Iraqi Journal for Computers and Informatics
Information Technology and Communications University
Vol. 50, No. 1, 2024, pp. 34-45

DOI: https://doi.org/10.25195/ijci.v50i1.475
Print ISSN: 2313-190X, Online ISSN: 2520-4912

Research Article

Identifying Researchers’ Interest using Text Mining

Zahraa Ali Katia Hassoun[1,*]
College of Business Informatics
University of Information Technology
and Communications
Baghdad, Iraq
ms202110671@iips.edu.iq

Safaa O. Al-mamory[2]
College of Information Technology,
University of Babylon
Babylon-Iraq
saffa@itnet.uobabylon.edu.iq

ARTICLE INFO
Article History
Received: 20/02/2023
Accepted: 02/03/2024
Published: 01/06/2024

ABSTRACT
Researchers' interests and academic journals are crucial for advancing scientific inquiry. Journals serve as platforms for sharing and validating discoveries, fostering a symbiotic relationship that advances our collective understanding and pushes the boundaries of human knowledge. Journals, which encompass natural edge research and establish benchmarks for academic rigor. In this paper, an analysis, using text mining, of the publications of Iraqi researchers in scientific journals is used to extract the researcher's interest. In more detail, this paper utilizes the following technologies: pre-processing (tokenization, POS (“Part Of Speech”), normalization, case folding, lemmatization) – filtering (stop word elimination) - feature Extraction (TF-IDF), as well as classification using deep neural network classifier (DNNC), to address the problem of identifying the researcher's interests through texts (title &abstract) analysis. The Iraqi researchers’ data in the field of computer science from the years 2010-2022. As obtained from the Scopus repository, a total of 1170 papers were collected via API- key and scrubber depending on the keyword of computer science and the year. Furthermore, these papers were manually classified based on the hierarchical classification of the ACM journal. Finally, the best results obtained from a classification using DNN and TF-IDF as classifying terms achieved a precision of 90%, Recall of 90%, f1-score of 90%, and accuracy of 90%.

Keywords:* *Researchers interest’s identification, DNN, classification, TF-IDF, Preprocessing.

1. INTRODUCTION

The paper introduces a structured approach, starting with the development of the MTAIMRAD framework (metadata, title, abstract, introduction, methodology, results, analysis, discussion), serving as the CAMTA (“Content Analysis of Metadata, Title, and Abstract “) method. It then offers practical guidelines, including an illustrative example, for implementing CAMTA. The subsequent critical examination explores the method's potential applications and benefits in business and management research, leveraging insights from content analysis applications in various fields [1, 2]**.** The CAMTA method, as introduced in this text, emphasizes the key elements of a journal article, namely metadata, title, and abstract. Emphasizing their accessibility, the author cites and highlights their significance as primary access points in research papers. Although existing content analysis applications have been predominantly in healthcare, the paper proposes extending CAMTA to the field of business and management due to its adaptable nature [3 , 6]**.**

Point out misconceptions in content analysis for literature reviews in business and management and also an earlier assertion that content analysis is underutilized. The current uses in business and management mostly revolve around Citation Context Analysis or Bibliometric Analysis, sidelining titles, abstracts, and other metadata components [7, 8].

The paper's motivation is to address the identified gap by presenting the CAMTA method for business and management research, aiming for rigor and validity. CAMTA's flexibility allows it to be used independently or in combination with other methods, responding to the need for methodological pragmatism and transparency in business and management research[2, 9, 10].

Scopus is a source-neutral abstract and citation database curated by independent subject matter experts who are recognized leaders in their fields, as well as it puts powerful discovery and analytics tools in the hands of researchers, librarians, research managers, and funders to promote ideas, people, and institutions**.** Scopus is designed to serve the information needs of researchers, educators, administrators, students, and librarians across the entire academic community. Whether searching for specific information or browsing topics, authors, journals, or books, Scopus provides precise entry points to peer-reviewed literature in the fields of science, technology, medicine, social sciences, and the arts and humanities [11].

