applied sciences

MDPI

Review

Single vs. Multi-Label: The Issues, Challenges and Insights of Contemporary Classification Schemes

Naseer Ahmed Sajid [1], Atta Rahman [2,*], Munir Ahmad [1], Dhiaa Musleh [2], Mohammed Imran Basheer Ahmed [3], Reem Alassaf [2], Sghaier Chabani [4], Mohammed Salih Ahmed [3], Asiya Abdus Salam [5] and Dania AlKhulaifi [2]

[1] Barani Institute of Information Technology (BIIT), Pir Mehr Ali Shah (PMAS) Arid Agriculture University, Rawalpindi 46000, Pakistan
[2] Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[3] Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[4] Department of Networks and Communication, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
[5] Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia
* Correspondence: aaurrahman@iau.edu.sa

Abstract: Over the decades, a tremendous increase has been witnessed in the production of documents available in digital form. The increased production of documents has gained so much momentum that their rate of production jumps two-fold every five years. These articles are searched over the internet via search engines, digital libraries, and citation indexes. However, the retrieval of relevant research papers for user queries is still a pipedream. This is because scientific documents are not indexed based on some subject classification hierarchies. Hence, the classification of these documents becomes a challenging task for the researchers. Classification of the documents can be two-fold: one way is to assign a single label to each document and the other is to assign multi-labels to each document based on its belonging domains. Classification of the documents can be performed by using either the available metadata or the whole content of the documents. While performing classification, there are many challenges which may belong to the dataset, feature selection technique, preprocessing methodology, and which classification model is suitable for the classification of the documents. This paper highlights the issues for single-label and multi-label classification by using either metadata or content of the documents and why metadata-based approaches are better than content-based approaches in terms of feasibility.

Keywords: classification; single label; multi-label; data mining and ML; digital libraries

Citation: Sajid, N.A.; Rahman, A.; Ahmad, M.; Musleh, D.; Basheer Ahmed, M.I.; Alassaf, R.; Chabani, S.; Ahmed, M.S.; Salam, A.A.; AlKhulaifi, D. Single vs. Multi-Label: The Issues, Challenges and Insights of Contemporary Classification Schemes. *Appl. Sci.* **2023**, *13*, 6804. https://doi.org/10.3390/app13116804

Academic Editors: Muhammad Zubair Asghar, Asad Masood and Shakeel Ahmad

Received: 26 April 2023
Revised: 30 May 2023
Accepted: 1 June 2023
Published: 3 June 2023

1. Introduction

Researchers are completely immersed in the discovery of innovative contraptions to minimize human labor. These innovative ideas are being introduced in the form of research publications which are considered a language of scientific communication as further elaborated by Bornmann et al. [1]. Over the decades, an incredible increase has been seen in the production of documents available in digital form which is nearly doubled every five years [2]. A major part of this plethora of documents comprises research publications due to the subsequent discoveries and inventions in science [3]. This continuous process of research publications has never been interrupted; on the contrary, it has been increasing rapidly and exponentially [4]. A report by Ware and Mabe [5] delineates that almost 28,100 active scholarly journals are publishing almost 2.5 million articles per year. These articles are searched over the internet via search engines such as Google, digital libraries such as IEEE Explore, and citation indexes such as Web of Science (WoS) and Scopus.

The vast number of these documents is unstructured in nature, due to which search systems are not efficient enough to retrieve the most relevant documents [6]. When the user poses a query, the search systems return a bulk of documents from which very few documents hold relevance to the query. Because of this disorganization of research publications, the problem of classifying research articles into the appropriate category has gained the attention of a lot of researchers in the document classification community. The researchers aimed to classify the research articles in such a way that guarantees maximum relevant information retrieval [7]. The availability of this huge corpus on the digital web has made it challenging for researchers to classify the publications into various categories.

In machine learning (ML), classification is regarded as a central concept that aims to classify items into two or more groups. The classification is performed on various ML problems, for instance, speech recognition [8], text categorization [9,10], etc. In scientific literature, document classification is beneficial to retrieve useful information [11]. The usual method of document classification comprises the selection of useful features from the data that could help to assign some target category. The classification can be of two forms: (1) single-label classification (i.e., classifying the items into a single class) and (2) multi-label classification (i.e., classifying the items into more than one class), since a research article can have an association with multiple categories. Therefore, multi-label classification has gained the attention of many researchers who have classified research articles into multiple categories [12,13]. Most of the multi-label classification schemes are of low accuracy and classify research articles into a limited number of categories [14–16]. The classification of research articles into multiple categories with high accuracy is a challenging task [17].

Of course, multi-label classification requires an immense effort to produce a diversified and comprehensive set of features that specifically belong to each category. This research work specifically focuses on the multi-label classification of research articles with good accuracy to overcome the existing gap. How to automatically assign an appropriate category to the document or research article? In the late 1980s, document classification was performed by manually building human-crafted rules for assigning a document to some predefined category. In the 1990s, the ML paradigm outperformed the manual system, because ML automatically assigns suitable categories via supervised learning [7].

To date, numerous approaches perform document classification by using supervised machine learning. These approaches classify documents into different categories [3,6,18,19], from which some of the approaches specifically address research articles' classification problems [3,6]. A research article holds an association with a category or categories. Being specific about the issue of "classification of research articles into a predefined category", mapping a research article into the specified category or categories can be beneficial in different scenarios (but not limited to) such as:

(1) conference/journal managements want to identify reviewers for the submitted papers.
(2) authors want to submit papers on a particular topic of conference.
(3) authors want to search relevant documents to their topics.
(4) citation indexes and digital libraries want to retrieve relevant papers for user queries.

The rest of the paper is organized as follows: Section 2 contains an overview of state-of-the-art document classification approaches; Section 3 highlights the issues and challenges with an in-depth insight into the contemporary classification techniques; Section 4 concludes the paper.

2. Overview of State-of-the-Art Approaches

This section encompasses a brief overview of state-of-the-art approaches which provides a fair idea about the current trends in the research articles' classification community as every scientific study is dependent upon the study of erudite peers in the field. The document classification community is focused on proposing innovative ideas for document classification as the number of documents in digital form is increasing. Text classification is a very old dilemma. As early as the 1800s, studies were completed on verifying the authorship of the works of Shakespeare [20]. When the first document classification ap-

proach was proposed, thereafter the process started to emerge into different branches. As a result, the community began the classification of a specific type of document, for instance: (1) magazines, (2) newspapers, [12,21–23] and hierarchical classification [24]. Since then, the document classification community has diverted its attention specifically to the research papers' classification due to subsequent inventions in scientific literature.

