IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 29, 2021, accepted September 16, 2021, date of publication September 23, 2021, date of current version October 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3115148

A Comprehensive Evaluation of Metadata-Based Features to Classify Research Paper's Topics

GHULAM MUSTAFA[1], MUHAMMAD USMAN[2], MUHAMMAD TANVIR AFZAL[3], ABDUL SHAHID[4], AND ANIS KOUBAA[5,6]

[1]Department of Computer Science, Capital University of Science & Technology, Islamabad 46000, Pakistan
[2]Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad 44000, Pakistan
[3]Department of Computer Science, Namal Institute Mianwali, Mianwali 42250, Pakistan
[4]Institute of Computing, Kohat University of Science and Technology, Kohat 26000, Pakistan
[5]Robotics and Internet of Things Lab, Prince Sultan University, Riyadh 12435, Saudi Arabia
[6]CISTER/INESC-TEC, Polytechnic Institute of Porto, 4200 Porto, Portugal

Corresponding author: Abdul Shahid (ashahid@kust.edu.pk)

The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

ABSTRACT The existing plethora of document classification techniques exploits different data sources either from the content or metadata of research articles. Various journal publishers like Springer, Elsevier, IEEE, etc., do not provide open access to the content of research articles, whereas metadata is freely available there. Metadata like title, keyword, and abstract can serve as a better alternative to the content in various scenarios. In the current literature, researchers have assessed the role of some of the metadata individually. We believe that the collective contribution of metadata parameters can play a significant role in classifying research papers. This paper presents a comprehensive evaluation of the role of metadata, individually as well as in combinations to achieve the objective of research paper classification. Moreover, we have classified the research articles into ACM hierarchy root categories (e.g. general literature, hardware, software, etc.). In this comprehensive evaluation, we have assessed all the possible combinations of metadata features against different classifiers such as Random Forest, K Nearest Neighbor, and Decision Tree. The results of this research reveal that the title & keywords combination outperforms other combinations with an F-measure score of 0.88.

INDEX TERMS Research paper classification, Word2Vector (W2V), metadata, association of computing machinery (ACM), k-nearest neighbor's (KNN), decision tree (DT), random forest (RF), term frequency (TF), term frequency and inverse document frequency (TFIDF), bag of word (BOW).

I. INTRODUCTION

Over the past several years, the research plethora over the web is briskly expanding. This considerable amount hinders the recommender systems from extracting the relevant research papers against the posed query. Besides, the area of research paper classification has grabbed the prime interest of the scientific community [1]. The classification of scholarly literature can assist in multifarious aspects by helping scholars like 1) Helping researchers to find the relevant papers, 2) Finding relevant literature to narrate the background concept of the proposed study, and so on. Typically, users explore different repositories to extract the relevant research papers, such as Digital Library, Google Scholar, etc. However, the existing data over the web is unstructured, which adversely affects finding relevant information against the posed query. These systems return millions of generic hits. Let us consider an example, when a user poses a query on Google Scholar, it yields around 2.8 million relevant research papers. Most of these papers do not even belong to the domain of the query. Reading all of these papers requires a lot of time. Almost 158 years are required to read five papers per day. This is because papers over these repositories are not correctly classified or indexed according to their respective classes. We believe that the performance of these systems can be enhanced if these papers are labeled to their respective domains. This will improve the functionality of existing recommender systems and help the scholars to find the relevant content or to conduct a literature review or survey of a particular domain efficiently. In literature, researchers have proposed diversified techniques for research papers' classification. These techniques are grouped into citations, meta-data, content-based and hybrid approaches [2]–[5]. Among these approaches, metadata holds significant

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

potential due to its free availability. Researchers have exploited different metadata-based features, including title, keyword, abstract and general terms, and then harnessed them individually or collectively to classify research papers into different categories [6], [7]. Based on critical analysis of contemporary approaches, we have identified that as per our knowledge none of them has combined useful metadata parameters like abstract, title, keywords, general terms, etc., to classify research papers using comprehensive and large datasets. Each metadata parameter of research papers holds significant potential, and their collective contribution can improve accuracy [8].

A. RESEARCH PROBLEM

This study presents a comprehensive evaluation of metadata of research papers individually and then collectively by forming different combinations to classify research papers into different categories specified by ACM. For this purpose, the comprehensive and large dataset is taken from ACM prepared by Santos [9]. It contains different metadata parameters of research articles from the domain of Computer Science. From this data, we have extracted title, abstract, general terms, and keywords. Similar to [4], [5], [10], this study also utilizes the ACM categorization system to assign the corresponding labels to research papers. The ACM Computing Classification System (CCS) is a subject classification system for computing devised by the Association for Computing Machinery (ACM). CSS is used by the various ACM journals to organize subjects. There are three levels of ACM Computing Classification System (ACM CCS). The proposed model classifies papers onto top-level ACM categories, as shown in Fig 1. We use words' embeddings as the latest semantic-based similarity computation approach for feature transformation instead of traditional count-based techniques [11]–[13]. The experimental results revealed that the combination of title and keywords metadata features had outperformed the rest of the metadata features by achieving 88% f-measure against K nearest Neighbor classifier followed by the combination of a title and abstract with 0.87 f-measure. We claim that the proposed approach is a valuable contribution to the research papers' classification community.

This paper is organized as follows. Section 2 explained the detailed study of contemporary state-of-the-art approaches proposed in the document classification area. Section 3 discussed the methodology of our proposed model in detail. Moreover, this section also describe datasets, a technique for processing data, and classifiers used in this paper. Section 4 describe the results of all possible combination of metadata features in detail. In the end, section 5 concludes our research.

II. RELATED WORK

This section sheds a light on contemporary state-of-the-art techniques about document classification. Different techniques have been proposed to automatically map the research papers onto different categories. Nanba [14] presented a study that classifies research papers using citation links and types based features. Based on this study, authors have developed a research papers classification tool named PRESRI. The tool employs authors' names or title words-based features. PRESRI'S current version takes the features into account and categorizes the papers based upon the cited paper mentioned in the bibliography of the query paper.