This study aims to build a dataset that is a standard that can later be relied upon to solve a problem, this dataset was obtained from the Scopus repository which consists of 1170 papers with 11 columns (According to the information retrieved), Afterward, manual classification will be applied, utilizing the ACM (Association for Computing Machinery) digital library tree classification. The data will then undergo preprocessing, involving techniques like tokenization, stemming, filtering, and feature extraction as a series of processes. Subsequently, the existing data will be transformed into numerical data known as vector features before being inputted into the construction of the DNN model. As well as another study aims to identify the researcher's interest based on the effectiveness in classifying papers and choosing the special keyword that is relevant. The study outcomes will yield a model. designed to categorize papers based on the ACM digital library tree classification**.**

Finally, the rest of this paper is structured as follows: section 2 contains the literature review, while section 3 provides details on the proposed approach. section 4 explores the results and discussion. The conclusion and recommendation are presented in section 5.

2. RELATED WORK

Researcher's interests: -

Marie *et al.*,(2016) [12]examine Highlights the importance of promoting cross-disciplinary research. Represent researchers’ interests in a large-scale academic database that includes not a specific, but a diverse range of research fields in Japan. As well as shows the applicability of the topic representation of researchers to author disambiguation and proposes a technique for identifying and summarizing researchers' interests across diverse fields by using LDA (Latent Dirichlet Allocation) to compute topic distributions across words based on the training dataset. Next, employing LDA to turn each of the target researcher's articles into a topic vector; last, determining the topic vectors' centroid to provide a profile of the researcher. Experiments carried out on the CiNii Articles reveal that the themes collected by this method cover a wide variety of academic fields. Furthermore, it is shown that the topic representation provided can be applied to author disambiguation.

Felipe et al.,(2021) [13] This study investigates the use of machine learning approaches to identify the primary research areas of specialization of researchers by utilizing a variety of numerical representations of their scientific production titles as the algorithms' data source. Further to address determining a researcher's specialty by utilizing a TF-IDF character n-gram representation of the text in the titles. The topic of determining a researcher's area of expertise has been approached from several angles, with the goal of determining whether or not employing a certain approach—like character-level n-grams with TF-IDF—would yield better results than the current state-of-the-art. Discovered that word-level representation is inferior to character-level n-grams with TF-IDF. The dataset used is (the titles of the researchers’ scientific productions and the Wikipedia pages (Lattes platform)), and the techniques used to process this problem are Text mining and information extraction. So, the result was Recognizing the research focus of a particular scholar solely based on the titles of their published works.

Nanna *et al.*,(2022) [14] study investigates explores and understands the creative work practices of academic researchers, specifically focusing on the workflows and tools they use to manage research ideas by using Thematic analysis. Overall, the findings suggest a diversity of approaches to idea management among researchers and emphasize the importance of supporting flexibility in tool usage and integration. The study also proposes avenues for future research to delve deeper into specific aspects of idea management in the research domain.

Jiangbo *et al.*,(2023) [15] This study description of a research work that focuses on information retrieval from social media platforms, feature selection using optimization algorithms, and the generation of association rules for applications such as recommendation systems by using those techniques DL(Deep Learning), NLP(Natural Language Processing) data mining, and feature selection, finally the result from this work was detecting the interests of social networking users.

Qiang Yang *et al.*,(2023) [8] This study introduces and elucidates a novel method for identifying changes in academic research interests, overcoming drawbacks in current approaches, and highlighting the benefits of the proposed model with experimental findings LSTM(Long Short Term Memory), GNN(graph neural network), and DNN(deep neural

DOI: https://doi.org/10.25195/ijci.v50i1.475
Print ISSN: 2313-190X, Online ISSN: 2520-4912

network) algorithms. So the main result is the development of a model that effectively addresses the problem of detection of interpretable shifts in researcher interests using temporal heterogeneous graphs.

Finally, we note that previous studies differ from this work in the dataset was collected from the Scopus repository by API and used, and no off-the-shelf data was used. Also, the data was manually classified based on the classification tree of the ACM scientific journal Computer Science. In addition, suitable preprocessing techniques have been applied to the text, which are (tokenization, POS, lemmatization, and stop-word elimination) to ensure accurate results. As well as Text vectorization by applying the TF-IDF method was used to obtain important words' characteristics(features). Text analysis was conducted on both the title and abstract, providing more detailed understanding and accuracy toward achieving the desired goal.