The contemporary approaches that address the issue of research articles' classification broadly rely on two categories: (1) a content-based approach; (2) a metadata-based approach, as described subsequently. It is also worth mentioning that the content-based approaches usually outperform the metadata-based approaches due to the rich text features, etc. However, the limitation is in terms of open-access articles only. Although many publishers and journals offer an open-access facility nowadays, a huge number of journals are still non-open-access. So, this is where the content-based schemes fail and the metadata-based approaches are used, because they only rely on the available metadata of the article such as the title, abstract, keywords, references, etc. Moreover, metadata-based approaches have proven comparable to the content-based approaches in both single label and multi-label approaches.

2.1. Content-Based Approaches

Currently, the document classification community is slightly biased when it comes to the data exploitation of research papers to categorize or classify them. Most of the contemporary approaches rely on the content of research articles due to the richness of features which can be constructed by exploiting the whole content. This section focuses on content-based state-of-the-art approaches.

In 2016, Tang et al. [25] proposed a novel Bayesian automatic text classification approach by exploiting different content-based features. They proposed a class-dependent set of features. They formulated classification rules by harnessing Baggenstoss's PDF Projection Theorem for the conversion of class-specific PDFs in low-dimensional feature into raw data space. They have also presented another approach based on a feature selection framework for Naïve Bayes [21]. These selected features are ranked for the classification. They presented a new divergence measure which is called "Jeffreys-Multi-Hypothesis (JMH) divergence, to measure multi-distribution divergence for multilabel classification". Another study by Shedbale et al. [22] is based on the survey of features' selection approaches for text classification. They highlighted the existing feature selection schemes and different methods of reducing the dimension of these features. These methods are categorized into two categories, (1) wrapper and (2) filter. The filter scheme provides significant performance improvement over the wrapper scheme without classifier feedback. The filter scheme has been used in most of the text classification/categorization problems in the literature.

Zhou et al. built a content-based classifier by using Naïve Bayes and Logistic Regression algorithms [26]. The classifier was built by relying on different features from which Bi-grams feature outperformed for both datasets. Similarly, Zong et al.'s [23] approach classified research papers based on different features by applying semantic similarity to them. Another content-based classification and visualization of scientific documents is proposed by Giannakopoulos et al. [27]. All modules of this approach were implemented by using the madIS system. The madIS system provides data evaluation functionalities via an extended relational database. The automatic clustering approach of scientific text and newspapers articles was proposed by Afonso and Duque [28]. A content level approach was proposed by Dendek et al. [13] for the classification of scientific documents. They applied different algorithms such as: Naïve Bayes, decision tree, k-nearest neighbor (KNN), neural network, Support Vector Machine (SVM) for the classification of documents. Likewise, Yaguinuma et al. [29] proposed a fuzzy ontology to represent and reason for fuzzy or vague information for more fine-grained classification by incorporating the fuzziness features.

Arash and Mahdi [30] presented an automatic subject-indexing approach for digital libraries and repositories. They proposed a concept matching approach by identifying these

concepts from the documents. Then, concept similarities are computed with the documents. Concept similarity is used for the classification of documents. Hingmire et al. [12] proposed a document classification algorithm based on the Latent Dirichlet Allocation (LDA) [31] and unlabeled dataset. The algorithm of this approach assigns one topic to one class label. A query extension method is proposed by Ortuño et al. [32], which extended information related to research papers by using their cited references. The evaluation of this approach was conducted on biomedical documents of the PubMed database. Another content-based hierarchical classification technique of textual data is presented by the authors of [33]. They proposed a classifier that was based on a modified version of the well-known k-nearest neighbor classifier (K-NN). The original classifier works only with the category representatives instead of the training documents. This category representation saved them effort and time, as they did not need to deal with all training documents and categories of different levels. They concluded that there is a need for an effective feature selection technique with the diversified dataset for the text classification [34,35].

Santos and Rodrigues [36] proposed an approach to assign a scientific document to one or more classes which is called multi-label hierarchy by using the content of the scientific documents. A similar approach was presented by Lijuan [37], based on ranking category relevance to evaluate the multi-label problems. Another similar approach is proposed by Wang and Desai [38] for the CINDI digital library. They formulated their method for ranking classes on the same level which can be helpful for text classification. The evaluation of this approach was performed on the collected dataset by using ACM98 classification scheme. They extracted research articles from the ACM digital library belonging to the computer science domain. Their method of text classification specifies and prepares the rank for the categories at the same level. Their method works from top downwards in the hierarchy until the suggested category is assigned. They used a flat local multi-label classifier which served as the basic block in their hierarchical classification system. Cai and Hofmann [39] presented another hierarchical approach to classify text documents by using an SVM classifier. They exploited the relationships among the classes which are commonly expressed in the form of hierarchy. Senthamarai and Ramaraj [40] proposed a technique for the classification of text documents based on text similarity. They presented a feature selection framework which calculates the score of selected words for text classification. They have also presented a learning model for text categorization, in which document collections were randomly selected and annotated by the domain experts. The evaluation of this classification approach is also presented by [41]; they evaluated different classification approaches with their merits and demerits [42–44].

Galke et al. [45] presented a systematic evaluation of classification approaches to explore how far semantic annotations can be conducted using just the metadata of the documents. The evaluation was completed with the classification obtained from analyzing only the metadata and with analyzing the semantic annotation of the whole text. Yan et al. [46] proposed a multi-label document-ranking model based on Long Short-Term Memory (LSTM). It consisted of two processes, one was repLSTM (an adapted representation process) and the other one was rankLSTM (a unified learning ranking process). Three datasets were used for the experiments to classify documents with reasonable performance of their proposed model. Baker and Korhonen [47] presented a method which performed hierarchical multi-label document classification by initializing a neural network model. They evaluated their approach on the biomedical domain using both sentences and document level classification. Wang et al. [48] proposed an ensemble classification method which groups together random forest and semantic core co-occurrence latent semantic vector space (CLSVSM). The Yahoo dataset was used for experiments which revealed the effectiveness of the proposed method with reasonable results. In the work by [49–53], the authors proposed automated text classification/categorization approaches. The document classification community is dominated with the content-based approaches.

Of course, these approaches have richness in terms of features and produce promising results. To make these schemes applicable to the content of the documents is a vital

requirement but most of the digital libraries are subscription-based such as ACM, IEEE, Springer, etc. [54–57]. There is a need for some alternative method to categorize documents when the content is not available. Such an alternate method is available in the form of metadata such as authors, title, keywords, etc. To date, there are very few document classifications approaches that exploit the metadata of research articles [58–60]. We discuss metadata-based document classification approaches in the next section.

