



## A Novel Metadata Based Multi-Label Document Classification Technique

Naseer Ahmed Sajid<sup>1</sup>, Munir Ahmad<sup>1</sup>, Atta-ur Rahman<sup>2,\*</sup>, Gohar Zaman<sup>3</sup>, Mohammed Salih Ahmed<sup>4</sup>,  
Nehad Ibrahim<sup>2</sup>, Mohammed Imran B. Ahmed<sup>4</sup>, Gomathi Krishnasamy<sup>6</sup>, Reem Alzahr<sup>2</sup>,  
Mariam Alkharraa<sup>2</sup>, Dania AlKhulaifi<sup>2</sup>, Maryam AlQahtani<sup>2</sup>, Asiya A. Salam<sup>6</sup>, Linah Saraireh<sup>5</sup>,  
Mohammed Gollapalli<sup>6</sup> and Rashad Ahmed<sup>7</sup>

<sup>1</sup>Barani Institute of Information Technology (BIIT), PMAS Arid Agriculture University,  
Rawalpindi, 46000, Punjab, Pakistan

<sup>2</sup>Department of Computer Science (CS), College of Computer Science and Information Technology (CCSIT), Imam  
Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia

<sup>3</sup>Faculty Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

<sup>4</sup>Department of Computer Engineering (CE), College of Computer Science and Information Technology (CCSIT), Imam  
Abdulrahman bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia

<sup>5</sup>College of Business Administration, Imam Abdulrahman bin Faisal University,  
P.O. Box 1982, Dammam, 31441, Saudi Arabia

<sup>6</sup>Department of Computer Information System (CIS), College of Computer Science and Information Technology  
(CCSIT), Imam Abdulrahman bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia

<sup>7</sup>ICS Department, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia

\*Corresponding Author: Atta-ur Rahman. Email: aaurrahman@iau.edu.sa

Received: 29 June 2022; Accepted: 13 November 2022

**Abstract:** From the beginning, the process of research and its publication is an ever-growing phenomenon and with the emergence of web technologies, its growth rate is overwhelming. On a rough estimate, more than thirty thousand research journals have been issuing around four million papers annually on average. Search engines, indexing services, and digital libraries have been searching for such publications over the web. Nevertheless, getting the most relevant articles against the user requests is yet a fantasy. It is mainly because the articles are not appropriately indexed based on the hierarchies of granular subject classification. To overcome this issue, researchers are striving to investigate new techniques for the classification of the research articles especially, when the complete article text is not available (a case of non-open access articles). The proposed study aims to investigate the multilabel classification over the available metadata in the best possible way and to assess, “to what extent metadata-based features can perform in contrast to content-based approaches.” In this regard, novel techniques for investigating multilabel classification have been proposed, developed, and evaluated on metadata such as the Title and Keywords of the articles. The proposed technique has been assessed for two diverse datasets, namely, from the Journal of universal computer science (J.UCS) and the benchmark dataset comprises of the articles published by the Association for computing machinery (ACM). The proposed technique yields encouraging results in contrast to the state-of-the-art techniques in the literature.



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Keywords:** Multilabel classification; indexing; metadata; content/data mining

## 1 Introduction

A research paper is regarded in scientific writing as a procedure for expressing scientific discoveries within a certain subject. Someone well said, “Communication in science is realized through research publications”. Since the advent of digital technologies, the number of articles published online has grown dramatically and is still growing. The growth almost doubles every five years [1]. Around thirty-thousand scholarly journals have been actively publishing nearly three million articles per year. Consequently, this published work has been searched over the web using indexing services like search engines and digital libraries. Nevertheless, it is still nearly impossible to accurately search for the relevant document in certain cases. This is because these articles are seldom indexed based on the hierarchies of subject classification such as the Computer Science ACM classification system and Elsevier [1–3]. This has created enough research gaps to be filled by researchers of closely related fields. Document classification is mainly done in two ways. First using complete article text which is certainly not possible for journals or periodicals that are not open access in nature. Second is using the metadata of the article that is available free for all types of articles, open-access, or non-open-access. Hence later approaches are more appropriate because of their independence of article access nature. Nonetheless, the proposed approach works for both types of articles since its required metadata is available in either type of article. This results in better and fine classification that is helpful for better indexing and searching over the web. Moreover, multilabel document classification is among the hottest areas of research and there is plenty of applications in text analysis and mining. For instance, a study conducted by Bathla et al. [4] proposed a recommender system for industry 5.0 that was mainly based on sentiment analysis and user behavior mining by in-depth analyzing the opinion given in the form of textual information.

This paper introduces a framework to classify research articles into the multiple categories (or classes) of the ACM classification system by employing freely available metadata such as title, keywords, and sometimes abstract. It is of particular interest to investigate the potential of metadata-based features by utilizing them in the best possible way to assist in scenarios when the content is not available. The primary observations and motivations that signify the proposed framework are listed below:

1. The content-based approaches attain more propitious results than metadata-based approaches because of their richness in features, which leads towards the biasness of the document classification community. Nonetheless, what should be done when there is no access to the research articles’ contents? The major publishing societies like ACM, Springer, Elsevier, etc. do not provide open access to many of their published articles as there are financial, legal, and technical barriers [2,4]. In such scenarios: there should be an alternative avenue to classify these articles. One of the most effective alternatives is freely available metadata. The metadata is defined as data about data or some external information about the actual data. Different kinds of useful research articles’ metadata such as titles, authors, keywords, categories, etc. are almost freely available in various digital libraries.
2. Based on the critical literature review, no scheme for the multi-label classification of scientific documents by relying fully on metadata; of course, it requires a comprehensive set of parameters. It can further be argued that an article may belong to multiple categories; the statement

is proven via experimentation on diversified datasets given in [Table 1](#). It is quite possible that a scientific document is partially related to one class and partially related to other classes. For instance, a scientific document on “Similarity Algorithm for Gene Ontology Terms” has three associations: one with the “Genes (Biology)”, second with the “Ontology”, and third with the “Similarity algorithms” class. Existing standard multi-label document classification techniques classify documents into a limited number of categories, and their performance deteriorates as the number of categories increases.

It is worth mentioning that the partial experimental results of the proposed approach have been published in [\[5\]](#). The proposed framework performs multi-label classification of scientific documents into pre-defined ACM subject hierarchy by using only metadata of the documents. The research aims to manipulate publicly accessible metadata adequately to stage multi-label classification and to estimate; “to what extent metadata-based features can perform like content-based approaches?” The *Title* and *Keywords* of research papers are exploited in different combinations in the proposed work. The system is intended to perceive the best metadata parameter for document classification.

The rest of the paper is organized as follows: Section 2 contains the research background and current state of the art. Section 3 presents the proposed scheme including dataset information and experimental setup and Section 4 presents the experimental results comprehensively by contrasting with the state-of-the-art techniques while Section 5 concludes the paper.

## 2 Research Background

The preceding section adequately emphasizes the research need and motivation that prompted the current study. This section concentrates on the comprehensive synopsis of the traditional methods in the field of document classification in general and multilabel classification in specific. As far as simple text classification is concerned, it is an old phenomenon. From the eighteenth century, investigations were conducted to verify the origin of Shakespeare’s work [\[6\]](#). When the first document classification approach was proposed, the process began to emerge into different areas, and as a result, communities began to classify specific types of documents, such as (1) magazines (2) newspapers, etc. [\[7\]](#). Due to significant discoveries in scientific bodies, the document classification community has recently redirected its attention specifically on the article’s classification. Contemporary approaches to the classification of research articles are broadly divided into two categories. (1) A method based on content (2) A method based on metadata. Subsequently, these categories are discussed step by step with the pros and cons of each category. As a top-line statement about the two categories, content-based approaches are more promising in terms of classification accuracy but are limited by open access. In contrast, the metadata-based approaches are independent of the open-access nature of the article but with reduced performance. A section is added for machine learning-based approaches as well.