3. DATASET COLLECTION

Selecting raw data from reliable sources is crucial to developing an innovative machine-learning model. It supports the model's effectiveness, ensures data accuracy, reduces biases, and promotes compliance with a range of regulations and standards in the field. Scopus, known for its comprehensive coverage of scientific literature, is a go-to resource for researchers looking for high-quality data. To access this repository, navigate to the Elsevier Developer Portal and create an API key after registration. This key is essential for validating data retrieval requests. This key is essential for validating data retrieval requests. The need to include the API key within request headers highlights the careful attention to detail required for secure and effective data access, as shown in Fig. 1. [16].

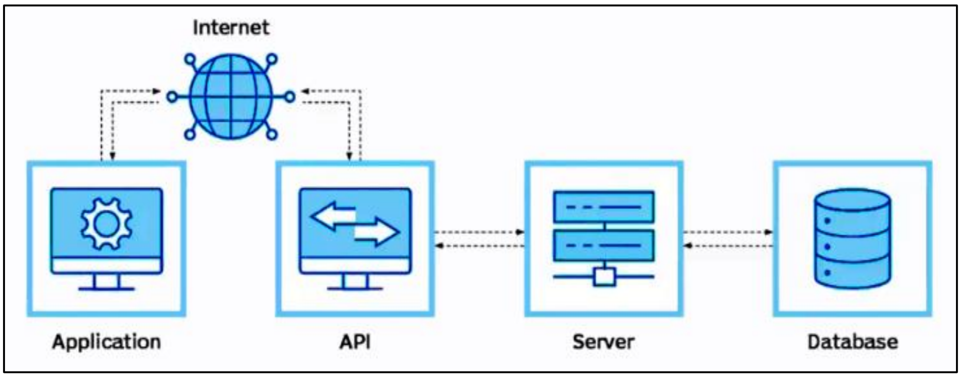

Fig. 1. Dataset collection framework.

To obtain comprehensive paper abstracts was challenging, mainly due to restrictive site policies that limit access to detailed researcher data. This issue, discussed in Chapter 4, required a strategic shift towards a scrubbing process, allowing the collection of 1,170 relevant papers from Iraqi computer science researchers from 2010 to 2022. This story details the technical steps involved and illustrates data science research's iterative and problem-solving nature, emphasizing the importance of ethical considerations and adaptability in pursuing knowledge and innovation. The papers were manually classified based on the ACM journal scientific website (digital library) within the computer science specialty and for ten specializations only, as shown in Table 1.

TABLE I: THE MANUAL CLASSIFICATION' CLASSES BASED ON ACM

No.	Class
1	Apply Computing
2	Information System
3	Computing Methodology
4	Mathematics of Computing
5	Social and Professional Topics
6	Theory of Computing
7	Network
8	Hardware
9	Human-Centered Computing
10	Software and its Engineering

DOI: https://doi.org/10.25195/ijci.v50i1.475
Print ISSN: 2313-190X, Online ISSN: 2520-4912

4. THE PROPOSED SYSTEM

This section provides a concise overview of both the proposed method and the gathered dataset. The approach encompasses five primary steps: dataset collection (from Scopus database using API, python package, scrapping), text pre-processing (labeling based on ACM, tokenization, POS, normalization, case folding, lemmatization, stop word eliminate), feature extraction (using TF-IDF), text classification (using DNN classifier), result evaluation (using precision, Recall, f1score, accuracy). Then, the suggested method for classification papers is based on the ACM digital library tree classification. Figure 2 illustrates the method proposed and employed in this study.