2.2. Metadata-Based Approaches

The contemporary metadata-based state-of-the-art research articles' classification schemes exploit the metadata of research articles for their classification into a pre-defined hierarchy. Metadata of scientific documents include title, authors, keywords, categories, funding/acknowledgement, references section, etc. These forms of metadata are almost freely available online as compared to the whole content of the articles. This section focuses on a brief overview of the metadata-based approaches.

Flynn proposed a metadata extraction scheme [42] for document classification. This was a "post hoc" classification system for document classification. After the metadata extraction of the document, the post hoc technique applied further to classify these documents. Khor and Ting [16] proposed a framework by using the Bayesian Network (BN) for the classification of conference papers. They used the keywords of research papers for classification. A feature selection algorithm is applied to automatically extracted keywords for each topic. To improve the performance of document classification approaches into predefined categories, Zhang et al. proposed another approach [43]. In this approach, they combined citation information and structural contents such as the title and abstract of the documents. Different similarity measures based on the structural contents and citation information are evaluated to improve the effectiveness of the classification. To address the document classification problem, the researchers employed different schemes on two data sources such as metadata and content.

In 2023, Sajid et al. [50] proposed a novel metadata-based approach for the classification of computer science published articles. Two diverse datasets were investigated in this regard. First, the dataset was obtained from the Journal of Universal Computer Science (J.UCS) and the second benchmark dataset was obtained from the Association for Computing Machinery (ACM) published articles [51]. The proposed approach was able to classify the articles based on metadata only and the performance was comparable to that of the content-based approaches in the literature.

The content-based schemes exploit the content of research articles for their classification [12–16,24,27,30,36,40,43,44]. Every scheme has its own pros and cons which depend on the size, pre-processing, and nature of the dataset. For these schemes' implementation, the content of research articles is an essential requirement. The content-based schemes provide better precision due to the rich number of features [13]; however, the content of scientific documents is not freely available most of the time. On the other hand, very few researchers have used only the metadata of the documents for the classification [16,42,43]. The metadata of the documents provide a limited number of features which may result in low accuracy as compared to the content-based document classification schemes. The objective of this paper is to use freely available metadata and to analyze to what extent the metadata-based features can behave like content-based features. Moreover, to what extent is the scheme effective for the sake of multi-label classification? The metadata are freely available in most scientific digital libraries such as IEEE (http://ieeexplore.ieee.org/ (accessed on 15 January 2023)), ACM (http://dl.acm.org/citation.cfm?id=2077531 (accessed on 15 January 2023)), and Springer (http://link.springer.com/chapter/10.1007%2F11925231_98 (20 January 2023)).

2.3. Evaluation Criteria

For a comprehensive understanding of the critical findings of the literature, this section has defined evaluation criteria on which all key papers from the literature have been evaluated and are shown as a comparative study in Tables 1–3.

Table 1. Critical Analysis of Content-based Approaches for Single Label Classification.

Approach	No. of Classes	Dataset	Algorithm/Methodology	Evaluation Parameters	Results
[45]	Econ (4), Polite (5), RCV1 (14), NVT (2)	Econ (62,924), Polite (27,576), RCV1 (100,000), NVT (100,000)	KNN	F-Measure	Econ (0.41), Polite (0.27), RCV1 (0.76), NVT (0.40)
[21]	20-Newsgroups (20), Reuters (135)	20-Newsgroups (20,000), Reuters (21,578)	Naïve Bayes	Accuracy, F-Measure	Accuracy (0.95), F-Measure(0.90)
[25]	20-Newsgroups (20), Reuters (135)	20-Newsgroups (20,000), Reuters (21,578)	Bayesian	F-Measure, G-Mean	Not Reported
[22]	C, Reuters (135)	20-Newsgroups (20,000), Reuters (21,578)	Survey	Accuracy, F-Measure	Accuracy (0.95), F-Measure (0.90)
[26]	Not Reported	CiteSeerX (665,483), arXiv (84,172)	Naive Bayes, Logistic Regression	F-Measure	CiteSeerX (0.76), arXiv (0.95)
[23]	20-Newsgroups (20), Reuters-10(10)	20-Newsgroups (16,391), Reuters-10 (7224)	SVM	F-Measure	20-Newsgroups (0.76), Reuters-10 (0.91)
[29]	4	100 Documents	Fuzz-Onto	Accuracy	Accuracy (0.44)
[30]	wiki-20 (5)	Wiki-20 (20)	Concept Matching-based Approach (CMA)	Precision, Recall	Precision (0.61), Recall (0.58), F-Measure (0.60)
[12]	20-Newsgroups (8), SRAA (10), WebKB(10)	20-Newsgroups, SRAA (73,218), WebKB (4199)	Latent Dirichlet Allocation (LDA)	F-Measure	20-Newsgroups (0.92), SRAA (0.85), WebKB (0.71)
[33]	Not Reported	100 Features	KNN	Precision, Recall	Precision (0.73), Recall (0.55)
[44]	Reuters (10), WebKB (7)	Reuter (21,578), WebKB (8282)	NPE and Particle Swarm Optimization (PSO)	F-Measure	Reuter (0.94), WebKB (0.89)
[40]	Not Reported	2000 Documents	PSO	Accuracy	Accuracy (0.9)

Table 2. Critical Analysis of Content-based Approaches for Multi-labels.

Approach	No. of Classes	Dataset	Algorithm/ Methodology	Evaluation Parameters	Results
[46]	Biomedicine (150), Email (6), News (103)	Biomedicine (100,000), Email (3021), News (800,000)	Long Short Term Memory (LSTM)	F-Measure	F-Measure (0.70)
[47]	PubMed (30)	PubMed (1852)	INIT-A, INIT-B	Precision, Recall, F-Measure	Precision(0.73, 0.68), Recall (0.77, 0.83), F-Measure (0.75, 0.75)

Table 2. *Cont.*

Approach	No. of Classes	Dataset	Algorithm/ Methodology	Evaluation Parameters	Results
[36]	11	5000 and 10,000 Documents	Binary Relevance, Naïve Bayes Multi-Nominal, Multi-label kNN	Accuracy	Accuracy (0.88)
[37]	WIPO-alpha (8), Newsgroups (5), OHSUMED (15), ENZYME (236)	Synthetic data, WIPO-alpha, Newsgroups(1000), OHSUMED (54,708), ENZYME (9455)	Hierarchical SVM, Hierarchical Perception	Accuracy, Precision	Accuracy (0.94), Precision (0.89)
[38]	6	45,000 Features	Naïve Bayes, Centroid	Accuracy	Accuracy (0.61)

Table 3. Critical Analysis of Metadata-based Approaches.