In another study [15], authors have presented a scheme to perform subject classification of scientific articles based on analysis of their interrelationship. The study has employed citations, common author, and common reference-based parameters. To do this, a relationship graph has been formed wherein research papers have been presented by nodes, and links among those nodes determine the relationship between papers. The outcomes of that study revealed that good results are produced against dense and close-packed graphs.

In [16] and [17], researchers have employed a reference segment of a research paper to locate the topics of the paper. The study follows an assumption that most of the time, authors cite the papers belonging to the same domain or similar category. To validate the claim, the authors have employed a data set from the Journal of Universal Computer Science (J.UCS). The stored references in the database have been matched with the extracted references of the paper.

In another study, Bayesian-based approach has been presented to classify research papers [18]. In this study, 400 research papers from education conferences have been considered as a data set and mapped to four different classes including e-learning, cognition issues, teacher instruction, and intelligent coaching system. The researchers have contended that there are keywords traits that can be used for categorizing the papers. The approach is solely based on keywords-based features.

In [11], research papers have been classified into multiple categories using phrase-to-text connectedness measures. Authors have employed three measures to evaluate the proposed approach: i) cosine relevance score between typical vector area representations of the texts coded with tf-idf weighting; ii) famous characteristic of the likelihood of term generation BM25; and iii) an in-house characteristic of the conditional probability of symbols average over matching fragments in suffix trees representing texts and phrases, CPAMF. Moreover, they have considered a collection of research papers abstract from the ACM digital library for experiments. Their experimental results showed that the CPAMF outperforms both the cosine measure and BM25 by a healthy margin.

Another text document classification technique based on similarity has been presented in [19]. In their similarity-based categorization framework, classification is based on tiny level documents and because of less range of documents achieve completeness. As there was small information (vocabulary), so their technique was quick however not appropriate for massive vocabulary. The performance of their technique (algorithm) was slightly lower than expected. Comparison against the performance with SVM, Rocchio

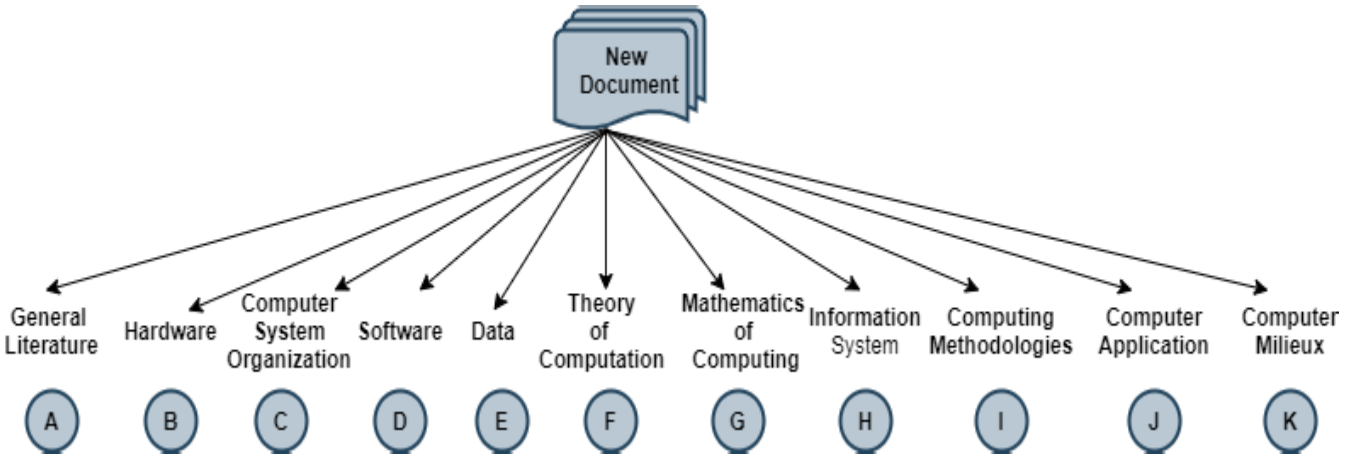


FIGURE 1. Top level categories in ACM hierarchy.

algorithm, Bayes, Naïve Bayes is mentioned in the paper, however, authors have not provided the table or graph results.

Some authors proposed hybrid approaches for textual document classification [20]–[22]. In hybrid approaches, the algorithms focused on both feature extraction using deep learning and classification using machine and deep learning. These approaches reported good results but employed the whole text of the documents. In the context of classification of research papers, Balys [23] proposed an automatic classification of scientific text-primarily based on applied math analysis of probabilistic distributions of scientific terms in texts. Some works like [24] focus on machine learning algorithms to develop subject classification rules for documents. However, most of the works of this kind, process whole text of papers to extract the options and develop a classifier which is time consuming [25]. Another study proposed the model using whole content as a feature [26]. This approach used naive Bayes and Logistic regression algorithms. They used two diversified datasets from the computer science domain which have already been annotated i.e. 1) CiteSeerX, 2) arXiv. They concluded that the achieved F1 Score on arXiv and CiteSeerX datasets are 0.95 and 0.75 respectively. XiaoyuLuo [27] used the SVM model for classifying English text in papers. They have conducted two analytical experiments to check the selected classifiers using English documents. The dataset used for the experiment contains 1033 text documents. The results presented that the Rocchio classifier provides the best performance when the size of the feature set is small while SVM outperforms the other classifiers. Moreover, they observed that the classification rate exceeds 90% when using more than 4000 features.

Some of the recent techniques [28], [29] employed different deep learning models to conduct the document classification and yielded good results. But these techniques are time-consuming. One of these techniques [30] proposed a novel multimodal deep learning architecture, called TechDoc, for technical document classification. This architecture utilizes both natural language and descriptive images to train hierarchical classifiers. As aforementioned, one of the strengths of our work is that we utilized the metadata instead of the whole text which makes it time-efficient. After the critical analysis of the state of the art techniques, there exist a few research papers' classification schemes that employ the metadata-based features. As per our knowledge, none of them have comprehensively evaluated the metadata features individually as well as the combination of metadata features to classify research papers into different categories. We have also observed that these techniques have also not employed comprehensive or large datasets. So after finding these gaps, we argue that metadata can play a pivotal role to hint at the category of a research paper individually as well as the combination of these features further improve the classification rate. Moreover, one of the advantages of using metadata is that it is freely available and holds significant potential [8] as compared to the whole text of the document. Table 1 presents an overview of existing document classification techniques.