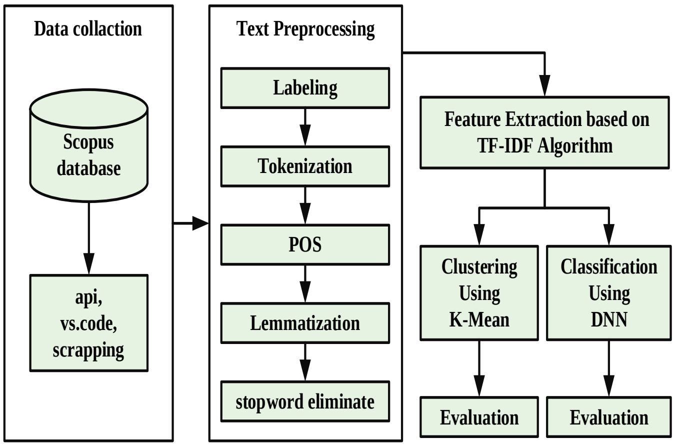

Fig. 2. The structure of the proposed model.

4.1. Text Preprocessing

Data preprocessing plays a pivotal role in the machine learning pipeline. It tackles diverse data challenges and readies the data for analysis, contributing significantly to the effectiveness of the entire process. this stage consists of several steps to clean data and make it suitable for the next stage of the proposed model.

4.2. Label

This step is very important because the quality of labeled data dramatically affects the performance of machine learning models. Therefore, attention to detail, consistency, and validation processes are essential to producing high-quality labeled datasets. Error! Reference source not found. shows the manual labeling process. For this reason, the data was classified based on the scientific journal ACM, so Only ten specializations were relied upon due to the suitability of the classification model, as shown in Figure 3.

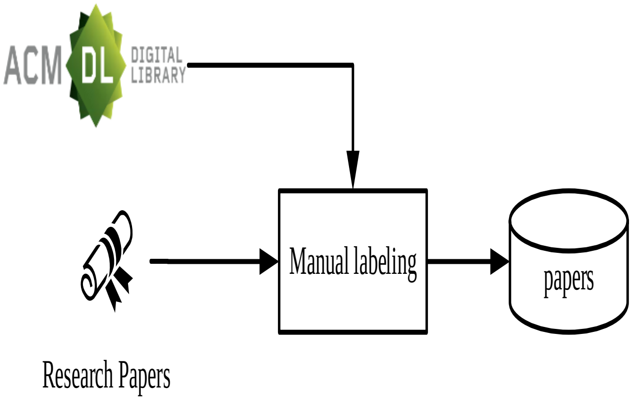

Fig. 3. Labeling dataset procedure.

4.3. Tokenization

Tokenization in text processing is fundamental in splitting text into smaller units, such as words or phrases. It serves as a precursor to more advanced natural language processing tasks, enabling the algorithm to understand the structure and meaning of the text. This process helps reduce the text's complexity, making it more manageable for algorithms. Figure 4 shows an example of the tokenization process:

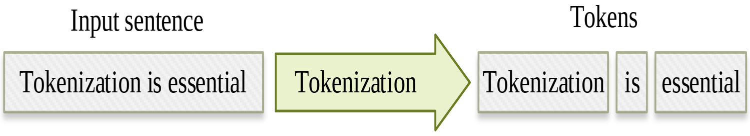

Fig. 4. The tokenization process example

4. 4. Part of Speech

Part of Speech (POS) is the essential algorithm in NLP. It classified words _ the words in the sentence_ based on their purpose of use in the sentence. The primary classification of English words is nouns, verbs, adjectives, and adverbs. Error! Reference source not found. illustrates the example process POS of two sentences.

Table III Example Of Pos Processing

Sentence	The function of the word "run"
The morning run always leaves me feeling energized for the day ahead.	noun
I plan to run the same route tomorrow to beat my personal best	verb

Technically, POS algorithms can follow set rules or deepen on the probabilistic models to classify words according to their possible sentence functions.

4.5. Lemmatization

Lemmatization is an NLP technique that simplifies words to their root form or lemma. The fundamental objective of the lemmatization in the NLP is to unify different forms of a word for cohesive analysis, contributing to standardized language and streamlined textual data examination.