Approaches	Classification Type	No. of Classes	Dataset	Algorithm/ Methodology	Evaluation Parameters	Results
[42]	Single-Class	99	2000 Documents	Independent Document Model (IDM) Framework	Precision, Recall, F-Measure	Precision (0.79), Recall (0.81), F-Measure (0.79)
[16]	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)
[43]	Single-Class	11	30,000 Features	Genetic Programming (GP)	Accuracy	Accuracy (0.61)
[50]	Multi-Class	11	J.UCS and ACM	Metadata title, metadata keywords and combined		Accuracy (0.88)
[58]	Multi-Class	11	J.UCS and ACM	Reference section	Accuracy	Accuracy (0.74)

2.3.1. Type of Data Source

The first evaluation criterion is the type of data source; researchers from the diversified domain have exploited data sources such as metadata and content of the documents. Some researchers used metadata of the documents and most of the researchers used the content of the documents.

2.3.2. Classification Type

The second criterion is the classification type. The single class means that we have many classes, but one document will be classified into only one class. Multi-label means we have many classes, and one document may be classified into one or more than one class.

2.3.3. Number of Classes

The next evaluation criterion is the number of different classes. This highlights how many classes to which a particular research paper belongs. Most of the researchers have used the standard classification scheme that is the ACM classification system, which contains eleven topics at its root.

2.3.4. Dataset

The evaluation criterion of the dataset will depict how many documents are used to evaluate the approaches from the literature. This will highlight the average number of documents we should pick for our experiments for the evaluation of the proposed approaches.

2.3.5. Algorithm/Methodology/Approach

This criterion will discuss the algorithms and methodologies used in the literature for the evaluation of the research documents. This will further help us to form an evaluation and comparison strategy.

2.3.6. Evaluation Parameters/Metrics

In the classification problems, usually, the target classes are given while the predicted/desired classes are obtained by the prescribed approaches/models and techniques. So, the evaluation parameters are used to see to what extent classification was successful. These are equally applicable to single and multi-label classification scenarios. Nonetheless, in the case of multi-label classification, more sophisticated approaches are used since the interclass boundaries/differences may blur which may result in potential misclassification.

The evaluation criterion will highlight which scientific documents have used which particular evaluation parameters, for example, accuracy, precision, recall, and F-Measure [59,60].

These are calculated by measuring true positive (TP), true negative (TN), false positive (FP) and false negative (FN) metrics, as given in Equations (1)–(4) [59,60].

- Precision: The number of true positive observations which belong to the total expected positive observations. It is represented by the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

- Recall: Represent the number of actual positive cases predicted as being positive. It is represented by the formula:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

- Accuracy: This is the percentage of correct predictions the model made. It is represented by the formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

- F1-Score: This is the weighted average of the recall and the precision. It is represented by the formula:

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precison} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

2.3.7. Results

The last criterion is the results, which will demonstrate how much value of accuracy, precision, and recall has been achieved so far in contemporary state-of-the-art approaches. It is usually measured in the form of a percentage. The higher the percentage means the approach is more accurate, precise, etc.

2.4. Critical Analysis of Contemporary Approaches Based on Evaluation Criteria

After the comprehensive analysis of the above-mentioned state-of-the-art approaches, we have concluded that these classification approaches exploited different data sources so that some of these have used metadata while others have used the content of the documents. Based on the above discussed observations, we classified state-of-the-art approaches into three types:

(1) Content-based approaches which exploited the content data source and classified documents into only one class (single label) from the multiple classes.
(2) Content-based approaches which exploit the content data source and classified documents into one or more than one class (multi-label).
(3) Metadata-based approaches which exploited the metadata source of the documents and classified documents into either single label or multi-label from the multiple classes.

2.4.1. Analysis of Content-Based Approaches (Single Label Classification)

The state-of-the-art approaches which exploited the content of the documents are shown in Table 1. Researchers of these approaches performed content-based document classification and classified documents into only one class. Different algorithms were used to predict the most relevant class. Similarly, different datasets were used for the classification of documents. For example, the datasets 20-Newsgroups and Reuters were used by many researchers [12,21–25,44]. Similarly, some other datasets have also been used for the document classification [31–33,40].

From Table 1, we can examine that the number of classes vary from dataset to dataset. Different researchers classified documents into a different number of pre-defined classes. By using these datasets and pre-defined number of classes, a variety of state-of-the-art approaches were presented in the last couple of decades and exploited content of the documents to classify documents into single label. These approaches have used different evaluation parameters such as accuracy, precision, recall, and F-measure. These approaches achieved accuracy from 0.4 to 0.95 by exploiting the content of the documents. Similarly, parameter precision achieved from 0.61 to 0.8, recall achieved from 0.55 to 0.76, and parameter F-measure achieved from 0.71 to 0.94 as mentioned in the literature. These values are significantly good because these techniques have exploited the content of documents which contain a huge bag of words (features) for the classification.

2.4.2. Analysis of Content-Based Approaches (Multi-Label Classification)

The state-of-the-art approaches which exploit the content of documents and have performed multi-label classification are shown in Table 2. These approaches have predicted one or more than one classes from the multiple classes. However, there are very few state-of-the-art approaches which perform multi-label classification. For multi-label classification, Santos [36] presented an approach which utilizes an ACM dataset and ACM classification system which contains eleven classes at its root level. They have applied different approaches (algorithms) for the multi-label classification and achieved accuracy up to 0.88. Similarly, Lijuan [37] also performed multi-label classification and applied an algorithm on different datasets such as WIPO-alpha, 20-Newsgroups, Enzyme, etc., and achieved accuracy up to 0.94 and precision up to 0.84. Wang and Desai [38] also presented a multi-label classification approach which uses 45,000 features to classify documents and classifies accurately up to 0.61. As already mentioned, multi-label classification is relatively more vulnerable to misclassification compared to the binary classification.

2.4.3. Analysis of Metadata-Based Approaches

The state-of-the-art approaches which exploit the metadata of the documents have performed single-label classification. These approaches have been shown in Table 3. These approaches predicted the most relevant class for a particular document from the multiple pre-defined classes. Very few state-of-the-art approaches have performed document classification by exploiting only metadata [16,42,43]. One important finding from the literature is that the systems which utilize the metadata of research papers were only able to classify papers into a single class. For single-label classification, Flyn [42] applied an algorithm on two thousand documents (2000) for the classification of documents into 99 predefined classes and achieved precision up to 0.79, recall value 0.81, and F-measure value 0.79, respectively.

Khor [16] applied different algorithms on a collection of 400 documents but they used very few generic classes (i.e., four classes) and achieved accuracy up to 0.84 for their document classification technique. Zhang [43] also used the genetic algorithm (GA) which is from the family of evolutionary algorithms to perform metadata-based document classification. They applied their techniques to a collection of 30,000 features to classify documents into eleven pre-defined classes.