III. METHODOLOGY

Figure 3 represents the methodology of our comprehensive evaluation. In this methodology, first of all, we acquired a comprehensive dataset from ACM prepared by Santos [9]. This dataset consists of metadata-based features of research papers, including Title, Abstract, Keywords, and General Terms. We selected these parameters based on the following justifications:

- The title of the paper holds potential terms that can assist in determining the category of a research article.
- The abstract is not part of metadata, but similar to different metadata parameters, it is also freely available and presents the main theme of the paper that can hint at the corresponding category.
- Keywords and general terms are explicitly assigned by the actual authors of the papers that are mostly from relevant areas.

TABLE 1. Summary of reviewed literature.

Title/Reference	Techniques/Methodology	Dataset	Results	Limitations
(Hidetsugu NANBA, 2000)	BCCT-C (Bibliographic coupling)and PERSI Tool as a Prototype	395 Paper in TEX Style	Recall= 0.80 Precision=0.76	Only limited to type C Citations for document Classification Limited dataset
(Mohsen Taheriyan, 2011)	Analysis of Interrelationships of authors, references, and citations	Extracted 255 Paper from ACM Digital Library	Precision =0.90 (On Small dataset). Precision 0.80 (On Largedataset)	Not much Effective for the Larger size of test cases (papers) Limited dataset
(Nasser, 2011)	Citation Based category identification	Journal of Universal Computer Science (J.UCS), 1460 document	Accuracy =0.70	Only limited to reference section, not using other metadata which may give better results. Limited dataset
(Kok-Chin Khor, 2006)	esiaBayesian Network. feature selection algorithm	400 conference papers	Average Accuracy = 83.75	Classification Approach Limited to Conference papers with Respect to topic Use only keyword metadata Limited dataset
(Ekaterina Chernyak, 2015)	CPAMF (characters averaged over matching fragments Measure) cosine and BM25	ACM Digital Library.	Nill	Used Limited dataset
(S. Senthamarai Kannan,2008)	Similarity-based learning algorithm and thresholding Strategies	Standard Collection for text classification the Reuters~(21578)	Recalland Precision =0.90 (For 2000 sample dataset)	This method is designed for only a small number of datasets.
(Vaidas BALYS,2010)	The scientific texts classification methodology	15,000 articles provided by the Institute of Mathematical Statistics	Average precsion upto = 0.67	For better evaluation, it needs very large size datasets.
(LioX et al, 2021)	UsedSVM model	1033text document	Accuracy 0.90	Limiteddataset Use the overall content of the document for classification
(Mustafa et al 2021) Proposed Approach	Used different classifiers	ACM Dataset	F-Measure: 0.88	Limited to Single Label classification

After extraction, we have performed some preprocessing steps on the dataset for the removal of noisy data. Afterward, we have transformed the features by making all the possible combinations of extracted metadata. Moreover, in the next step, we transformed these features into numeric form. Then we divided our dataset into two sections, training dataset and testing dataset in 80:20 ratios. The training data is used for classifier training and the testing data is used to evaluate the classifier model. After training the classifier, we provided the new test sample to the classifier model, so that it can predict the label for the given sample. Then the system compares the predicted label with the actual label of that sample and reports the result in the form of different evaluation parameters.

A. DATASET

The selection of an appropriate data set plays a significant role. We employed a comprehensive data set prepared by Chernyak [11] from the ACM digital library. We have chosen this dataset because it contains data from different research areas. This will help in the concrete assessment of the proposed methodology. It consists of 11 classes A (General Literature), B (Hardware), C (Computer system organization),

D(Software), E (Data), F (Theory of Computation), G (Mathematics of Computing), H (Information Systems), I (Computing methodologies), J (Computers Application), and K (Computing Milieux). Each class is represented by many instances, whose total is 86116 data items or records. In each record there are five parameters such as 1) Title of the Paper, 2) Abstract, 3) Keywords, 4) General Terms, and 5) Classification Label. The first four parameters are used as features and the last parameter is used as a classification label. These parameters are evaluated individually as well as all possible combinations of these parameters. Some of the data items belong to multiple categories, for which we converted the records into a single label by duplicating the records with the second or third label of the same record. Table 2 and figure 2 show some statistics and distribution of all parameters of the dataset individually as well as in combination respectively.

TABLE 2. Dataset statistics.

Dataset Statistics	
Number of documents	86116
Number of parameters	5
Number of documents with the title	86116
Number of documents with abstract	53963
Number of documents with one or more categories	54994
Number of documents with one or more keywords	23971
Number of documents with one or more general terms	51574

B. PRE-PROCESSING

Before starting experiments we converted the data into a specific format. For this we applied the following pre-processing techniques:

1) NOISE REMOVAL

The first preprocessing technique that we applied is Noise Removal. Noise is an unavoidable problem, which affects the data collection and data preparation processes in Data Mining applications that can result in errors. Noise has two main sources [31] one is implicit errors introduced by measurement tools, such as different types of sensors. The other is random errors introduced by batch processes or experts when the data are gathered, such as in a document digitalization process. In our dataset, there are some random errors introduced during data gathering from real-world problems that often suffer from corruption that may hinder the performance of the system in terms of accuracy [32]. There are two broad categories of noise [31]. The first is Class Noise which contains (Contradictory examples and Mislabeled Examples). The second class is Attributes Noise which contains (Erroneous values, Missing values and Don't care values). In our case, the dataset contains the noise is in the form of missing values. In literature, the missing values are handled by multiple techniques such as deletion of records [33], [33], Mean Substitution [34], Last observation carried forward [35], etc. As there are very few records with missing values in our dataset so we used the deletion technique as it is deemed as the simplest and efficient method for handling the missing data. Following this, we deleted the instances with missing values.

2) STOP WORDS REMOVAL

For the Optimization of data analytic processes, stop word removal becomes the most important factor. For achieving better accuracy extraction of redundant words with low or no semantic meaning must be filtered out. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK (Natural Language Toolkit) in python has a list of stop words stored in 16 different languages. In this paper, we used English based dictionary in which list of stop words is predefined, compared to the target text on which removal is required.