4.6. Stop Word Eliminate

The stop words in language usually include pronouns, articles, prepositions, conjunctions, and other words that carry little meaning or occur frequently. Examples of stop words in English include "the," "and," "but," "how," "or," ,"what". So on. Eliminating stop words is one of the basics of natural language processing because stop words or letters do not have a meaning. Figure 5 shows the example of a stop-word elimination process in the sentence "The Artificial *intelligence is a branch of computer science."*.

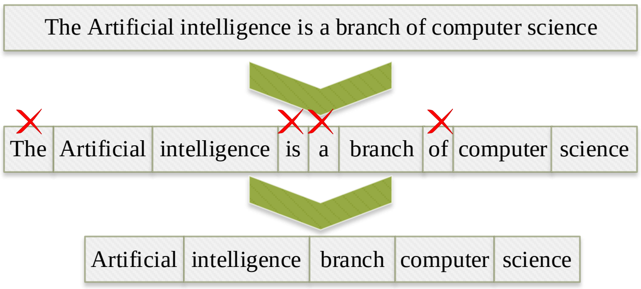

Fig. 5. Stop Word Elimination Process Example.

4.1. Feature Extraction by Using TF-IDF

Feature extraction entails the process of carefully choosing and transforming pertinent data from its raw form into a concise and comprehensible representation. In the domains of machine learning and data analysis, this process entails converting incoming data, regardless of its intricacy or absence of organization, into a set of characteristics that may be utilized for detecting patterns or training models. When working with textual data, various strategies are available for extracting features. These include translating documents into numerical forms using methods such as word embedding, TF-IDF, and bag-of-words. These techniques are commonly used in natural language processing.

model, particularly the term frequency-inverse document frequency (TF-IDF) model, is a widely used method for feature extraction in text categorization [17] [18] [19]**.** The (TF-IDF) methodology is a Term Weighting method that is employed for categorization. The usage of TF-IDF is supported by research which demonstrates that the use of TFIDF for word weighting has a significant impact on categorization and obtaining better and optimized results [20].

Figure 6 shows the basic operations in the feature extraction in the proposed model. The datasets were collected from the step text preprocessing.

Term Frequency (TF) measures how frequently a term occurs in a document. For a term t in document d, the term frequency $TF(t,d)$ is calculated as:

$$TF(\, t,d) = \frac{N_t}{N_{Total}} \quad \dots \ (1)$$

Where N_t is refer to the Number of times term t appears in document d, N_{Total} and refer to the Total number of terms in document d
Inverse Document Frequency (IDF) measures how important a term is within the entire corpus. It is calculated as:

$$\mathrm{IDF}\ (\mathrm{t},\mathrm{D}) = \log\left(\frac{\mathrm{N_D}}{\mathrm{N_t}}\right) \quad \dots (2)$$

Where N_D refer to Total number of documents in corpus D, and N_t refer to the Number of documents containing term t. Combining these two metrics calculates the TF-IDF score for a term. For term t in document d from the document set D, the TF-IDF score is computed as:

$$\mathrm{F_1\ score} = \frac{2\mathrm{TP}}{2\mathrm{TP+FP+FN}} \quad \dots (3)$$

The higher the TF-IDF score, the rarer the term is in the entire document set and the more significant it is to the individual document.

Euclidean Distance, Cosine Similarity: We use measures like Euclidean distance and cosine similarity to compare documents or terms. Euclidean distance is calculated as the square root of the sum of the squared differences between corresponding elements of the vectors. For two documents represented by vectors A and B, it is calculated according to Error! Reference source not found.):

$$\mathrm{d(A,y)} = \sqrt{(\mathrm{A_1} - \ \mathrm{B_1})^2 + \ (\mathrm{A_2} - \ \mathrm{B_2})^2} \dots (4)$$

Cosine similarity measures the cosine of the angle between two non-zero vectors of an inner product space, which provides the cosine of the angle between them, calculated according to Eq. :

$$\mathrm{Similarity(P,Q)} = \frac{\mathrm{P{\cdot}Q}}{|\mathrm{P}||\mathrm{Q}|} \quad \dots (5)$$

The resultant vectors, compared through Euclidean distance or cosine similarity, help identify the useful features that contribute most significantly to the learning task. The selected useful features improve the ability of the machine learning model to understand and classify documents based on the significance of terms