The authors in [50,58] investigated their approaches on J.UCS and ACM standard datasets. In [50], the authors investigated the multi-label classification using metadata title,

metadata keywords, and metadata title plus keywords and achieved the highest accuracy of 0.88. Similarly, the authors in [58] exploited the references section as the metadata of the articles to classify them into multiple classes and achieved the highest accuracy of 0.74.

The main advantage of the metadata-based approaches over the content-based approaches is minimum dependency on the content availability. However, in contrast to the content-based approaches, metadata-based approaches exhibit relatively degraded performance and that is quite understandable.

2.5. Deep Learning-Based Approaches

The state-of-the-art deep learning methods to perform document classification are shown in Table 4. These approaches either predicted one or more classes. Zhao et al. [61] designed a framework that captures the hierarchal relationships in semi-structured documents to extract the multilevel semantics. The results show the advantage of using pretrained word embedding and deep learning compared to classical machine learning techniques. However, their work is limited to the classification of semi structured documents containing a clear hierarchically semantic structure. The results of [62] confirm the performance advantage of using pretrained word embedding such as GloVe-300 and Word2Vec-100. However, Zhao et al. [61,62] suggested in both references exploring other options; potentially more powerful alternatives include Bidirectional Encoder Representations from Transformers (BERT) [63], RoBERTam, and GPT3. Compared to GloVe, it is expected to improve performance in the cost of longer processing and training time. Limsopatham [64] used BERT in their work, and their findings in legal document classification show that pretraining models by using in-domain documents improved the performance. However, if in-domain documents are limited, then pretraining using a large corpus also leads to an improved performance. Behera and Kumaravelan [65] proposed an FRS-RNN+CNN model and compared the results with classical machine learning models. Their model achieved an accuracy of 98.5% in the Reuters dataset and 96.98% accuracy in the 20-Newsgroups dataset. Compared to classical machine learning techniques, the results were significantly higher. However, the tuning of hyperparameters takes longer. Almuzaini [66], studied the impact of stemming on Arabic datasets; the results of their work show that in Arabic NLP applications, stemming is not an essential step to improve performance. However, by stemming the vocabulary is significantly reduced and therefore the training time is also reduced. Kim [67] performed a multi-label classification on a Korean translated dataset and achieved accuracy up to 71%. Huang et al. [68] used four datasets, Amazon Mobile Phone reviews, Amazon fine food reviews, and Yelp reviews 1 and 2. The authors observed that the accuracy of their method was higher in the Amazon mobile phone reviews and the fine food reviews compared to Yelp. A possible explanation is the more professional words typically used in the first two. It is commonly observed that precision, recall, and F-measure [69,70] are among the most widely used criteria to evaluate the performance of deep learning approaches as shown in the last rows of Table 4.

Table 4. Deep Learning-Based Approaches.

Approach	No. of Classes	Dataset	Algorithm	Evaluation Parameters	Results
[61]	MEDLINE (29)	MEDLINE (143,842)	DL BASE MODEL	Precision, Recall, F-measure	P (51.68) R (54.29) F1 (52.95)
	OHSUMED (23)	OHSUMED (13,929)			P (65.72) R (69.38) F1 (67.50)

Table 4. *Cont.*

Approach	No. of Classes	Dataset	Algorithm	Evaluation Parameters	Results
[66]	Arabic News Texts (8)	Arabic News Texts (6114)	CNN, CNN-LSTM CNN-GRU BiSTM BiGRU Att-LSTM Att-Gru	Weighted average F-measure	81.68, 80.5, 80.11, 81.686, 83.63, 81.9, 83.22
	Saudi Press Agency (6)	Saudi Press Agency (13,402)			96.71, 97.01, 96.78, 97.44, 97.37, 97.38, 97.96
[67]	Annals of the Joseon Dynasty (40)	Annals of the Joseon Dynasty (380,009)	HAN	Accuracy, Hamming Loss, Micro F1, Macro F1	Accuracy (71%), Hamming Loss (0.044), Micro F1 (0.83), Macro F1 (0.75)
[64]	ECHR Violation (40)	ECHR Violation (11,000)	(MaxPool-ECHR-Legal-BERT)	Micro F1	micro F1 (0.7213),
			BigBird		micro F1 (0.7308)
	Overruling Task Dataset (2)	Overruling Task Dataset (2400)	Harvard-Law-BERT	F-measure	F1 (0.9756)
[62]	20-Newsgroups (20)	20-Newsgroups (18,846)	CNN-BiFaGRU	Accuracy, Precision, Recall, F-measure	A (73.5), P (75.87), R (72.21), F1 (73.95)
	AG-News (4)	AG-News (127,600)			A (88.05), P (88.41), R (87.79), F1 (89.09)
	R8 (8) of Reuters	R8 (7674) of Reuters			A (96.8), P (97.02), R (96.67), F1 (96.84)
	R52 (52) of Reuters	R52 (9100) of Reuters			A (92.64), P (94.28), R (91.82), F1 (93.01)
	WebKb (4)	WebKb (4199)			A (90.47), P (90.92), R (89.97), F1 (90.44)
[65]	20 Newsgroups (20)	20-Newsgroups (18,846)	FRS-RNN +CNN	Accuracy, Precision, Recall, F-measure	A (96.98), P (97.09), R (96.98), F1 (97)
	Reuters-21578 (8)	Reuters-21578 (10,788)			A (98.5), P (98.56), R (98.5), F1 (98.5)
[68]	Yelp1 (5)	Yelp1 (1,990,636)	HMAN, HMAN-no DVA	Accuracy	73.4
	Yelp2 (5)	Yelp2 (1,894,817)			73.4
	Food Reviews (5)	Food Reviews (110,000)			with DVA (82.6) without DVA (82.8)
	Phone Reviews (5)	Phone Reviews (110,000)			with DVA (83.5) without DVA (82.3)

3. The Issues and Challenges and Insight of Contemporary Classification Approaches

The following are the issues or challenges which were observed from the above discussion in Section 2.

The existing research articles' classification schemes depend upon the content of the articles. In this context, most of the time, the non-availability of research articles (complete research paper) makes those schemes non-applicable. There is a need for some best alternative way to classify research articles that produce results closer or better than content-based approaches. Most state-of-the-art approaches focus on single-label classification, while research articles may belong to multiple categories. There is a need for such a multi-label classification system that utilizes the best possible alternate of the content-based approaches with closer or improved accuracy. The existing multi-label classification schemes classify citations into a limited number of categories. While a research article may belong to multiple categories, for instance, in the computer science domain, the research articles may belong to more than one category of ACM classification system. The ACM categorization system has 11 topics on its root level. There is a need for an approach that is efficient enough to classify research articles at least to the root level of the ACM classification system.