### 2.1 Content-Based Approaches

Currently, the document classification community is kind of biased when it comes to data exploitation of research papers to categorize or classify them. Due to the wealth of characteristics that may be produced by utilizing the entire text, most contemporary article classification techniques rely on research article content. This section focuses on content-based state-of-the-art approaches.

The authors in [\[7\]](#) presented an approach based on a feature selection framework for Naïve Bayes. These selected features are ranked for classification. They presented a new divergence measure which is called “Jeffreys-Multi-Hypothesis (JMH) divergence, to measure multi-distribution divergence for

multi-label classification". Another study by [8] is based on the survey of feature selection approaches for text classification. They have highlighted the existing feature selection schemes and different methods of reducing the dimension of these features. Similarly, an approach in [9] classifies research papers based on different features by applying semantic similarity to them. Li et al. also proposed an automated hierarchy approach for document classification [10,11]. They have utilized the linear discriminate projection method to generate the intermediate level of the hierarchy. In [12], authors proposed a novel Bayesian automatic text classification approach by exploiting different content-based features. In 2021, authors built a document classifier by exploiting the paper's references section and named it citation-based category identification (CBCI) [13]. Another content-based classification and visualization of scientific documents are proposed by [14]. The automatic clustering approach of scientific text and newspaper articles was proposed by [15]. A content-level approach was proposed by [16] for the classification of scientific documents. They applied different algorithms like Naive Bayes, decision trees, k-Nearest neighbor (kNN), neural networks, and Support Vector Machines (SVM) for the classification of documents. In [17] authors presented an automatic subject indexing approach for digital libraries and repositories. They proposed a concept-matching approach by identifying these concepts from the documents. Another content-based hierarchical classification technique of textual data is presented by [18]. They proposed a classifier that is based on a modified version of the well-known kNN classifier. They concluded that there is a need for an effective feature selection technique with a diversified dataset for text classification [19,20]. In [21] authors proposed an approach to assign a scientific document to one or more classes which are called multi-label hierarchies by using the content of the scientific documents. They have extracted these scientific documents from the ACM library which contain scientific documents from different workshops, conferences, and Journals from the domain of Computer Science. A similar approach had been presented by [22], based on ranking category relevance to evaluate the multi-label problems. Another similar approach is proposed by [23] for the Concordia indexing and discovery (CINDI) digital library [24]. They have extracted research articles from ACM digital library belonging to the Computer Science domain. In [25], authors presented another hierarchical approach to classify text documents by using an SVM classifier. In [26] authors proposed a technique for the classification of text documents based on text similarity. The evaluation of this classification approach is also presented by [27]; they evaluated different classification approaches with their merits and demerits. For improving the performance of document classification another approach proposed Multi-co-training (MCT) and for features selection, they used the Term frequency-inverse document frequency (TF-IDF), topic distribution based on Latent Dirichlet allocation (LDA), and neural-network-based document embedding known as the document to vector (Doc2Vec) [28]. Another novel method called k-Best-Discriminative-Terms was used to collect terms from the whole text of the document instead of only collecting from the title or abstract [29]. Another approach used hierarchical and transparent representation named (HIEPAR) for Multi-label classification. They introduced it for presenting semantic information about the reviewer and the paper [30]. The content-based techniques dominate the field of document classification. Although the outcomes of these approaches are encouraging, the document's substance is a crucial prerequisite. However, most digital libraries, including Elsevier, ACM, IEEE, and Springer, are subscription-based. When the contents of are unavailable, there needs to be another way to categorize the papers. Such an alternative is available in the form of metadata like the list of authors, titles, keywords, and other structural information inside the paper such as table of contents and references. To date, there are very few classification approaches that exploit the metadata of research articles.

## **2.2 Metadata-Based Approaches**

Modern metadata-based standard-published article classification methods manipulate article metadata for classification into a predefined hierarchy. The title, authors, keywords, indexed terms, and bibliography are all part of the research article's metadata. This type of metadata is widely available to the public. Subsequently, a brief overview of the metadata-based methods has been presented. The author presented a novel approach [31] by manipulating the references portion of the article for the classification and achieved a reasonable accuracy of 70%. A natural language processing approach was proposed by Yohan et al. [32] in which they identified named entities and classified them into their different categories. Another metadata extraction scheme is proposed by Flynn [33], for document classification. The author proposed a "post hoc" classification system for document classification. Khor et al. [34], proposed a framework by using the Bayesian Network (BN). For classification, four hundred conference articles are collected and classified into four major topics. Musleh et al. proposed an approach [35]. In this approach, they developed an Arabic key-phrases extraction method for Arabic text documents. Another approach used structural content like an abstract for an automated keyword assignment system collected from a local publication database [36]. A novel similarity measure was proposed which can be used with k-NN and Rocchio algorithms [37]. Document classification outcome also depends highly on the method of information extraction [38] and the type of classification used, for example, an approach used in [39] for behavior classification. To solve the article classification problem, various schemes have been investigated in the literature based on the available data such as completed contents of the article and just the metadata or both. The content-based approaches, such as the work presented in [40–42], manipulate the entire article text for classification purposes. Several studies in this area have been conducted. Based on the dataset attributes, each approach has strengths and weaknesses. Open-access research articles are required for the implementation of such approaches. The content-based approaches exhibit higher accuracy due to their rich feature set, but only if the article is open access. In contrast, the studies on metadata-based document classification are somewhat limited. This is primarily due to the metadata's limited set of features, which are incapable of providing the same accuracy as content-based approaches but are not reliant on open-access articles. Current research aims to exploit the publicly available article metadata in a novel way to enhance the accuracy of the metadata-based approach nearer to the content-based approach and to investigate its effectiveness for multi-label classification. Overall, the classification problem is not limited to the documents, but it encompasses several other areas of research like automated text classification (ATC) for repositories, digital libraries, ontology-based information extraction [43], human behavior classification and other problems in engineering and health sciences, etc.

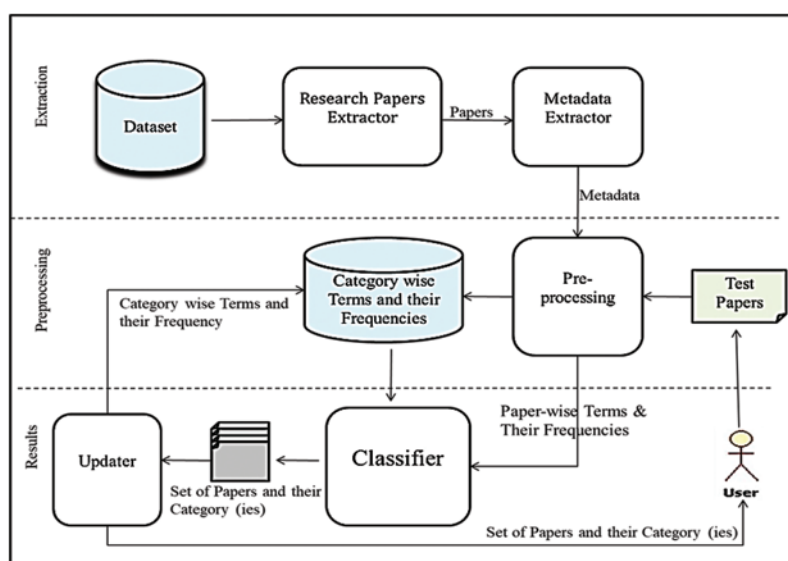
## **2.3 Machine Learning-Based Approaches**

This section discusses the most recent contributions. The first approach is MATCH's incorporation of metadata awareness and large-scale label hierarchy into their model, which resulted in a significant improvement in the multi-label document classification field, outperforming all previous studies [44]. Furthermore, this paper discussed how Ye et al. [45] took a similar approach to MATCH by incorporating label hierarchy and metadata in their model, with the addition of representing those using heterogeneous graphs, showing further improvement in their results compared to MATCH. The authors proposed a different approach from the previous two by focusing only on extracting significant parts of long documents using multi-layer attention, producing meaningful segments by a convolution-based encoder with multiple Res-SE blocks, then assigning the most relevant ones for each label [46]. Xu et al. [47] model focused on solving the lack of labeled data ready for training by

creating a dynamic self-training semi-supervised classification method, which combines self-training into the process of Multi-Label Classification. Labrak et al. [48] developed a more efficient version of the ML-Net classifier for classifying COVID-19-related documents in the LitCovid platform. Lastly, Sholikah et al. [49] contributed to the field by conducting a comparative study between various methods for implementing Multi-Label Document Classification via an automated ethical protocol review, concluding that the traditional machine learning approach and Naïve-Bayes create better results than the rest.