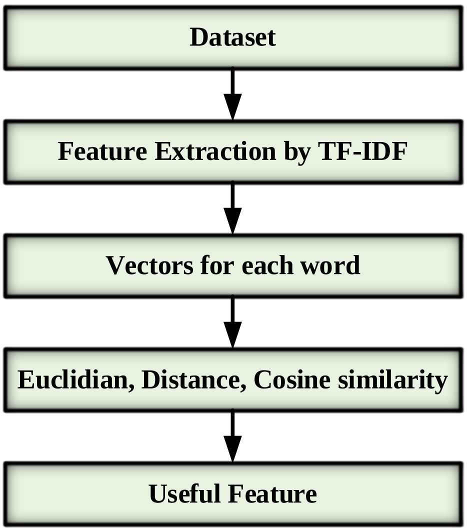
Dataset
Feature Extraction by TF-IDF
Vectors for each word
Euclidian, Distance, Cosine similarity
Useful Feature

Fig. 6. Features extraction's basic operations

A. Text Classification by Deep Neural Network Classifier (DNNC)

A DNN classifier, belonging to the broader family of artificial neural networks (ANNs), is a machine learning model employed for classification purposes. It is characterized by its deep architecture, featuring multiple layers. Deep neural network classifiers have proven effective across various domains such as image recognition, natural language processing, speech recognition, and beyond. Their success is attributed to their capacity to autonomously acquire hierarchical representations from data, the constructed Deep Neural Network (DNN) comprises an input layer, an output layer, and three intermediary hidden layers. The input encompasses the 11,152 features, while the output produces probabilities corresponding to the 10 distinct class labels. The hidden layers contain 700, 128, and 10 nodes, as illustrated in Figure 2. The Rectified Linear Unit (ReLU) activation function is applied between the first pair of consecutive layers, while for the last layer using activation function (Softmax) as shown equation below:- [21]

$$\mathrm{f}\,(\mathrm{x}) = \mathrm{x}^{+} = \max(\,0, \mathrm{x})\ \ldots\ (6)$$

In this context, 'x' represents the output from the preceding node, and '$\boldsymbol{f}\ (\boldsymbol{x})$' denotes the input for the subsequent node. The activation function was initially introduced to a dynamic network.

Supported by mathematical validations, this activation function has been proven to facilitate more effective training of deep networks when contrasted with commonly employed activation functions such as the logistic sigmoid function and its more practical equivalent, the hyperbolic tangent. In this architecture, the loss function is represented by the cross-entropy cost function as shown in this equation below. [20]

$$\mathrm{J(w)} = -\frac{1}{\mathrm{N}}\sum_{\mathrm{n=1}}^{\mathrm{N}}[\,\mathrm{y_n}\log \mathrm{y}_{\mathrm{n}}^{\wedge} + (1 - \mathrm{y_n}\,)\log(1 - \mathrm{y}_{\mathrm{n}}^{\wedge})\,]\ \ \ldots\ (7)$$

In this context, y_n corresponds to the actual class of each sample, $y_n^{\wedge}$ represents the output from the Deep Neural Network (DNN), and 'N' denotes the total number of inputs

B. Classifier Performance

Performance evaluation involves not only determining the best classification but also assessing whether it is capable of doing the purpose it was designed to do. As well as it used in the optimal model for the current problem can be chosen by comparing the performance of many models that use the same dataset. This practice allows academics and practitioners to contrast the benefits and drawbacks of several models and select the most suitable one for their particular use case. The metrics used to evaluate model performance are Precision, Recall, f1score, and accuracy, which are calculated using a confusion matrix.