C. FEATURE TRANSFORMATION

For comprehensive evaluation of metadata features of research articles, we have transformed the individual metadata features title, keyword, abstract, general terms into all of their possible combinations. The process has formed total 15 combinations, which are listed below.

- Title
- Abstract
- Keywords
- Keywords
- Title + Abstract
- Title + Keywords
- Title + Generals Term
- Abstract + Keywords
- Abstract + Generals Term
- Keywords+ general Terms
- Title + Abstract + Keywords
- Abstract + Keywords + Generals Term
- Title + Keywords + Generals Term
- Title + Abstract + Generals Term
- Title + Abstract + Keywords + Generals Term

D. WORD EMBEDDING MODEL

Most of the similarity measures and machine learning algorithms often take a numeric vector as an input. However, before performing any operation on a text, we need a way to convert each document into numeric vectors. This is one of the fundamental problems in data mining, which aims to numerically represent the unstructured text documents to make them mathematically computable. For this, numerous techniques have been presented in the literature. These techniques are mainly divided into two broad Categories such as Count based approaches and Semantic-based approaches. Some of the widely used count-based techniques in research articles classification approaches are: 1) One Hot Encoding 2) Bag of Word (BOW) or Term Frequency (TF), 3) Term Frequency and Inverse Document Frequency (TFIDF).

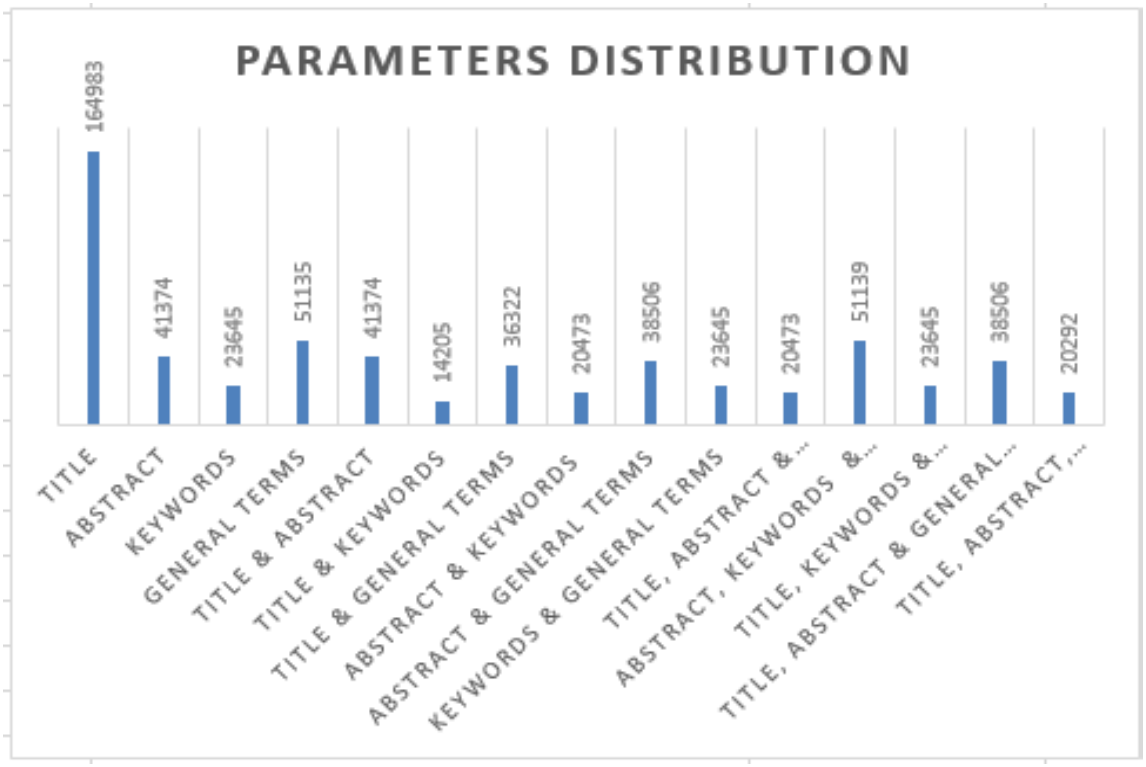


FIGURE 2. Feature combination distribution (Total records).