### 3 Proposed Multi-Label Classification Framework

The proposed system's framework [50] is depicted in Fig. 1. The research paper extractor module extracts research papers from the dataset. The system inputs these research papers into the metadata extractor module. The metadata extractor takes metadata from research articles such as titles, keywords, and categories, and passes it to the preprocessor module as input.



**Figure 1:** Multi-label computer science research papers' classification framework

#### 3.1 The Dataset

To evaluate the proposed framework, two types of datasets were chosen from [51], one consisting of J. UCS articles standard metadata and the other of ACM articles duly published by [21]. The first type of dataset was chosen because it contains hardcore computer science articles published by diverse domain researchers from around the world. Similarly, the second dataset was chosen because it contains a wide range of scientific publications such as conference proceedings, journal articles, book chapters, and many other periodicals in the computer science domain. The detailed statistics of both data sets are presented in Table 1. A detailed description of both datasets is presented subsequently explaining the number and type of features and the number of instances in each dataset. The benchmark datasets were carefully selected based on their appropriateness to the study and to make a fair comparison with other techniques.



**Table 1:** Dataset statistics

Features	ACM [21]	JUCS [51]
Research papers count	86,116	1,460
Classes (categories) at root level count	11	13
Category's level count	03	03
Papers without keyword's section count	3860	31
Single-label papers count	54%	51%
Multi-label papers count	46%	49%
Journals, conferences & workshops Count	2,240	01
References for research papers count	96,138	16,404
Averages of references in each paper	18	11
Topic reference (TR) pairs	1,59,681	47,947

These metadata parameters are parsed after some pre-processing steps and sent to the Category wise metadata merger (CMM) algorithm, which merges category-wise metadata of each research paper and stores category-wise terms and their frequencies in a database. The MLC predicts which of these research publications will fall into one of several categories. The category updater is used to improve the performance of the proposed approach by enriching the knowledge base (dataset) for article categorization. The Experiments on a variety of datasets reveal that many articles fall into many categories, as shown in Table 1. A research paper may be primarily affiliated with one category while also being partially related to other categories. In this study, a method is devised for detecting several classes of this nature. To deal with this type of overlapping, first, the study finds the membership (here the terms like frequency weights or simply weight or association for each category are used) of scientific documents concerning each category, and then the proposed study applies the alpha-cut  $\phi$  (threshold) to that membership to find the most relevant set of categories for the scientific documents. The formal representation of identifying categories for scientific documents is in Eq. (1):

$$\forall \mu_{C_i}(\mathcal{D}) \geq \phi \rightarrow \mathcal{D} : C_i \quad (1)$$

where  $\mathcal{D}$  is the scientific document,  $C_i$  is the set of categories or classes (ACM classification system),  $\phi$  is an alpha cut (threshold) assigned any value determined by the experts;  $\mu_{C_i}(\mathcal{D})$  is the membership (terms frequency weights) of  $\mathcal{D}$  in category  $C_i$  and  $\mathcal{D} : C_i$  represents the document  $\mathcal{D}$  belongs to category set  $C_i$

### 3.2 Pre-Processing

For the proposed framework, the metadata-based features are extracted from the research publications of the JUCS dataset. Per Table 1, there is a reasonable ratio of papers with multiple labels which motivates the idea of multi-label classification further. Some pre-processing operations, such as normalization (converting all terms to lowercase), removal of stop words from the Title and Keywords of all research papers, and conversion of all compound words into single words, are conducted to make metadata parameters ready for experimentation [46]. CMM Algorithm merges the metadata of all research papers that belong to the same category. The CMM algorithm is presented in Fig. 2.

The CMM algorithm extracts already processed Titles and Keywords from each research paper and merges (concatenates) them according to their categories. The Titles of all research papers

relating to each category are separately concatenated and the Term frequency (TF) weights for these Titles strings are computed. Similarly, the Keywords of all research papers relating to each category are concatenated and the TF weights for these Keyword strings are computed. Both, the Titles & Keywords are collectively concatenated for all research papers relating to each category, and the TF weights for these Titles and Keywords string are computed. These TF weights concerning Title, Keywords, and Title & Keywords strings are stored in the database concerning their categories and are computed by using the algorithm presented in Fig. 2. This algorithm works as follows. In pre-processing steps, the research papers according to their already annotated categories are stored in the database. These categories-wise research papers are given as input to this algorithm, and it returns category-wise TF weights. In Fig. 2, the algorithm picks research paper one by one from each category (in Lines 4 & 5) and extracts the Title of each research paper in a category and concatenates all Titles (in Line 9) of that category. Similarly, it extracts the Keywords of each research paper (if exists) in a category and concatenates all Keywords (in Line 8) of that category. At the end of Line 11, the resultant strings of Titles and Keywords for each category are obtained. Then these resultant strings of Title and Keywords are concatenated for each category (in Line 12) and these categories-wise resultant strings are given as inputs to the *Count\_Terms\_Frequency* module (in Line 13) to compute TF weights for each category. At the end of Line 14, all the terms and their frequencies are computed category-wise. In the same way, as in Fig. 2 algorithm, computing Term Frequency Weights of Test Document, an algorithm is proposed to compute the test document's term frequency weights. The working of this algorithm is depicted in Fig. 3.

**Input:** Research papers (documents) belong to the same category of dataset  $\mathcal{D}$ .  
**Output:** Category wise Term Frequency (TF) for each key feature (term)

```

1.  $\kappa \leftarrow$  Array of KEYWORDS of category  $c$  //  $c$  is a set of categories
2.  $\tau \leftarrow$  Array of TITLE of category  $c$ 
3.  $\mathfrak{K} \leftarrow$  Resultant array of  $\kappa$  and  $\tau$  of category  $c$ 
4. For each  $C_i$  in  $c$  //  $c$  is a set of categories
5.   For each  $d_j$  in  $\mathcal{D}$  //  $\mathcal{D}$  is a set of documents in a  $C_i$ 
6.     Locate KEYWORDS or TITLE in  $d_j$ 
7.     If found then
8.        $\kappa =$  Concatenate ( $\kappa$ , KEYWORDS in  $d_j$ )
9.        $\tau =$  Concatenate ( $\tau$ , TITLE in  $d_j$ )
10.    End if
11.  End For loop
12.   $\mathfrak{K} =$  Concatenate ( $\kappa, \tau$ ) // for each category  $C_i$ 
13.  Count_Terms_Frequency( $\mathfrak{K}$ ) // for each category  $C_i$ 
14. End For Loop
```