So, evaluation based on labels examines each label independently, transforming a multi-label classifier into a binary classifier for a specific label. This results in four potential prediction outcomes: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Accuracy, Precision, and Recall are then defined as follows:

DOI: https://doi.org/10.25195/ijci.v50i1.475
Print ISSN: 2313-190X, Online ISSN: 2520-4912

i. Precision

within the realm of classification, serves as a metric to gauge the accuracy of a model's positive predictions. It is computed as the ratio of true positive predictions to the sum of true positives and false positives, according to Error! Reference source not found.. Essentially, precision assesses the exactness or precision of the positive predictions made by the model[22]

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \ldots (8)$$

ii. Recall

is a classification metric that evaluates a model's ability to recognize all relevant instances of a given class, calculated by dividing true positive predictions by the sum of false negatives according toError! Reference source not found.[21]

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \ldots (9)$$

iii. Accuracy

is a widely used metric to assess the effectiveness of a model or system, evaluating the overall correctness of its predictions, according toError! Reference source not found.[23]

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{N}} \quad \ldots (10)$$

iv. F1 Score

is a metric that balances the trade-off between precision and recall, providing a more comprehensive evaluation of a classifier's performance, especially in scenarios with imbalanced classes [24]. The $F1score$ based on labels is characterized by the following definition: [22]

$$\text{F}_1\text{Score} = \frac{\text{2TP}}{\text{2TP+FP+FN}} \quad \ldots (11)$$

5. EXPEREMENTAL RESULTS

This section identifies the Application Programming Interface (API), the method of retrieving the dataset, its size, and the source from which it is retrieved. It also explains a general description of the nature of that dataset, its classification, and the source relied upon in the classification process, in addition to the number of classes used.
Finally, the results are presented in a Error! Reference source not found., and the findings are discussed.

5.1 Dataset

In this study, the dataset used is a collection of metadata (author's papers which show details in Table 1) for computer science and Iraqi research for years from 2010 -2022. The data will then be classified manually based on the ACM journal and then entered into preprocessing such as techniques in tokenizing, POS, lemmatization, filtering, and also feature extraction is made as a set of processes and the text (title & abstract) will be converted into numerical data called vector features, after that, it will be inputted into the DNN model. The results of the study will be an approach for classifying the analysis text of papers based on the tree classification for an ACM journal. This section consists of two steps:

***API*: -** is a set of defined rules that explain how computers or applications communicate with one another. (Typically these rules are documented in an API specification). APIs sit between an application and the web server, acting as an intermediary layer that processes data transfer between systems [25].

***API-Key:* -** is a unique digital string that is obtained from the website from which you want to obtain data, as it is considered an essential part of digital security, so the purpose of using an API key is to allow only the authorized person to access the database [23].

After getting the API-key from Scopus (the steps and all details exist in [1])by using it the data was retrieved from the Scopus database by Python, VS. code(platform), and scrapping. For the subject area is" computer science "and the affiliation country is "Iraq" From (2010-2022), additionally the statistics of the retrieved results reached

[1] https://dev.elsevier.com/

(1170) papers. So, the retrieved fields are shown Error! Reference source not found.. (note: we can retrieve another field as needed for academic research)

TABLE III. Example Dataset Retrieval

No.	terms	Description
1	Author name	First Author (author first name entry)
2	title	Paper title
3	Abstract	paper abstract content (text of papers abstract)
4	Affiliation name	Name of university
5	URL	Content Abstract Retrieval API URI
6	coverDate	Publication Date (YYYY-MM-DD)
7	aggregation type	Source Type (if paper journal or conference)
8	subtype Description	Document Type Description (if the paper is article, journal, review, book or conference)
9	Source-id	identifier for a source/publication venue (identifier of each paper) in Scopus
10	identifier	Scopus ID (It is a unique number assigned to each researcher in the Scopus database)
11	Class label	Special branch in computer science
12	link ref=search	Scopus abstract detail page URL (It refers to a general summary of the research, meaning it does not show details of the entire paper)

After data is obtained then manually classified according to the an ACM journal scientific website (digital library) such as (general and reference, Following this, the text undergoes preprocessing to enhance its suitability for subsequent operations. Important word features are then extracted using the TF-IDF technique, converting the text into numerical data and vectors. Finally, these values will be utilized to construct a deep neural network model for developing a paper classification model based on text (title & abstract) analysis. After collecting the papers, a determination of the class label (precise specialization) will be made for each paper. Labeling papers based on the ACM journal scientific website (digital library) tree classification which contain (10) classes as shown in Error! Reference source not found..