There are also other challenges which were observed and can make sophisticated problems for the researcher who is working in the domain of the text classification by using either the content or metadata of the research papers. Feature selection is an important part of text classification, feature vectors must have some meaning which represents the text, and these features must be free of noise. Due to the presence of outliers and unknown classes, the classification of the text may become more subjective. According to the range of text, features from the text may vary from hundreds of thousands to thousands of thousands of features. An accurate feature selection can significantly yield a great contract of mileage in the text classification process. The content of the research papers may yield a large amount of data; stemming may turn down the performance of the classifiers.

There are very few state-of-the-art approaches that rely on freely available metadata as shown in Table 3. These schemes classify documents into single label [16,42,43]. Only the approaches proposed by Santos and Rodrigous [36], Lijuan [37], and Wang and Desai [38] classify documents to multiple classes but by exploiting the content of the documents. All other approaches have not dealt with the multi-label classification problem. The existing multi-label classification schemes classify documents into a limited number of categories. The researchers who used the metadata of the papers only performed the single-label classification and achieved up to almost 0.84 accuracy by using a few numbers of classes.

Deep learning-based approaches have become more popular in terms of better performance. For instance, according to Table 4, these schemes exhibit an average accuracy of 90% and above [62,64–66] with multi-label classification. However, again, the dependency is mainly on the availability of the contents. Nonetheless, when it comes to the limited metadata of the research articles, the accuracy and other measures are seriously degraded, as seen in the case of [61].

Document classification is more challenging than that of other text classifications such as tweets, reviews, news articles, etc. That is mainly because of the following:

1. Research articles are subject to open-access issues. Non-open-access journal articles do not provide the content of the paper but only the metadata.
2. Research articles are mainly available in PDF format. Parsing the PDF document and converting it into text, especially when the document is not structured and/or requires optical character recognition (OCR), mainly results in textual errors that eventually may affect the classification accuracy of the employed model.
3. Other textual documents are not restricted in terms of open-access and structuring; hence, NLP approaches are more successful in terms of accuracy and other measures.

4. Conclusions

This research paper highlights the single-label and multi-label classification and the need for multi-label classification for documents. The document classification community is dominated with the content-based approaches. Of course, these approaches have richness in features and produce promising results. To make these schemes applicable the content of the documents is a vital requirement but most of the digital libraries are subscription-based, such as ACM, IEEE, and Springer, etc. There is a need for some alternative method to categorize documents when the content is not available. Such an alternative is available in the form of metadata such as authors, title, keywords, etc. There are very few document classifications approaches that exploit the metadata of research articles and perform single-label classification. However, exploiting metadata for the multi-label classification is also a challenging task for the researchers, by using freely available metadata in the best possible way to perform multi-label classification and to evaluate to what extent metadata-based features can perform in the same way as content-based approaches. Furthermore, it can be gleaned from the extensive literature review that there is still a significant gap in the investigation of deep learning models for the metadata-based approaches in a multi-label documents' classification paradigm. In this regard, other variants of deep learning must be investigated such as transfer learning, ensemble approaches, and the fused models, in particular, when more than one pre-trained model is used to classify the document based on a joint consensus.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bornmann, L.; Mutz, R. Growth rates of Modern Science: A Bibliometric Analysis based on the Number of publications and Cited References. *J. Assoc. Inf. Sci. Technol.* **2015**, *66*, 2215–2222. [CrossRef]
2. Larsen, P.O.; Ins, M. The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index. *Sci. Metr.* **2010**, *84*, 575–603. [CrossRef] [PubMed]
3. Davis, J.; Weeks, R.; Revett, M. Jasper: Communicating Information Agents for WWW. In Proceedings of the Fourth International World Web Conference, Boston, MA, USA, 11–14 December 1995; pp. 11–14.
4. Hodgson, A.; Schlager, L. Closing the PDF Gap: ReadCube's Experiments in Reader Focused Design. *Learn. Publ.* **2017**, *30*, 65–69. [CrossRef]
5. Ware, M.; Mabe, M. *The STM Report: An Overview of Scientific and Scholarly Journal Publishing*; International Association of Scientific, Technical and Medical Publisher: The Hague, The Netherlands, 2015.
6. Koller, D.; Sahami, M. Hierarchically Classifying Documents using very few Words. In Proceedings of the 14th International Conference on Machine Learning (ICML-97), Nashville, TN, USA, 8–12 July 1997; pp. 170–178.
7. Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* **2002**, *34*, 1–47. [CrossRef]
8. Jelinek, F. *Statistical Methods for Speech Recognition*; The MIT Press: Cambridge, MA, USA, 1998.
9. Apte, C.; Damerau, F.; Weiss, S.M. Automated Learning of Decision Rules for Text Categorization. *Inf. Syst.* **1994**, *12*, 233–251. [CrossRef]
10. Dagan, I.; Karov, Y.; Roth, D. Mistake-driven Learning in Text Categorization. In Proceedings of the EMNLP-97, The Second Conference on Empirical Methods in Natural Language Processing, Providence, RI, USA, 1–2 August 1997; pp. 55–63.
11. Shin, K.; Abraham, A.; Han, S. Enhanced Centroid-Based Classification Technique by Filtering Outliers. In *Text, Speech and Dialogue*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 4188, pp. 159–163.
12. Hingmire, S.; Chougule, S.; Palshikar, G.K.; Chakraborti, S. Document Classification by Topic Labeling. In Proceedings of the SIGIR '13—36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 28 July–1 August 2013; pp. 877–880.
13. Dendek, P.J.; Czeczko, A.; Fedoryszak, M.; Kawa, A.; Wendykier, P.; Bolikowski, L. Content Analysis of Scientific Articles in Apache Hadoop Ecosystem. In *Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation Studies in Computational Intelligence*; Springer: Cham, Switzerland, 2014; pp. 157–172.
14. Salton, G. Developments in Automatic Text Retrieval. *Science* **1990**, *253*, 974–980. [CrossRef]