**Figure 2:** Category-wise metadata merger algorithm

**Input:** Set of Test Documents  $\mathcal{D}$   
**Output:** Document wise Term Frequency (TF) for each key feature (term)

```

1.  $\kappa \leftarrow$  Array of KEYWORDS of test document  $\mathcal{D}$ 
2.  $\tau \leftarrow$  Array of TITLE of test document  $\mathcal{D}$ 
3.  $\mathfrak{K} \leftarrow$  Resultant array of  $\kappa$  and  $\tau$  of test document  $\mathcal{D}$ 
4. For each  $d_i$  in  $\mathcal{D}$  //  $\mathcal{D}$  is a set of test documents &  $i=0, 1, 2, \dots$ 
5.   Locate KEYWORDS or TITLE in  $d_i$ 
6.   If found then
7.      $\kappa =$  KEYWORDS in  $d_i$ 
8.      $\tau =$  TITLE in  $d_i$ 
9.   End if
10.   $\mathfrak{K} =$  Concatenate ( $\kappa, \tau$ ) // for each test document  $d_i$  in  $\mathcal{D}$ 
11.  Count_Terms_Frequency( $\mathfrak{K}$ ) // for each test document  $d_i$  in  $\mathcal{D}$ 
12. End For loop
```

**Figure 3:** Algorithm for computing term frequency weights of the test document



When a user enters a research paper (test document) to determine its set of categories, the same pre-processing methods as described above are used. The phrase frequency weights for each term of the research article are calculated using the title and keywords of the research paper. This algorithm takes the test documents as input and returns TF against each test document. It extracts the Title and Keywords (in lines 7 & 8) from the test documents and then concatenates Title & Keywords (in line 10) to form the resultant string. Then the resultant string is given to the *Count\_Terms\_Frequency* module (in line 11) to compute TF weights for that test document. These weights return as an output to the user.

### 3.3 Multi-Label Classification (MLC)

The MLC is the core algorithm of the proposed framework [51]. It helps a user find out the most relevant category or class for the input research article. When the user gives input of a test document to identify a set of categories. For this purpose, to compute TF weights from the metadata of the test documents like Title and Keywords, the algorithm presented in Fig. 4 is used. Similarly, category-wise TF weights have already been computed by using the CMM algorithm (Fig. 2). There are three inputs for the MLC algorithm; the first is the test document's terms and their frequencies, the second input is the category-wise terms and their frequencies and third input the threshold value which is selected by the domain expert. The next task is to compare the test document's TF weights with all categories-wise TF weights. For this comparison, the test document TF weights and category-wise TF weights are given as input to the MLC algorithm, and it returns a set of categories as the output of the MLC algorithm. After the comparison of these TF weights, the weights ( $W_c$ ) for each category are calculated from Line 6 to Line 8 (the comparison is also presented above in Table 1). The total weight (*Weight\_Sum*) of all the categories is also computed (in Line 9). The next task is to compute membership ( $\mu_{vj}(\mathcal{D})$ ) (some weights or associations of the test document with each group) (in Lines 11 & 12). Finally, how to assign the most applicable category to the test document is projected (in Lines 13 & 14). Based on the test document's membership in each category, MLC predicts the most suitable category or combination of categories for it. The higher the membership, the more likely the test documents will be assigned to that group. There are two ways to forecast the most relevant category or categories for the test document at this point. One option is to choose the top category with the highest membership among all categories, and another is to use a  $\alpha$ -cut (threshold) to predict one or more categories. This threshold value is selected by the domain expert (which is discussed in the results section). Following the prediction of a category or group of categories for the test document, the next step is to update the proposed framework's knowledgebase (repository) (in Line 15). MLC does this by updating the word frequency weights for the specific category or group of categories retrieved from the test document. As a result, the knowledge base for document categorization is enriched, which improves the performance of the proposed classification approach.

## 4 Results, Discussion, and Analysis

The proposed framework is implemented and tested on two diversified datasets, one is the J.UCS dataset [51] and another is the ACM dataset developed by Santos et al. [21]. A comprehensive set of experiments has been performed on the diversified datasets and the results are evaluated by applying well-known evaluation measures for multi-label classification. These evaluation metrics are accuracy, precision, recall, and F-measure, with Godbole et al. [52] proposing formulas for these evaluation criteria for multi-label document categorization. The confusion matrix is not applicable for multi-label classification as it is used for single-label classification. That's why Godbole et al. [52] proposed formulas for evaluation criteria are applied. For the extrapolation of one or more than one category,

the proposed MLC algorithm (see Fig. 4) requires a threshold value that acts as a barrier for the prediction of categories. The MLC algorithm computes membership ( $\mu_{yj}(\mathcal{D})$ ) values, and the threshold value is chosen among them. If the computed membership value for a category is equal to or greater than the threshold value for a test paper, the category is included in the collection of predicted categories. Now the question crops up how the optimal threshold value could be selected by domain experts? There is no standard for providing the threshold value. Also, there is no such relationship between the threshold value and the number of papers. A comprehensive series of experimental rounds is performed to select the optimal threshold value. For this purpose, the evaluation process is started from a minimum threshold value, for instance, 0.05, and evaluates the performance of the proposed technique. Similarly, the threshold value has gradually increased from 0.05 to 0.4, and so on. After an in-depth analysis of results, it is analyzed that good results are achieved at a threshold value ranging from 0.1 to 0.3. But at the threshold value less than 0.1 ( $<0.1$ ), too many categories are predicted. For instance, if a research paper has one original category, the MLC algorithm predicted five categories at a 0.1 threshold value. Similarly, there are multiple cases in which predicted categories are more than the original categories at threshold values less than 0.1 ( $<0.1$ ). The prediction of many categories is reported due to a very low value of threshold value. On the other hand, when the threshold value is set to more than 3.0 ( $>3.0$ ), the numbers of predicted categories became very low or even there are many cases where the MLC algorithm predicted no category against the research paper due to the high value of the threshold. For example, if a research paper has 5 original categories and at a threshold value of 3.0, in most cases the proposed model predicted only one category. After the selection of the range of values for the threshold, the next task is to acquire a suitable threshold value that could be used by the proposed algorithm. For this, experiments have been performed on a dataset comprising 100 research papers randomly selected from J.UCS and ACM. These 100 research papers are taken from eleven different ACM categories (classes). The evaluation parameters like precision, recall, F-measure, and accuracy are computed for different data sizes of research papers like 10, 20, 30 . . . , and 100 at different threshold values like 0.10, 0.11, 0.12 . . . , and 0.30. After the detailed analysis of the first type of experiment, the most suitable threshold values are identified at which the MLC algorithm yields significant results. Using the evaluation parameters [47], each set of research papers was evaluated on a threshold range of 0.1 to 0.3. Moreover, examined the optimal threshold value for each set at which the best accuracy was attained. In Fig. 5, the x-axis represents each group of papers while optimum thresholds (those produced the best accuracy) values are depicted by the y-axis. The proposed model performed best at a threshold value of 0.2.

The MLC algorithm has predicted a set of categories for 100 research papers and evaluated them by utilizing well-known evaluation measures [53]. Each type of research article is assessed using a threshold range of 0.1 to 0.3 and the best threshold value for each group of research papers is analyzed to determine the highest level of accuracy. As a result of the first type of experiment on 100 research papers, the proposed model yielded the best results at a threshold value of 0.2. The results of all the best threshold values on which the best results are produced for the 10 sets of research papers are presented in Fig. 6. The evaluation parameters are plotted on the y-axis and the best threshold values (membership values) are plotted on the x-axis, respectively. Likewise, the best results for evaluation parameters such as accuracy, precision, and F-measure are attained at the threshold value of 0.2. Nonetheless, the best result for the recall is achieved at the threshold value of 0.15 in the first set. From the first type of experiment, it is concluded that the optimum threshold value is 0.2. For sake of

validation, this optimum threshold value is further investigated with two different and larger datasets i.e., 12 times and 15 times bigger in size, respectively. The proposed multilabel classification algorithm uses this threshold value as a reference point when predicting the categories.