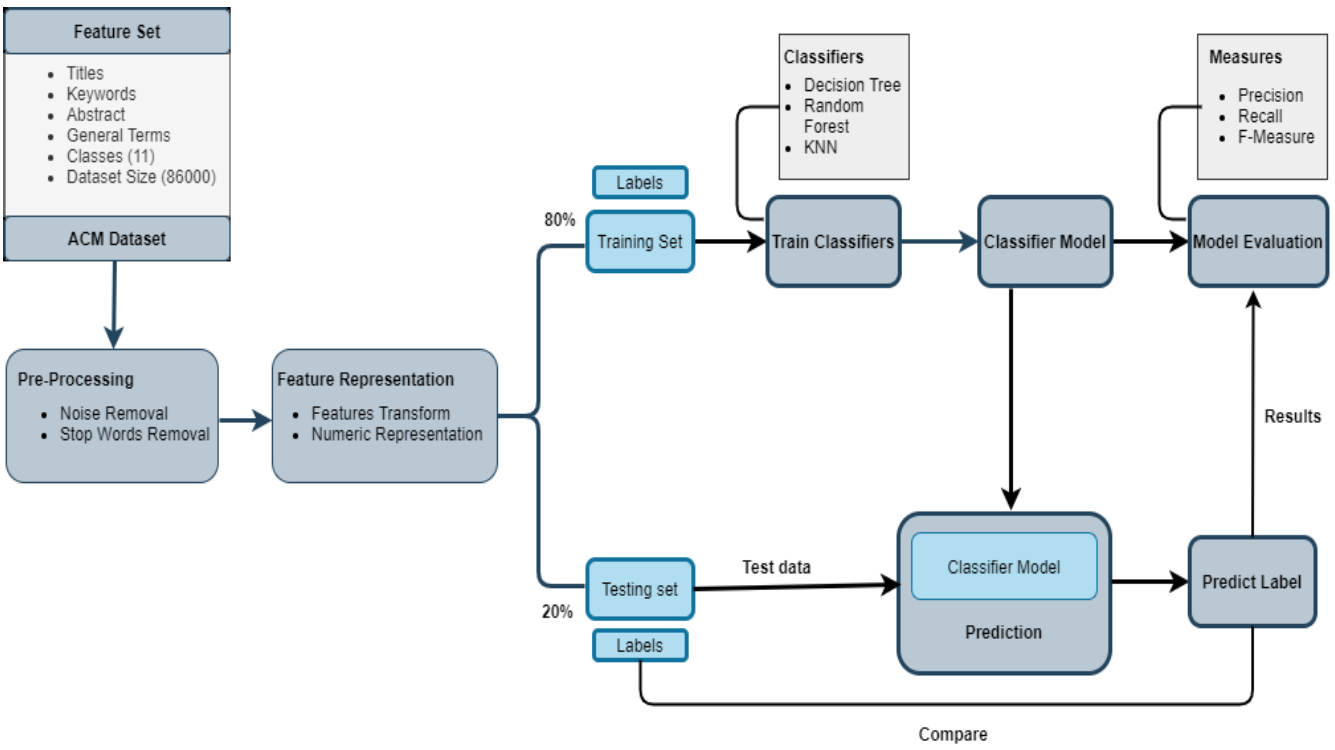


FIGURE 3. Architecture diagram.

For capturing information these techniques rely on the frequency of terms and ignored the semantic and context of the term. The current state-of-the-art approaches [5], [7], [19], [36]–[38] for research article classification have employed these count based techniques measures like TF, BOW, and TFIDF, etc. due to which they have ignored the semantic and

contextual information of terms and it might be assigned a wrong category to the research articles.

To address the above-mentioned issue we have to use some alternative techniques for representation which is considered semantic and contextual information. By literature survey, we found word embedding technique that is one of the most well-known techniques used in different domains [39]–[41]. Word embedding is used to represent document vocabulary. It is capable of capturing the context of a word in a document, semantic and syntactic similarity, relation with other words, etc. For this, Word2Vec is one of the most popular techniques to learn word embedding's using a shallow neural network. It was developed by Mikolov at Google [42]. These models are shallow, two-layer neural networks equipped to reconstruct linguistic contexts of words.Word2vec takes as its source a wide corpus of text and generates a vector space, usually several hundred dimensions, with a corresponding vector in space being allocated to each specific word in the corpus. Word vectors are placed in the space of the vector so that terms sharing common meanings in the corpus are located in space near each other. Afterward, the similarities score between identical words is greater than two different words which show that these models also capture the semantic and context of words.

So that's why we have used the word2vec -pre-trained (trained on Google news) model to generate the vector for each instance of each combination. The word embedding model generated a vector of 75* 4 lengths, which consists of 300 elements. Each instance of record consists of a different number of sentences, which are further transformed into a single vector.,In the end, all of the metadata combinations are represented by meaningful vectors.

E. CLASSIFIER

To evaluate the proposed methodology, three machine learning classifiers, (1) K Nearest Neighbor, (2) Random Forest, and (3) Decision tree classifier, are applied to the features using PyCharm. The k-nearest-neighbor classifier [43] is a supervised classification algorithm, which takes a bunch of labeled points and uses them to learn how to map other points. To map a point on a label, it considers its nearest neighbor. Random forest [44] is a tree-based classification algorithm, which involves the building of several decision trees. It combines the output produced by all the decision trees to improve the generalization ability of the model. Decision Tree Classifier [45] is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is found. For obtaining the best user-defined values or parameters for the above-mentioned classifiers we repeated the experiments multiple times by slightly changing the parameters' values. In the end, we achieved the best parameters values on which yielded maximum results on classifiers. The values of parameters are presented in table 3.

TABLE 3. Parameter values.

Classifiers	Parameters Value
K-Nearest Neighbor	n_neighbors=4 weights=uniform algorithm=auto
Random Forest	n_estimators=50 max_depth=none min_samples_split=3 min_samples_leaf=2 max_features=auto
Decision Tree	Splitter= random max_depth=none min_samples_split=3 min_samples_leaf=2

F. TRAINING

The data set is divided in the ratio of 0.6:0.4. 0.6 is for training and 0.4 for testing. We have applied a supervised learning method to train our model.

G. CLASSIFICATION

The classification is a process of taking a classifier and running it on unknown content to determine class membership for the unknown content. The remaining data set is reserved for classification given to the trained classifier, and the classifier can predict the label based on its training. After that, the system compared the predicted labels with the actual labels and reports the result in the form of different evaluation measures.

H. EVALUATION

To evaluate the proposed methodology, we have employed the standard evaluation measures, precision-recall, and F-measure. The reason for the selection of these evaluation parameters is the frequent reporting of these parameters in literature [16], [18], [26].

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{1}$$

Precision is the fraction of the relevant documents that are relevant to the query.

$$Recall = \frac{TruePositive}{TruePositive + TrueNegative} \tag{2}$$

The recall is the fraction of the relevant documents that are successfully retrieved. We have also calculated the F-measure, which is the harmonic mean of precision and recall.

$$F1measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

The values of these evaluation parameters are in the range of 0 to 1. Where 0 represents the lowest result and 1 represents the highest result i.e 100 percent.

IV. RESULT

This section presents the results attained after conducting the comprehensive evaluation of metadata features in different ways i.e. evaluation of individual features and evaluation of their multiple combinations.

A. INDIVIDUAL FEATURES

In individual feature assessment, the individual role of abstract, keywords, general terms, and title are analyzed. The outcomes of the metadata feature against different classifiers are shown in Fig. 4. As it can be inferred from the figure the metadata feature ''abstract'' has performed extraordinary by achieving 0.87 f-measures against the KNN classifier. The metadata feature ''keyword'' has achieved the second-highest score of f-measure i.e., 0.86 against KNN neighbor followed by the ''Title'' feature with f-measure 0.83.