15. Gerstl, P.; Hertweck, M.; Kuhn, B. Text Mining: Grundlagen, Verfahren und Anwendungen. *HMD-Prax. Wirtsch.* **2001**, *38*, 38–48.
16. Khor, K.; Ting, C. A Bayesian Approach to Classify Conference Papers. In Proceedings of the 5th Mexican International Conference on Artificial Intelligence, Apizaco, Mexico, 13–17 November 2006; pp. 1027–1036.
17. Har-Peled, S.; Roth, D.; Zimak, D. Constraint Classification for Multiclass Classification and Ranking. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2002; pp. 809–816.
18. Kononenko, I. Comparison of Inductive and Naïve Bayesian Learning Approaches to Automatic Knowledge Acquisition. In *Current Trends in Knowledge Acquisition*; IOS Press: Amsterdam, The Netherlands, 1990.
19. Sajid, N.A.; Ali, T.; Afzal, M.T.; Qadir, M.A.; Ahmed, M. Exploiting Reference Section to Classify Paper's Topics. In Proceedings of the International Conference on Management of Emergent Digital EcoSystems (MEDES'2011), San Francisco, CA, USA, 21–23 November 2011; pp. 220–225.
20. Zechner, N. The Past, Present and Future of Text Classification. In Proceedings of the Intelligence and Security Informatics Conference (EISIC), Uppsala, Sweden, 12–14 August 2013.
21. Tang, B.; Kay, S.; He, H. Toward Optimal Feature Selection in Naive Bayes for Text Categorization. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 2508–2521. [CrossRef]
22. Shedbale, S.; Shaw, K.; Mallick, P.K. Filter Feature Selection Approaches for Automated Text Categorization. *Int. J. Control Theory Appl.* **2016**, *10*, 763–773.
23. Zong, W.; Wu, F.; Chu, L.K.; Sculli, D. A Discriminative and Semantic Feature Selection Method for Text Categorization. *Int. J. Prod. Econ.* **2015**, *165*, 215–222. [CrossRef]
24. Li, T.; Zhu, S.; Ogihara, M. Hierarchical Document Classification Using Automatically Generated Hierarchy. *J. Intell. Inf. Syst.* **2007**, *29*, 211–230. [CrossRef]
25. Tang, B.; He, H.; Baggenstoss, P.M.; Kay, S. A Bayesian Classification Approach using Class-specific Features for Text Categorization. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1602–1606. [CrossRef]
26. Zhou, T. Automated Identification of Computer Science Research Papers. Ph.D. Thesis, University of Windsor, Windsor, ON, Canada, 2016.
27. Giannakopoulos, T.; Stamatogiannakis, E.; Foufoulas, I.; Dimitropoulos, H.; Manola, N.; Ioannidis, Y. Content Visualization of Scientific Corpora using an Extensible Relational Database Implementation. In *Theory and Practice of Digital Libraries*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 101–112.
28. Afonso, A.R.; Duque, C.G. Automated Text Clustering of Newspaper and Scientific Texts in Brazilian Portuguese: Analysis and Comparison of Methods. *J. Inf. Syst. Technol. Manag.* **2014**, *11*, 415–436. [CrossRef]
29. Yaguinuma, C.A.; Santos, M.T.P.; Camargo, H.A.; Nicoletti, M.C.; Nogueira, T.M. A Meta-Ontology for Modeling Fuzzy Ontologies and its Use in Classification Tasks based on Fuzzy Rules. *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.* **2014**, *6*, 89–101.
30. Arash, J.; Mahdi, A.H.E. Classification of Scientific Publications According to Library Controlled Vocabularies: A new concept matching-based Approach. *Libr. Hi Tech* **2013**, *31*, 725–747.
31. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
32. Ortuño, F.M.; Rojas, I.; Navarro, M.A.A.; Fontaine, J.F. Using Cited References to Improve the Retrieval of Related Biomedical Documents. *BMC Bioinform.* **2013**, *14*, 113. [CrossRef]
33. Duwairi, R.; Al-Zubaidi, R. A Hierarchical K-NN Classifier for Textual Data. *Int. Arab. J. Inf. Technol.* **2011**, *8*, 251–259.
34. Eyheramendy, S.; Madigan, D. A Novel Feature Selection Score for Text Categorization. In Proceedings of the Workshop on Feature Selection for Data Mining, in Conjunction with the SIAM International Conference on Data Mining, Newport Beach, CA, USA, 21–23 April 2005; pp. 1–8.
35. Tang, B.; Shepherd, M.; Milios, E.; Heywood, M. Comparing and Combining Dimension Reduction Techniques for Efficient Text Clustering. In Proceedings of the Workshop on Feature Selection for Data Mining, in Conjunction with the SIAM International Conference on Data Mining, Newport Beach, CA, USA, 21–23 April 2005; pp. 17–26.
36. Santos, A.P.; Rodrigues, F. Multi-label Hierarchical Text Classification using the ACM Taxonomy. In Proceeding of 14th Portuguese Conference on Artificial Intelligence, Aveiro, Portugal, 12–15 October 2009; pp. 553–564.
37. Lijuan, C. Multi-Label Classification over Category Taxonomies. Ph.D. Thesis, Department of Computer Science, Brown University, Providence, RI, USA, 2008.
38. Wang, T.; Desai, B.C. Document Classification with ACM Subject Hierarchy. In Proceedings of the 2007 Canadian Conference on Electrical and Computer Engineering, Vancouver, BC, Canada, 22–26 April 2007; pp. 792–795.
39. Cai, L.; Hofmann, T. Hierarchical Document Categorization with Support Vector Machines. In Proceedings of the CIKM '04—Thirteenth ACM International Conference on Information and Knowledge Management, Washington, DC, USA, 8–13 November 2004; pp. 78–87.
40. Senthamarai, K.; Ramaraj, N. Similarity based Technique for Text Document Classification. *Int. J. Soft Comput.* **2008**, *3*, 58–62.
41. Brucher, H.; Knolmayer, G.; Mittermayer, M. Document Classification Methods for Organizing Explicit Knowledge. In Proceedings of the Third European Conference on Organizational Knowledge, Learning, and Capabilities, Athens, Greece, 5–6 April 2002.
42. Flynn, P.K. Document Classification in Support of Automated Metadata Extraction from Heterogeneous Collections. Ph.D. Thesis, Faculty of Old Dominion University, Norfolk, VA, USA, 2014.

