputing, Springer, Volume 79, 2023, Pages 2251–2280
11. Neha Agarwal, Geeta Sikka, Lalit Kumar Awasthi, "Enhancing web service clustering using Length Feature Weight Method for service description document vector space representation", Expert Systems with Applications, Elsevier, Volume 161, 2020, Pages 1-11
12. Peng Yang, Yu Yao, Huajian Zhou, "Leveraging Global and Local Topic Popularities for LDA-Based Document Clustering", IEEE Access, Volume 8, 2020, Pages 24734 – 24745
13. Chongyi Liu, Xiangyu Wang and Honglei Xu, "Text Classification Using Document-Relational Graph Convolutional Networks", IEEE Access, Volume 10, November 2022, Pages 123205 - 123211
14. Jing Peng and Shuquan Huo, "Few-Shot Text Classification Method Based on Feature Optimization", Journal of Web Engineering, Volume 22, Issue 3, May 2023, Pages 497 - 514
15. Yongchun Gu, Yi Wang, Heng-Ru Zhang, Jiao Wu and Xingquan Gu, "Enhancing Text Classification by Graph Neural Networks With Multi-Granular Topic-Aware Graph", IEEE Access, Volume 11, February 2023, Pages 20169 - 20183
16. Maroua Louail, Chafia Kara-Mohamed Hamdi-Cherif and Aboubekeur Hamdi-Cherif, "Tasneef: A Fast and Effective Hybrid Representation Approach for Arabic Text Classification", IEEE Access, Volume 12, August 2024, Pages 120804 - 120826
17. Rasim Cekik, "A New Filter Feature Selection Method for Text Classification", IEEE Access, Volume 12, September 2024, Pages 139316 – 139335
18. Xinjie Sun and Xingying Huo, "Word-Level and Pinyin-Level based Chinese Short Text Classification", IEEE Access, Volume 10, November 2022, Pages 125552 - 125563
19. Qiang Liu, Jingzhe Chen, Fan Chen, Kejie Fang, Peng An and Yiming Zhang, "MLGN: A Multi-Label Guided Network for Improving Text Classification", IEEE Access, Volume 11, July 2023, Pages 80392 – 80402
20. Thanakorn Thaminkaew, Piyawat Lertvittayakumjorn and Peerapon Vateekul, "Prompt-Based Label-Aware Framework for Few-Shot Multi-Label Text Classification", IEEE Access, Volume 12, February 2024, Pages 28310 – 28322
21. Jinpeng Bai and Xinfu Li, "Chinese Multilabel Short Text Classification Method Based on GAN and Pinyin Embedding", IEEE Access, Volume 12, June 2024, Pages 83323 – 83329
22. Xinlei Cao, Jinyang Yu and Yan Zhuang, "Injecting User Identity into Pretrained Language Models for Document-Level Sentiment Classification", IEEE Access, Volume 10, March 2022, Pages 30157 – 30167
23. Hyun Kwon, "Dual-Targeted Textfooler Attack on Text Classification Systems", IEEE Access, Volume 11, October 2021, Pages 15164 – 15173

24. Tengfei Liu, Yongli Hu, Boyue Wang, Yanfeng Sun, Junbin Gao and Baocai Yin "Hierarchical Graph Convolutional Networks for Structured Long Document Classification", IEEE Transactions on Neural Networks and Learning Systems, Volume 34, Issue 10, October 2023, Pages 8071 – 8085
25. Kun Zhang, Le Wu, Guangyi Lv, Enhong Chen, Shulan Ruan and Jing Liu, "Description-Enhanced Label Embedding Contrastive Learning for Text Classification", IEEE Transactions on Neural Networks and Learning Systems, Volume 35, Issue 10, October 2024, Pages 14889 - 14902
26. Yunlong Gao and Xiaogang Peng, "Same-Radical Character Association Model in Text Classification", IEEE Transactions on Consumer Electronics, Volume 70, Issue 1, February 2024, Pages 1000 – 1009
27. Arda Can Aras, Tuna Alikasifoglu and Aykut Koc, "Graph Receptive Transformer Encoder for Text Classification", IEEE Transactions on Signal and Information Processing over Networks, Volume 10, March 2024, Pages 347 – 359
28. Lin Xiao, Pengyu Xu, Mingyang Song, Huafeng Liu, Liping Jing and Xiangliang Zhang, "Triple Alliance Prototype Orthotist Network for Long-Tailed Multi-Label Text Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume 31, Pages 2616 - 2628