**Input:** Term Frequency of Test Document ( $\mathcal{T}\mathcal{D}$ ),  
Term Frequency of each Category  $c$ ,  
 $\alpha$ -cut (threshold)

**Output:** Set of categories

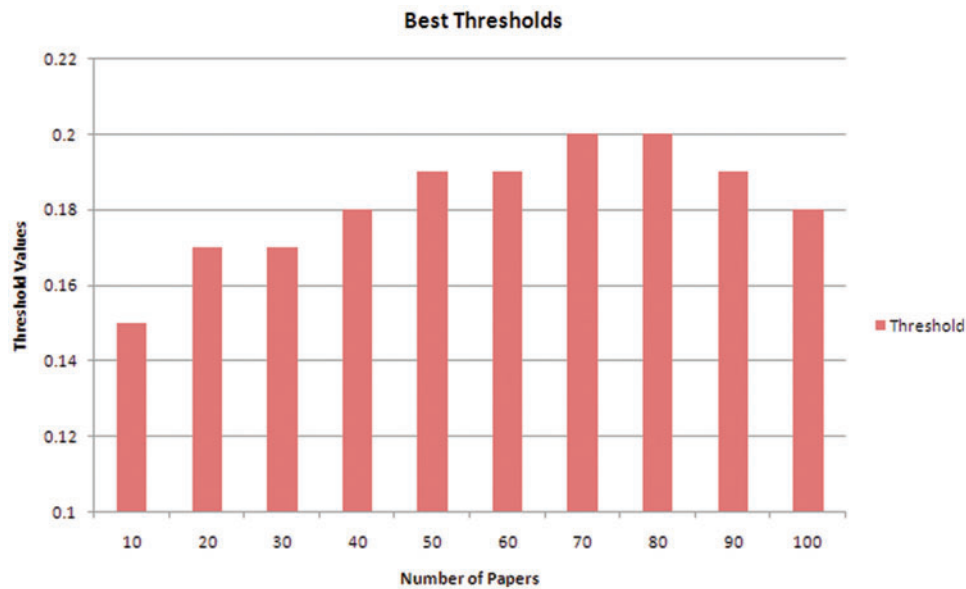
```

1.  $W \leftarrow$  Array of category's weights // to store each category weight
2.  $Weight\_Sum \leftarrow 0$ 
3. For each  $y_j$  in  $c$  // terms in each category.
4.   For each term  $x_i$  in  $\mathcal{T}\mathcal{D}$  // terms in test document
5.   If  $x_i$  is found in  $y_j$  then // matching  $\mathcal{T}\mathcal{D}$  term with  $c$  terms
6.     //Calculate the term weight of test document in each category.
       
$$W_c = \sum_{k=0}^n [TermWeight(x_i) * TermWeight(y_j)]$$

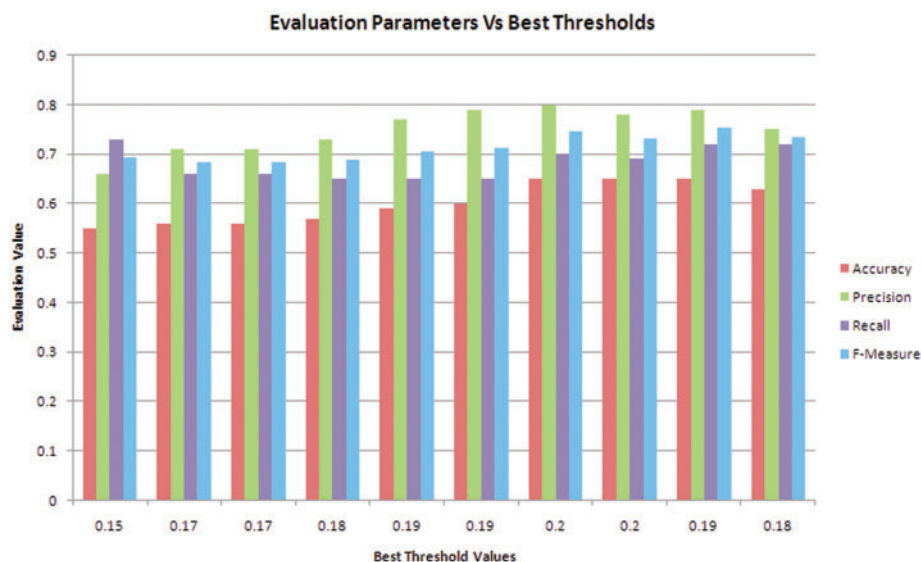
7.   End if
8. End For loop
9.  $Weight\_Sum = \sum_{b=1}^n W_c$  // total terms weights of all category  $c$  against test document
10. End For Loop
11. For each weight  $W_c$  in  $W$ 
12.    $\mu_{y_j}(\mathcal{D}) = (W_c / Weight\_Sum)$  //membership of test document in each category
13.   If  $\mu_{y_j}(\mathcal{D}) \geq \phi$  then //applying  $\alpha$ -cut(threshold)
14.      $c_i \leftarrow \mathcal{D}$  // assign document to that category
15.      $Update\_Category(c_i)$ 
16.   End if
17. End For loop

```

**Figure 4:** Multi-label classification (MLC) algorithm



**Figure 5:** Number of papers vs. threshold values



**Figure 6:** Evaluation parameters vs. best threshold values

#### 4.1 Evaluation of Threshold on J.UCS Dataset

In comparison to the first type of experiment on 100 research publications, the findings of the second type of test on a nearly 12 times bigger J.UCS dataset are critically assessed. This dataset contains a total of 3,039 pairs of paper categories, which are then merged according to their classifications. Experiments on the five threshold values (0.1, 0.15, 0.2, 0.25, and 0.3) are evaluated on the J.UCS dataset in three ways: *Title*, *Keywords*, and *Title & Keywords* of research papers. To see how well the study performs and which metadata and threshold value yields better evaluation parameters results for multi-label document classification. Table 2 shows the results of the J.UCS dataset. It has been described above that the experiments are performed on five different threshold values and the result is presented in Table 2. Standard evaluation parameters were used for the performance evaluation of these predicted results. The experimental results are analyzed in detail by applying different threshold values to the MLC algorithm. It yields different results at different threshold values. As indicated earlier, in testing results figures for the J.UCS dataset, MLC algorithm performance is notably good at a threshold value of 0.25. At this threshold value, the MLC algorithm's performance is good for both metadata (*Title*) and metadata (*Title & Keywords*). The results of the MLC algorithm are good in the range from 0.2 to 0.25 threshold values. When the threshold value is less than 0.2, it predicts more categories, lowering the accuracy and precision values. On the other side, it has the potential to boost recall values. If the threshold value is more than 0.25, the number of projected categories is too low, causing all evaluation parameter values to decrease.

**Table 2:** J.UCS dataset results

Metadata	Threshold	Accuracy	Precision	Recall	F-measure
Title	0.1	0.42	0.44	0.93	0.60
	0.15	0.59	0.72	0.69	0.70
	0.2	0.64	0.77	0.74	0.75

(Continued)

**Table 2:** Continued

Metadata	Threshold	Accuracy	Precision	Recall	F-measure
	0.25	0.85	0.95	0.88	0.91
	0.3	0.36	0.46	0.37	0.41
Keywords	0.1	0.43	0.44	0.92	0.60
	0.15	0.59	0.73	0.68	0.70
	0.2	0.62	0.76	0.71	0.73
	0.25	0.62	0.74	0.65	0.69
	0.3	0.37	0.49	0.38	0.43
Title & keywords	0.1	0.39	0.41	0.91	0.57
	0.15	0.56	0.71	0.65	0.68
	0.2	0.60	0.75	0.69	0.72
	0.25	0.65	0.76	0.66	0.71
	0.3	0.34	0.45	0.35	0.39

Hence for this dataset, the most optimum threshold value is 0.25 which produces the most significant results for multi-label document classification. It has further validated the optimal threshold value (0.2) on a 15-time bigger dataset (ACM dataset) as compared to the first experiment on 100 papers. A detailed description of the ACM dataset and the evaluation of the optimum threshold value are presented.