43. Zhang, B.; Goncalves, M.; Fan, W.; Chen, Y.; Fox, E.; Calado, P.; Cristo, M. Combining Structural and Citation-Based Evidence for Text Classification. In Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04), ACM, New York, NY, USA, 8–13 November 2004; pp. 162–163.
44. Wang, Z.; Sun, X. Document Classification Algorithm Based on NPE and PSO. In Proceedings of the 2009 International Conference on E-Business and Information System Security, EBISS'09, Wuhan, China, 23–24 May 2009.
45. Galke, L.; Mai, F.; Schelten, A.; Brunsch, D.; Scherp, A. Using titles vs. full-text as source for automated semantic document annotation. In Proceedings of the Knowledge Capture Conference, ACM, Austin, TX, USA, 4–6 December 2017.
46. Yan, Y.; Wang, Y.; Gao, W.C.; Zhang, B.W.; Yang, C.; Yin, X.C. Lstm2: Multi-label ranking for document classification. *Neural Process. Lett.* **2018**, *47*, 117–138. [CrossRef]
47. Baker, S.; Korhonen, A. *Initializing Neural Networks for Hierarchical Multi-Label Text Classification*; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; Volume BioNLP 2017, pp. 307–315.
48. Wang, R.; Chen, G.; Sui, X. Multi-label text classification method based on co-occurrence latent semantic vector space. *Procedia Comput. Sci.* **2018**, *131*, 756–764. [CrossRef]
49. Musleh, D.; Ahmed, R.; Rahman, A.; Al-Haidari, F. A Novel Approach to Arabic Keyphrase Extraction. *ICIC Express Lett. B* **2019**, *10*, 875–884.
50. Sajid, N.A.; Ahmad, M.; Rahman, A.-U.; Zaman, G.; Ahmed, M.S.; Ibrahim, N.; Krishnasamy, G.; Alzaher, R.; Alkharraa, M.; AlKhulaifi, D.; et al. A novel metadata based multi-label document classification technique. *Comput. Syst. Sci. Eng.* **2023**, *46*, 2195–2214. [CrossRef]
51. Shahid, A.; Afzal, M.T.; Abdar, M.; Basiri, M.E.; Zhou, X.; Yen, N.Y.; Chang, J.-W. Insights into relevant knowledge extraction techniques: A comprehensive review. *J. Supercomput.* **2020**, *76*, 1695–1733. [CrossRef]
52. Rahman, A. Knowledge Representation: A Semantic Network Approach. In *Handbook of Research on Computational Intelligence Applications in Bioinformatics*, 1st ed.; IGI Global: Hershey, PA, USA, 2016; Chapter 4.
53. Rahman, A.; Dash, S.; Luhach, A.K.; Chilamkurti, N.; Baek, S.; Nam, Y. A Neuro-Fuzzy Approach for User Behavior Classification and Prediction. *J. Cloud Comput.* **2019**, *8*, 17. [CrossRef]
54. Rahman, A.; Alhaidari, F.A. The Digital Library and the Archiving System for Educational Institutes. *Pak. J. Inf. Manag. Libr. (PJIML)* **2019**, *20*, 94–117. [CrossRef]
55. Zaman, G.; Mahdin, H.; Hussain, K.; Rahman, A. Information Extraction from Semi and Unstructured Data Sources: A Systematic Literature Review. *ICIC Express Lett.* **2020**, *14*, 593–603.
56. Alamoudi, A.; Alomari, A.; Alwarthan, S.; Rahman, A. A Rule-Based Information Extraction Approach for Extracting Metadata from PDF Books. *ICIC Express Lett. Part B Appl.* **2021**, *12*, 121–132.
57. Zaman, G.; Mahdin, H.; Hussain, K.; Rahman, A.; Abawajy, J.; Mostafa, S.A. An Ontological Framework for Information Extraction from Diverse Scientific Sources. *IEEE Access* **2021**, *9*, 42111–42124. [CrossRef]
58. Sajid, N.A.; Ahmad, M.; Afzal, M.T.; Rahman, A. Exploiting Papers' Reference's Section for Multi-Label Computer Science Research Papers' Classification. *J. Inf. Knowl. Manag.* **2021**, *20*, 2150004. [CrossRef]
59. Alghamdi, A.S.; Rahman, A. Data Mining Approach to Predict Success of Secondary School Students: A Saudi Arabian Case Study. *Educ. Sci.* **2023**, *13*, 293. [CrossRef]
60. Alqarni, A.; Rahman, A. Arabic Tweets-Based Sentiment Analysis to Investigate the Impact of COVID-19 in KSA: A Deep Learning Approach. *Big Data Cogn. Comput.* **2023**, *7*, 16. [CrossRef]
61. Zhao, W.; Fang, D.; Zhang, J.; Zhao, Y.; Xu, X.; Jiang, X.; Hu, X.; He, T. An effective framework for semistructured document classification via hierarchical attention model. *Int. J. Intell. Syst.* **2021**, *36*, 5161–5183. [CrossRef]
62. Belherazem, A.; Tlemsani, R. Boosting Convolutional Neural Networks Using a Bidirectional Fast Gated Recurrent Unit for Text Categorization. *Int. J. Artif. Intell. Mach. Learn.* **2022**, *12*, 1–20. [CrossRef]
63. Alotaibi, A.; Rahman, A.; Alhaza, R.; Alkhalifa, W.; Alhajjaj, N.; Alharthi, A.; Abushoumi, D.; Alqahtani, M.; Alkhulaifi, D. Spam and sentiment detection in Arabic tweets using MARBERT model. *Math. Model. Eng. Probl.* **2022**, *9*, 1574–1582. [CrossRef]
64. Limsopatham, N. Effectively Leveraging BERT for Legal Document Classification. In *Proceedings of the Natural Legal Language Processing Workshop*; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 210–216.
65. Behera, B.; Kumaravelan, G. Text document classification using fuzzy rough set based on robust nearest neighbor (FRS-RNN). *Soft Comput.* **2021**, *25*, 9915–9923. [CrossRef]
66. Almuzaini, H.A.; Azmi, A.M. Impact of Stemming and Word Embedding on Deep Learning-Based Arabic Text Categorization. *IEEE Access* **2020**, *8*, 127913–127928. [CrossRef]
67. Kim, D.K.; Lee, B.; Kim, D.; Jeong, H. Multi-Label Classification of Historical Documents by Using Hierarchical Attention Networks. *J. Korean Phys. Soc.* **2020**, *76*, 368–377. [CrossRef]
68. Huang, Y.; Chen, J.; Zheng, S.; Xue, Y.; Hu, X. Hierarchical multi-attention networks for document classification. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1639–1647. [CrossRef]

69. Gollapalli, M.; Rahman, A.; Alkharraa, M.; Saraireh, L.; AlKhulaifi, D.; Salam, A.A.; Krishnasamy, G.; Alam Khan, M.A.; Farooqui, M.; Mahmud, M.; et al. SUNFIT: A Machine Learning-Based Sustainable University Field Training Framework for Higher Education. *Sustainability* **2023**, *15*, 8057. [CrossRef]
70. Rahman, A.; Musleh, D.; Nabil, M.; Alubaidan, H.; Gollapalli, M.; Krishnasamy, G.; Almoqbil, D.; Khan, M.A.A.; Farooqui, M.; Ahmed, M.I.B.; et al. Assessment of information extraction techniques, models and systems. *Math. Model. Eng. Probl.* **2022**, *9*, 683–696. [CrossRef]