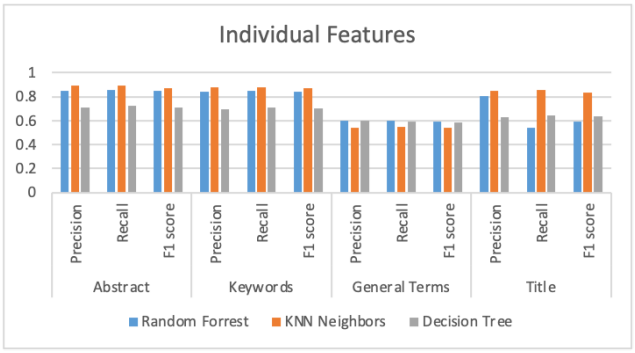


FIGURE 4. Individual features result.

B. COMBINATION OF BI-FEATURES

After scrutinizing the individual contribution of each metadata parameter, we have formed their bi-combinations. The bi-combinations are (Title + Abstract), (Title + Keywords), (Title + Generals Term), (Abstract + Keywords), (Abstract + Generals Term). The outcomes against bi-features are shown in Fig. 5. As it can be seen in the figure that that title + keywords combination has outperformed all other bi-combinations by attaining 0.88 f-measures against KNN classifier followed by title + abstract with 0.87 f-measure and Abstract + keywords with 0.85 f-measures against KNN neighbor.

C. COMBINATION OF TRI-FEATURES

The tri combinations of employed metadata parameters are (Title + Abstract + Keywords), (Abstract + Keywords + Generals Term), (Title + Keywords + Generals Term), (title + abstracts+ Generals Term). The results of these tri-combinations against different classifiers are shown in Fig. 6. The results revealed that (title + abstracts+ Generals Term) combination has performed extraordinary by achieving 0.88 f-measures. The (Title + keywords +General Term) tri-combination has achieved second-highest value of f-measure i.e., 0.85 against KNN classifier followed by the

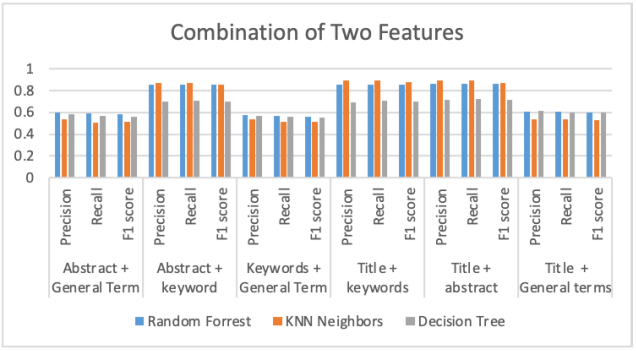


FIGURE 5. Combination of two feature result.

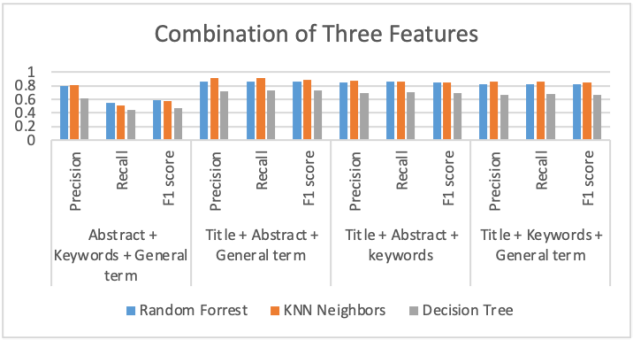


FIGURE 6. Combination of three feature result.

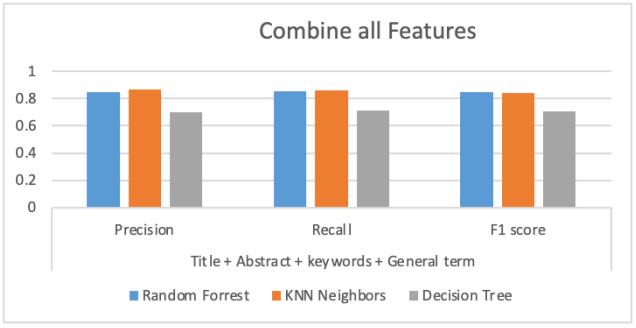


FIGURE 7. Combination of all feature result.

tri-combination (title + abstract + keywords) with 0.84 f-measure.

D. COMBINATION OF TETRA METADATA FEATURES

The results attained against a single tetra combination of metadata parameters Title + Abstract+ Keyword+ General Term are shown in Fig 7. It can be inferred from the figure that precision, recall, and F-measure score for tetra combination of metadata parameters remained higher than 0.8, and the performance of random forest and KNN classifiers are almost identical. However, the decision tree performed comparatively worse.

V. CONCLUSION

Classification of documents into predefined categories is deemed as an important research problem for the past

several years. An accurate classification model to label the research papers into different categories can boost the efficiency of various digital libraries and can also assist the scholarly community by providing them content to conduct a literature review on a particular topic or domain. Critical analysis of state-of-the-art document classification has revealed that most of the schemes have employed the content of research articles and a few of them have harnessed the metadata to classify research papers into different categories; however, the metadata-based approaches have not been assessed in the form of determining the collective role of useful metadata parameters such as title, keywords, categories, etc. In this study, we have presented a classification model that classifies research papers onto the top level of ACM categories with the help of metadata combinations based on features. For classification, we have picked Random Forest, KNN, and Decision tree Classifiers. The empirical results have revealed that title + abstracts+ Generals Term combination has outperformed by attaining an F-measure of 0.88. Among all the classifiers, KNN performed significantly better than other classifiers. The outcome of this study revealed that our approach outperformed the existing metadata-based schemes with an F-measure of 0.88. In the future, we intend to enrich the model by mapping the research papers onto the second level of ACM categories.

REFERENCES

[1] J. Beel, S. Langer, M. Genzmehr, and B. Gipp, "Research paper recommender system evaluation: A quantitative literature survey," in *Proc. Int. Workshop Reproducibility Replication Recommender Syst. Eval.*, 2013, pp. 15–22.

[2] B. Tang, H. He, P. M. Baggenstoss, and S. Kay, "A Bayesian classification approach using class-specific features for text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 6, pp. 1602–1606, Jun. 2016.