#### 4.2 Evaluation of Threshold on ACM Dataset

There are 86,116 research publications in the ACM collection [21] from various workshops, conferences, and journals. At the top level of the ACM Computing Classification System (CCS), it includes 11 separate categories. The results of these experiments for the ACM dataset are presented in Table 3. These experimental results are presented in three ways to analyze the effectiveness of the proposed work and to analyze that which metadata and threshold value yield better results for multi-label document classification. The tests are carried out on the ACM dataset, and the experimental outcomes are evaluated using the multi-label document classification evaluation parameters [54]. By using multiple threshold levels in the proposed MLC algorithm, the experimental data were critically assessed. At different threshold values, it produces varied outcomes. As seen in the above-mentioned experimental results figures for the ACM dataset, the proposed MLC algorithm performs very well at a threshold value of 0.2. The performance of the MLC algorithm is good for all metadata (*Title*), metadata (*Keywords*), and metadata (*Title & Keywords*) at this threshold value. The results of the MLC algorithm are good when the threshold value ranges from 0.2 to 0.25. If a threshold value less than 0.2 was picked, it predicts more categories which may decrease in accuracy as well as precision values. On the other hand, it may increase the recall value with an increase in the threshold value from 0.25 then the number of predicted categories is very small which results in decreasing all evaluation parameter values. It has been noticed from the above experiments that the proposed MLC algorithm produced better results at a threshold value of 0.2 for all metadata (*Title*), metadata (*Keywords*), and metadata (*Title & Keywords*). So, three different types of experiments are performed one on a small dataset (100 research papers) and two experiments on large datasets, one is 12 times (J.UCS dataset) and the other is 15 times (ACM) bigger than the dataset on which threshold was tuned. The above experimental results show that for the first and third types of experiments the most optimum threshold value is the

same which is 0.2. But for the second type of experiment, the results at threshold value 0.25 remained the best; however, the results are close enough to threshold value 0.2. As a result of extensive testing on various data sizes, it has been shown that at a threshold value of 0.2, the proposed model produces substantial results for a variety of datasets simply using metadata from research papers. Similarly, the proposed model yields significant numbers of results for metadata (*Title*) at threshold value 0.2 as compared to the other metadata such as metadata (*Keywords*) and (*Title & Keywords*).

**Table 3:** ACM dataset results

Metadata	Threshold	Accuracy	Precision	Recall	F-measure
Title	0.1	0.55	0.56	0.98	0.71
	0.15	0.68	0.70	0.95	0.81
	0.2	0.88	0.91	0.94	0.92
	0.25	0.71	0.79	0.74	0.76
	0.3	0.49	0.56	0.50	0.53
Keywords	0.1	0.53	0.53	0.98	0.69
	0.15	0.59	0.60	0.88	0.71
	0.2	0.73	0.73	0.73	0.73
	0.25	0.64	0.66	0.64	0.65
	0.3	0.49	0.50	0.49	0.49
Title & keywords	0.1	0.51	0.51	0.98	0.67
	0.15	0.64	0.65	0.95	0.77
	0.2	0.79	0.86	0.86	0.86
	0.25	0.72	0.75	0.72	0.73
	0.3	0.53	0.55	0.53	0.54

### 4.3 Comparison

This study presents a novel method for performing multi-label categorization of research publications in the Computer Science sector that is based entirely on openly available metadata-based criteria. The aim of current research is to evaluate and compare proposed mechanisms to existing procedures to determine their status in the state-of-the-art. The proposed approach has made the best use of freely available metadata. The approach was tested using two different datasets: (1) J.UCS [51] and (2) ACM [21]. Their thorough descriptions and outcomes have already been discussed. The comparison and critical evaluation of the obtained results have been presented. There should be some parallels between the proposed and existing approaches to support the comparisons and in this regard, the literature was thoroughly reviewed. Based on the critical literature review, no technique exists that performs multi-label categorization using solely publicly available metadata. Only single-label classification is used by those who use metadata from research publications. Because content-based approaches are so popular in the document categorization field, these schemes are abundant in the literature. Almost most of these content-based classification methods used single labels. Only one method for multi-label classification that relies on content-based features was found. The main goal of this work is to see how much metadata can be used to categorize research papers into many classifications. To compare the proposed technique to state-of-the-art, multi-label categorization using research paper content-based schemes were carefully selected. Since the proposed scheme used publicly available metadata



to accomplish multi-label classification. A detailed description of these state-of-the-art approaches is presented in Table 4. Three approaches utilized metadata of the documents and performed single-label classification. Similarly, three approaches used the content of the documents and performed multi-label classification. There are some other approaches (as described in the Background) that used the content of the documents and performed single-label classification. These approaches are not included in Table 4 because the focus of this dissertation is on two essences; one is metadata of the documents and another one is the multi-label classification. In Table 4, an approach proposed by Flynn [33] used miscellaneous documents. There are only two approaches that have utilized scientific documents for the classification. One approach was proposed by Khor et al. [34] which used metadata of the scientific documents but performed single-label classification. Another approach proposed by Santos et al. [21] used the content of the scientific documents and performed multi-label classification as well. This was the most relevant to the proposed work, hence, it was selected for comparison and evaluation.

**Table 4:** State-of-the-art approaches

Ref.	Data	Class type	# of classes	Dataset details	Algorithm/ methodology	Evaluation	Results
[21]	Content	Multiclass	11	5,000 and 10,000 documents	Binary relevance (BR), Label powerset (LP), Multi-label kNN (MLkNN)	Accuracy	Accuracy (0.88)
[22]	Content	Multiclass	WIPO (8), News (5), OHSUMED (15), ENZYME (236)	Synthetic data, WIPO-alpha, newsgroup (1000), OHSUMED (54, 708), ENZYME (9,455)	Hierarchical SVM, Hierarchical perception	Accuracy, precision	Accuracy (0.94), precision (0.89)
[23]	Content	Multiclass	6	45,000 features	Naïve Bayes, centroid	Accuracy	Accuracy (0.61)
[30]	Metadata	Single class	11	30,000 features	Genetic programming (GP)	Accuracy	Accuracy (0.61)
[33]	Metadata	Single class	99	2000 documents	Independent document model (IDM) framework	Precision, recall, F-measure	Precision (0.79), recall (0.81), F-measure (0.79)
[34]	Metadata	Single class	4	400 documents	Bayesian network (BN), Naïve Bayes (NB), Bayesian network learner (BNL)	Accuracy	Accuracy (BN, 0.84; NB, 0.83, BNL, 0.76)

First, the proposed approaches are compared with the content-based approach proposed by Santos et al. [21]. Secondly, the proposed approaches are compared with the metadata-based single-label classification scheme proposed by Khor et al. [34]. Multi-label document classification approach is proposed by Santos et al. [21] which utilizes the content of scientific documents. In this approach, scientific documents are extracted from the ACM digital library and multi-label classification is performed by using the content of these scientific documents. This system has achieved an accuracy of 0.88 by using the Binary Relevance (BR) classifier. The proposed study is interested to scrutinize to

what extent a remarkable accuracy can be achieved by using just metadata-based features instead of content-based features. For the experiments, the metadata parameters such as *Title* and *Keywords* are extracted from the research papers to classify into one or more categories. Santos et al. [21] offered an identical ACM dataset for the studies, and it was discovered that the ideal threshold value is (0.2) at which the proposed approach yields meaningful results. The experiments for the metadata *Title* and *Keywords* based strategy are carried out using the ACM dataset provided by Santos et al. [21], and the most suitable threshold value (0.1) at which the suggested approach delivers significant results is identified. Fig. 7 shows a comparison of the outcomes of these multi-labels recommended ways at the optimal threshold value with the multi-label strategy proposed by Santos et al. [21]. The proposed approach achieved the same accuracy as Santos by exploiting the metadata (*Title*). Moreover, remarkable results are found by using metadata (*Keywords*) and metadata (*Title & Keywords*). Results for metadata (*Title*) are better than metadata (*Keywords*) and metadata (*Title & Keywords*) because many research papers have no metadata (*Keywords*) section.

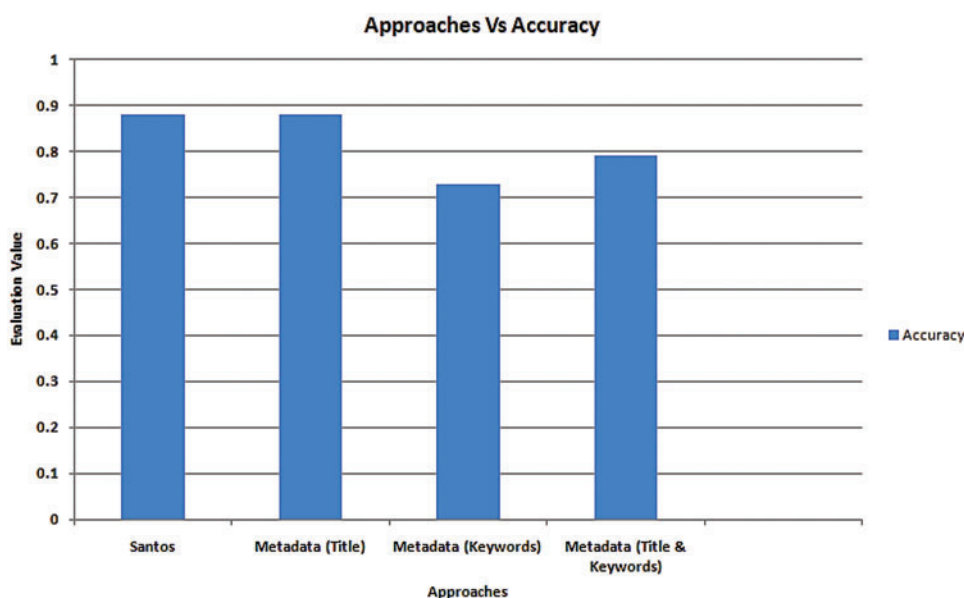


Figure 7: Comparison with a state-of-the-art approach [21]

## 5 Conclusions and Future Scope

This study uses a proposed complete process to classify research articles into several categories, which is a novel approach. Various experiments for multi-label document classification using only metadata such as *Title*, *Keywords*, and *Title & Keywords* of research articles are carried out and assessed on two different datasets: ACM and J.UCS. Further, different data sizes are investigated to find out the optimum threshold values and metadata for the proposed model to predict one or more categories. The experimental findings showed that the suggested model produces acceptable results for 100 research papers and a 15-fold bigger dataset when the threshold value is set at 0.2. (ACM). The best findings for a 12-time larger dataset (J.UCS) with a low error rate are also significantly closer to the results of this threshold setting. Consequently, it was found that the optimum threshold value of 0.2 and metadata (*Title*) at which the proposed model produces the best results when simply using metadata from research papers and performing multi-label document classification. These results are compared

to state-of-the-art techniques for multi-label document categorization at a threshold value of 0.2. It is concluded that by using only metadata of the research papers remarkable results can be obtained for multi-label classification. Moreover, the proposed approach achieved the same accuracy as the Santos et al. [21] scheme by exploiting the metadata (*Title*), and promising results are obtained by exploiting metadata (*Keywords*) as well as metadata (*Title & Keywords*). Results for metadata (*Title*) are better than metadata (*Keywords*) and metadata (*Title & Keywords*) because many research papers have no metadata (*Keywords*) section. In the future, more metadata elements can be investigated for classification such as references.

**Acknowledgement:** Acknowledged the group work.

**Funding Statement:** The authors received no specific funding for this study.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

- [1] T. Li, S. Zhu and M. Ogihara, "Hierarchical document classification using automatically generated hierarchy," *J. Intell. Inf. Syst.*, vol. 29, no. 1, pp. 211–230, 2007.
- [2] G. Zaman, H. Mahdin, K. Hussain, A. Rahman, J. Abawajy *et al.*, "An ontological framework for information extraction from diverse scientific sources," *IEEE Access*, vol. 9, no. 9, pp. 1–21, 2021.
- [3] H. Morrison, "From the field: Elsevier as an open access publisher," *The Charleston Advisor*, vol. 18, no. 3, pp. 53–59, 2017.
- [4] G. Bathla, P. Singh, S. Kumar, M. Verma, D. Garg *et al.*, "Recop: Fine-grained opinions and sentiments-based recommender system for industry 5.0," *Soft Computing*, 2021. <https://doi.org/10.1007/s00500-021-06590-8>.
- [5] N. A. Sajid, M. T. Afzal and M. A. Qadir, "Multi-label classification of computer science documents using fuzzy logic," *Journal of the National Science Foundation of Sri Lanka*, vol. 44, no. 2, pp. 155–165, 2016.
- [6] N. Zechner, "The past, present, and future of text classification," in *Proc. European Intelligence and Security Informatics Conf.*, Uppsala, Sweden, pp. 230–230, 2013.
- [7] B. Tang, S. Kay and H. He, "Toward optimal feature selection in naïve Bayes for text categorization," *IEEE Transactions on Knowledge & Data Engineering*, vol. 28, no. 9, pp. 2508–2521, 2016.
- [8] S. Shedbale, K. Shaw and P. K. Mallick, "Filter feature selection approaches for automated text categorization," *International Journal of Control Theory and Applications*, vol. 10, no. 8, pp. 763–773, 2016.
- [9] W. Zong, F. Wu, L. K. Chu and D. Sculli, "A discriminative and semantic feature selection method for text categorization," *International Journal of Production Economics*, vol. 165, no. 1, pp. 215–222, 2015.
- [10] S. Hingmire, S. Chougule, G. K. Palshikar and S. Chakraborti, "Document classification by topic labeling," in *Proc. 36th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Dublin, Ireland, pp. 877–880, 2013.
- [11] H. Park, Y. Vyas, and K. Shah, "Efficient classification of long documents using transformers," in *Proc. 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, pp. 702–709, 2022.
- [12] 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, 2016.
- [13] N. A. Sajid, M. Ahmad, M. T. Afzal and A. Rahman, "Exploiting papers' reference's section for multi-label computer science research papers' classification," *Journal of Information & Knowledge Management*, vol. 20, no. 2, pp. 2150004–2150021, 2021.