[3] H. N. L. Nguyen and Q. H. Bao, "A combined approach for filter feature selection in document classification," in *Proc. IEEE 27th Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2015, pp. 169–177.

[4] K. Chekima, C. K. On, R. Alfred, G. K. Soon, and P. Anthony, "Document categorizer agent based on ACM hierarchy," in *Proc. IEEE Int. Conf. Control Syst., Comput. Eng.*, Nov. 2012, pp. 386–391.

[5] T. Wang and B. C. Desai, "Document classification with ACM subject hierarchy," in *Proc. Can. Conf. Electr. Comput. Eng.*, 2007, pp. 792–795.

[6] P. K. Flynn, *Document Classification Support Automated Metadata Extraction Form Heterogeneous Collections*. Norfolk, VA, USA: Old Dominion Univ., 2014.

[7] N. A. Sajid, M. T. Afzal, and M. A. Qadir, "Multi-label classification of computer science documents using fuzzy logic," *J. Nat. Sci. Found. Sri Lanka*, vol. 44, no. 2, p. 155, Jun. 2016.

[8] F. Qayyum and M. T. Afzal, "Identification of important citations by exploiting research articles' metadata and cue-terms from content," *Scientometrics*, vol. 118, no. 1, pp. 21–43, Jan. 2019.

[9] G. D. Santos, "Classificação multi-etiqueta hierárquica de textos segundo a taxonomia ACM," Ph.D. dissertation, Dept. Comput. Sci., Inst. Politécnico do Porto. Inst. Superior de Eng. do Porto, Porto, Portugal, 2008.

[10] B. Zhang, M. A. Gonïalves, W. Fan, Y. Chen, E. A. Fox, P. Calado, and M. Cristo, "Combining structural and citation-based evidence for text classification," in *Proc. 13th ACM Conf. Inf. Knowl. Manage.*, 2004, pp. 162–163.

[11] E. Chernyak, "An approach to the problem of annotation of research publications," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, Feb. 2015, pp. 429–434.

[12] C. Goller, J. Löning, T. Will, and W. Wolff, "Automatic document classification-a thorough evaluation of various methods," in *Proc. ISI*, 2000, pp. 145–162.

[13] W. Zong, F. Wu, L.-K. Chu, and D. Sculli, "A discriminative and semantic feature selection method for text categorization," *Int. J. Prod. Econ.*, vol. 165, pp. 215–222, Jul. 2015.

[14] H. Nanba, N. Kando, and M. Okumura, "Classification of research papers using citation links and citation types: Towards automatic review article generation," *Adv. Classification Res. Online*, vol. 11, no. 1, pp. 117–134, 2011.

[15] M. Taheriyan, "Subject classification of research papers based on interrelationships analysis," in *Proc. 2011 Workshop Knowl. Discovery, Modeling Simulation*, 2011, pp. 39–44.

[16] N. A. Sajid, T. Ali, M. T. Afzal, M. Ahmad, and M. A. Qadir, "Exploiting reference section to classify paper's topics," in *Proc. Int. Conf. Manage. Emergent Digit. EcoSyst.*, 2011, pp. 220–225.

[17] N. A. Sajid, M. Ahmad, M. T. Afzal, and Atta-Ur-Rahman, "Exploiting papers' reference's section for multi-label computer science research papers' classification," *J. Inf. Knowl. Manage.*, vol. 20, no. 1, Mar. 2021, Art. no. 2150004.

[18] T. C. Y. Khor and K. C. Ting, "A Bayesian approach to classify conference papers," in *Proc. Mexican Int. Conf. Artif. Intell.* Berlin, Germany: Springer, 2006, pp. 1027–1036.

[19] R. N. Karman and S. Ramaraj, "Similarity-based techniques for text document classification," *Int. J. SoftComput*, vol. 3, no. 1, pp. 58–62, 2008.

[20] M. N. Asim, M. U. G. Khan, M. I. Malik, A. Dengel, and S. Ahmed, "A robust hybrid approach for textual document classification," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1390–1396.

[21] J. Abreu, L. Fred, D. Macêdo, and C. Zanchettin, "Hierarchical attentional hybrid neural networks for document classification," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, 2019, pp. 396–402.

[22] S. A. Devi and S. Siva, "A hybrid document features extraction with clustering based classification framework on large document sets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 7, pp. 1–11, 2020.

[23] V. Balys and R. Rudzkis, "Statistical classification of scientific publications," *Informatica*, vol. 21, no. 4, pp. 471–486, Jan. 2010.

[24] S. B. Cunningham, "Applying machine learning to subject classification and subject description for information retrieval," in *Proc. 2nd New Zealand Int. Two-Stream Conf. Artif. Neural Netw. Expert Syst.*, 1995, pp. 243–246.

[25] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, Jun. 2019.

[26] T. Zhou, "Automated identification of computer science research papers," Ph.D. dissertation, Dept. Comput. Sci., Univ. Windsor, Windsor, ON, Canada, 2016.

[27] X. Luo, "Efficient English text classification using selected machine learning techniques," *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021.

[28] M. N. Asim, M. U. Ghani, M. A. Ibrahim, W. Mahmood, A. Dengel, and S. Ahmed, "Benchmarking performance of machine and deep learning-based methodologies for Urdu text document classification," *Neural Comput. Appl.*, vol. 33, pp. 5437–5469, Sep. 2020, doi: 10.1007/s00521-020-05321-8.

[29] S. Rahman and P. Chakraborty, "Bangla document classification using deep recurrent neural network with BiLSTM," in *Proc. Int. Conf. Mach. Intell. Data Sci. Appl.* Singapore: Springer, 2021, pp. 507–519.

[30] S. Jiang, J. Luo, J. Hu, and C. L. Magee, "Deep learning for technical document classification," 2021, *arXiv:2106.14269*. [Online]. Available: http://arxiv.org/abs/2106.14269

[31] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artif. Intell. Rev.*, vol. 22, no. 3, pp. 177–210, Nov. 2004.

[32] X. Wu and X. Zhu, "Mining with noise knowledge: Error-aware data mining," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 38, no. 4, pp. 917–932, Jul. 2008.