- [14] T. Giannakopoulos, E. Stamatogiannakis, I. Foufloulas, H. Dimitropoulos, N. Manola *et al.*, “Content visualization of scientific corpora using an extensible relational database implementation,” in *Proc. Int. Conf. on Theory and Practice of Digital Libraries*, Valletta, Malta, pp. 101–112, 2014.
- [15] A. R. Afonso and C. G. Duque, “Automated text clustering of newspaper and scientific texts in Brazilian Portuguese: Analysis and comparison of methods,” *Journal of Information Systems and Technology Management*, vol. 11, no. 2, pp. 415–436, 2014.
- [16] P. J. Dendek, A. Czeckzo, M. Fedoryszak, A. Kawa, P. Wendykier *et al.*, “Content analysis of scientific articles in apache hadoop ecosystem,” in *Proc. Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation Studies in Computational Intelligence*, Warsaw, Poland, pp. 157–172, 2014.
- [17] K. Golub, “Automated subject indexing: An overview,” *Cataloging & Classification Quarterly*, vol. 59, no. 8, pp. 702–719, 2021.
- [18] R. Duwairi and R. A. Zubaidi, “A hierarchical KNN classifier for textual data,” *The International Arab Journal of Information Technology*, vol. 8, no. 3, pp. 251–259, 2011.
- [19] J. J. G. Adeva, J. M. P. Atxa, M. U. Carrillo and E. A. Zengotitabengoa, “Automatic text classification to support systematic reviews in medicine,” *Expert Systems with Applications*, vol. 41, no. 4, pp. 1498–1508, 2014.
- [20] X. Deng, Y. Li, J. Weng and J. Zhang, “Feature selection for text classification: A review,” *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3797–3816, 2019.
- [21] A. P. Santos and F. Rodrigues, “Multi-label hierarchical text classification using the ACM taxonomy,” in *Proc. 14th Portuguese Conf. on Artificial Intelligence*, Aveiro, Portugal, pp. 553–564, 2009.
- [22] C. Lijuan, “Multi-label classification over category taxonomies,” Ph.D. dissertation, Department of Computer Science, Brown University, 2008.
- [23] T. Wang and B. C. Desai, “Document classification with ACM subject hierarchy,” in *Proc. Canadian Conf. on Electrical and Computer Engineering*, Vancouver, BC, Canada, pp. 792–795, 2007.
- [24] N. R. Shayan, “Cindi: Concordia indexing and discovery system,” Doctoral dissertation, Concordia University, 1997.
- [25] J. Risch, S. Garda and R. Krestel, “Hierarchical document classification as a sequence generation task,” in *Proc. ACM/IEEE Joint Conf. on Digital Libraries*, Virtual Event, China, pp. 147–155, 2020.
- [26] J. Dai and X. Liu, “Approach for text classification based on the similarity measurement between normal cloud models,” *The Scientific World Journal*, vol. 2014, no. 1, pp. 1–9, 2014.
- [27] A. Bilski, “A review of artificial intelligence algorithms in document classification,” *International Journal of Electronics and Telecommunications*, vol. 57, no. 3, pp. 263–270, 2011.
- [28] D. Kim, D. Seo, S. Cho and P. Kang, “Multi-co-training for document classification using various document representations: Tf-idf, lda and doc2vec,” *Information Sciences*, vol. 477, no. 1, pp. 15–29, 2019.
- [29] C. A. Gonçalves, E. L. Iglesias, L. Borrajo, R. Camacho, A. S. Vieira *et al.*, “Comparative study of feature selection methods for medical full-text classification,” in *Proc. IWBBIO’19*, Granada, Spain, pp. 550–560, 2019.
- [30] D. Zhang, S. Zhao, Z. Duan, J. Chen, Y. Zhang *et al.*, “A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation,” *ACM Transactions on Information Systems*, vol. 38, no. 1, pp. 1–20, 2020.
- [31] N. A. Sajid, T. Ali, M. T. Afzal, M. A. Qadir and M. Ahmed, “Exploiting reference section to classify paper’s topics,” in *Proc. MEDES’11*, San Francisco, California, USA, pp. 220–225, 2011.
- [32] P. M. Yohan, B. Sasidhar, S. A. H. Basha and A. Govardhan, “Automatic named entity identification and classification using heuristic-based approach for Telugu,” *International Journal of Computer Science Issues*, vol. 11, no. 1, pp. 173–180, 2014.

- [33] P. K. Flynn, "Document classification in support of automated metadata extraction from heterogeneous collections," Ph.D. dissertations, Faculty of Old Dominion University, 2014.
- [34] K. Khor and C. Ting, "A Bayesian approach to classify conference papers," in *Proc. 5th Mexican Int. Conf. on Artificial Intelligence*, Apizaco, Mexico, pp. 1027–1036, 2006.
- [35] D. Musleh, R. Ahmed, A. Rahman and F. Al-Haidari, "A novel approach for Arabic keyphrase extraction," *ICIC Express Letters B*, vol. 10, no. 10, pp. 875–884, 2019.
- [36] M. Kretschmann, A. Fischer and B. Elser, "Extracting keywords from publication abstracts for an automated researcher recommendation system," *Digitale Welt*, vol. 4, no. 1, pp. 20–25, 2020.
- [37] M. Eminağaoğlu and Y. Gökşen, "A new similarity measure for document classification and text mining," *KnE Social Sciences*, vol. 7, no. 1, pp. 353–366, 2020.
- [38] K. Adnan and R. Akbar, "An analytical study of information extraction from unstructured and multidimensional big data," *J. Big Data*, vol. 6, no. 91, pp. 1–38, 2019.
- [39] A. Rahman, S. Dash, A. K. Luhach, N. Chilamkurti, S. Baek *et al.*, "A neuro-fuzzy approach for user behavior classification and prediction," *Journal of Cloud Computing*, vol. 8, no. 17, pp. 1–15, 2019.
- [40] M. Rivest, E. Vignola-Gagné and É. Archambault, "Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling," *PLoS One*, vol. 16, no. 5, pp. e0251493, 2021.
- [41] Y. Lin, Z. Jun, M. Hongyan, Z. Zhongwei and F. Zhanfang, "A method of extracting the semi-structured data implication rules," *Procedia Computer Science*, vol. 131, no. 1, pp. 706–716, 2018.
- [42] D. Liangxian, Q. Junxia and G. Pengfei, "The application of semantics web in digital library knowledge management," *Physics Procedia*, vol. 24, no. 1, pp. 2180–2186, 2012.
- [43] S. A. Khan, and R. Bhatti, "Semantic web and ontology-based applications for digital libraries: An investigation from LIS professionals in Pakistan," *The Electronic Library*, vol. 36, no. 5, pp. 826–841, 2018.
- [44] Y. Zhang, Z. Shen, Y. Dong, K. Wang and J. Han, "MATCH: Metadata-aware text classification in a large hierarchy," in *Proc. WWW'21 Web Conf.* [online], pp. 3246–3257, 2021.
- [45] C. Ye, L. Zhang, Y. He, D. Zhou and J. Wu, "Beyond text: Incorporating metadata and label structure for multi-label document classification using heterogeneous graphs," in *Proc. Conf. on Empirical Methods in Natural Language Processing*, Online and at Punta Cana, Dominican Republic, pp. 3162–3171, 2021.
- [46] Y. Liu, H. Cheng, R. Klopfer, M. R. Gormley and T. Schaaf, "Effective convolutional attention network for multi-label clinical document classification," in *Proc. Conf. on Empirical Methods in Natural Language Processing*, Online and at Punta Cana, Dominican Republic, pp. 5941–5953, 2021.
- [47] Z. Xu and M. Iwaihara, "Integrating semantic-space fine-tuning and self-training for semi-supervised multi-label text classification," in *Proc. ICADL'21*, Virtual Conference, Asia Pacific region, pp. 249–263, 2021.
- [48] Y. Labrak and R. Dufour, "Multi-label document classification for COVID-19 literature using keyword-based enhancement and few-shot learning," in *Proc. BioCreative VII Challenge Evaluation Workshop*, Virtual Conference, USA, pp. 1–5, (hal-03426326), 2021.
- [49] R. Sholikah, D. Purwitasari and M. Hamidi, "A comparative study of multi-label classification for document labeling in ethical protocol review," *Techno. Com.*, vol. 21, no. 2, pp. 211–223, 2022.
- [50] N. A. Sajid, "Multi-label classification of computer science research papers using papers' metadata," Ph.D. dissertation, Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan, 2018.
- [51] A. Shahid, M. T. Afzal, M. Abdar, E. Basiri, X. Zhou *et al.*, "Insights into relevant knowledge extraction techniques: A comprehensive review," *The Journal of Supercomputing*, vol. 76, no. 9, pp. 1695–1733, 2020.
- [52] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification," in *Proc. 8th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Sydney, Australia, pp. 22–30, 2004.

- [53] S. Abbas, S. A. Raza, M. A. Khan, A. Rahman, K. Sultan *et al.*, “Automated file labeling for heterogeneous files organization using machine learning,” *Computers, Materials & Continua*, vol. 74, no. 2, pp. 3263–3278, 2023.
- [54] S. M. Alotaibi, A. Rahman, M. I. B. Ahmed and M. A. Khan, “Ensemble machine learning based identification of pediatric epilepsy,” *Computers, Materials & Continua*, vol. 68, no. 1, pp. 149–165, 2021.