[33] J.-O. Kim and J. Curry, "The treatment of missing data in multivariate analysis," *Sociol. Methods Res.*, vol. 6, no. 2, pp. 215–240, Nov. 1977.

[34] N. K. Malhotra, "Analyzing marketing research data with incomplete information on the dependent variable," *J. Marketing Res.*, vol. 24, no. 1, pp. 74–84, Feb. 1987.

[35] R. M. Hamer and P. M. Simpson, "Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials," *Amer. J. Psychiatry*, vol. 166, no. 6, pp. 639–641, Jun. 2009.

[36] A. P. Santos and F. Rodrigues, "Multi-label hierarchical text classification using the ACM taxonomy," in *Proc. 14th Portuguese Conf. Artif. Intell. (EPIA)*, 2009, vol. 5, no. 5, pp. 553–564.

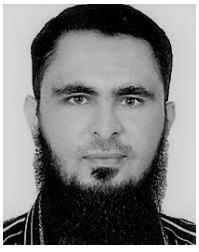
[37] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proc. Can. Conf. Electr. Comput. Eng.*, vol. 6, no. 1, 2004, pp. 22–30.
[38] P. K. Flynn, "Document classification in support of automated metadata extraction form heterogeneous collections," *Int. J. SoftComput*, vol. 2, no. 3, pp. 1–189, 2014.
[39] A. Ujjal Dey, S. Kumar Ghosh, E. Valveny, and G. Harit, "Beyond visual semantics: Exploring the role of scene text in image understanding," 2019, *arXiv:1905.10622*. [Online]. Available: http://arxiv.org/abs/1905.10622
[40] L. Xiao, G. Wang, and Y. Zuo, "Research on patent text classification based on Word2Vec and LSTM," in *Proc. 11th Int. Symp. Comput. Intell. Design (ISCID)*, vol. 1, Dec. 2018, pp. 81–84.
[41] Y. Z. C. Q. Pan, H. Dong, and L. Zhang, "Recommendation of crowdsourcing tasks based on Word2Vec semantic tags," *Wireless Commun. Mobile Comput.*, vol. 19, no. 1, pp. 61–66, 2019.
[42] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: http://arxiv.org/abs/1301.3781
[43] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[44] T. Kam Ho, "Random decision forests," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, 1995, pp. 278–282.
[45] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

GHULAM MUSTAFA received the B.S. degree (Hons.) in software engineering from COMSATS University, in 2017, and the M.S. degree (Hons.) in computer science from the Capital University of Science & Technology, Islamabad, Pakistan, in 2020. He is currently working as a Lecturer with the Computer Science Department, Capital University of Science & Technology. His research interests include data science, data mining, and text mining. He received two Gold Medals for his B.S. degree.

MUHAMMAD USMAN received the B.S. degree in computer science from Pir Mehr Ali Shah Arid Agriculture University Rawalpindi (PMAS-AAUR), in 2017, and the M.S. degree in computer science from the Capital University of Science & Technology, Islamabad, Pakistan, in 2020. He is currently working as a Lecturer with the Computer Science Department, FAST University, Islamabad. His research interests include data science, data mining, text mining, and scientometrics.

MUHAMMAD TANVIR AFZAL received the M.Sc. degree (Hons.) in computer science from Quiad-i-Azam University, Islamabad, Pakistan, and the Ph.D. degree (Hons.) in computer science from Graz University of Technology, Austria. Previously, he worked with the Capital University of Science & Technology, Islamabad, as a Professor, an Associate Professor, and an Assistant Professor of computer science. Furthermore, he has worked at NESCOM, COMSATS University Islamabad, and JinTech Islamabad. He worked as a master trainer and the program director of a national level training for a public sector organization in Pakistan on Human Factors Engineering and conducted the training of over 100 hours for experts from the industry. He has been associated with academia and industry at various levels for the last 20 years. He is currently serving as a Professor and the Director QEC of the Department of Computer Science, Namal Institute Mianwali. He conducted more than 100 curricular, co-curricular, and extra-curricular activities in the last five years, including seminars, workshops, national competitions (ExcITeCup), and invited international and national speakers from Google, Oracle, IICM, IFIS, and SEGA Europe. Under his supervision, more than 60 post graduated students (M.S. and Ph.D.) have defended their research theses successfully and a number of Ph.D. and M.S. students are pursuing their research with him. He has authored more than 120 research articles, including 50 published articles in impact factor leading journals in the field of data science, information retrieval and visualization, semantics, digital libraries, and scientometrics. He has authored two books and has edited two books in computer science. His cumulative impact factor is more than 110, with citations over 770. He played pivotal role in making collaborations between MAJU-JUCS, MAJU-IICM, and TUG-UNIMAS. He was a recipient of multiple international research fundings. He served as the Ph.D. symposium chair, the session chair, the finance chair, a committee member, and an editor for several IEEE, ACM, Springer, and Elsevier international conferences and journals. He is serving as the Editor-in-Chief for reputed impact factor journal, such as *Journal of Universal Computer Science*.

ABDUL SHAHID received the Ph.D. degree in computer science from the Capital University of Science & Technology, Islamabad, Pakistan. He is currently associated as a Faculty Member with the Institute of Computing, Kohat University of Science and Technology, Pakistan. Besides his research activities, he is a Professional Software Engineer and working as a consultant with software companies for the last 13 years. In this field, he has published a number of good quality papers in different international conferences and journals. His research interests include information systems and digital libraries. The core topic of his interest is recommending relevant documents with the help of in-text citation frequencies and patterns.



ANIS KOUBAA is currently a Professor in computer science and the Leader of the Robotics and Internet of Things Research Lab, Prince Sultan University. He is also a Research and Development Consultant at Gaitech Robotics, China, and a Senior Researcher at CISTER/INESC-TEC and ISEP-IPP, Porto, Portugal. He is also an ACM Distinguished Speaker. He is also a Senior Fellow of Higher Education Academy (HEA), U.K. He received several distinctions and awards, including the Rector Research Award at Imam Mohammad Ibn Saud Islamic University, in 2010, and the Rector Teaching Award at Prince Sultan University, in 2016. He has been the Chair of the ACM Chapter, Saudi Arabia, since 2014.