TABLE IV. Example Of The Paper Labeling Using (10) Classes

No.	Class
1	Apply Computing
2	Information System
3	Computing Methodology
4	Mathematics of Computing
5	Social and Professional Topics
6	Theory of Computing
7	Network
8	Hardware
9	Human-Centered Computing
10	Software and its Engineering

After completing the classification of the papers manually through text (title& abstract) analysis based on the tree classification of an ACM (journal scientific website (digital library)), each paper will be classified according to its research orientation using ten classifications from the ACM journal.

5.2 Result

The processed data from the corpus will be separated into two groups for training and testing purposes during the splitting corpus step. Data Training serves as a classification model, and the classification outcomes are utilized to create a model to forecast the labels on Data training 80% of the data is used for training, while 20% is used for testing. The dataset for this study was divided into 300 epochs that would be used for categorization.

In this study, the DNN Method will undergo two distinct phases: Training Stage and Testing Phase. During the training stage, an analysis of the sample document will be conducted using Training Data. This analysis involves selecting vocabulary, which comprises words likely to occur in the sample collection, serving as representations for papers corresponding to each label. Following this, during the testing phase, Testing Data will be employed to ascertain the paper's label based on the vocabulary, which already represents scientific terminology pertinent to the researcher's area of interest and the paper's type for each label. The study employs the TFIDF feature combined with DNN layers, which yield average accuracy, Micro average, and Weighted average to determine precision, recall, and F1 score. The resulting average accuracy is subject to variation based on the inclusion of additional layers.

After getting an average accuracy value based on the usage, to see the effectiveness of the classifier performance, an accuracy calculation based on precision, recall, and f1 score is calculated based on the confusion matrix. After that, the model is built, consisting of 11 layers, each containing a certain number of units (neurons). Also, the activation function used in all layers is Relu, and Bayas is True, while for the last layer (the number of labels is 10), it is Softmax. So, overall, the model contains an optimizer named Adam, and the loss function used is cross-entropy. The metrics used are accuracy, precision, recall, and f1 score. In addition, the epoch used is 300 epochs, and the batch size is 32. Finally, the execution time it takes is 305,143 milliseconds or approximately 7 minutes.

Error! Reference source not found. shows the results obtained through the proposed model and methods, it consists of 11 layers, each layer measured with certain values. It is noted that as the number of layers increases, the accuracy of all proposed measurements decreases, and it is possible to consider this partial detail as a suggestion that can be known in the future.

Table V. Result In The Execution Of The Dnn Model

No. of layer	Metrics	Precision	Recall	F1 score	Accuracy
3	Micro average.	90%	90%	90%	90%
	Weighted average.	90%	90%	90%	
4	Micro average.	88%	88%	87%	88%
	Weighted average.	88%	88%	88%	
5	Micro average.	85%	85%	85%	86%
	Weighted average.	86%	86%	85%	
6	Micro average	85%	83%	84%	83%
	Weighted average	86%	83%	84%	
7	Micro average	84%	82%	83%	83%
	Weighted average	83%	83%	83%	
8	Micro average	79%	74%	76%	74%
	Weighted average	79%	74%	76%	
9	Micro average	79%	76%	77%	77%
	Weighted average	80%	77%	78%	
10	Micro average	64%	56%	55%	55%
	Weighted average	66%	55%	55%	
11	Micro average	58%	52%	51%	52%
	Weighted average	59%	52%	51%	

Finally, the best accuracy achieved by the model is in three layers, which represent the best architecture that suits this data and the problem to be solved.as accuracy is 90%, recall is 90 %, precision is 90% and f1score is 90 %.

5 CONCLUSIONS

This research focused on collecting 1170 computer science papers from the Scopus repository in Iraq from 2010-2022. Text preprocessing was used to enhance data validity and performance. The data was processed using text frequency features and a deep neural network algorithm, resulting in a 90% accuracy rate for the proposed model. This approach promotes knowledge expansion, professional growth, and reduced research paper acquisition time. Based on the prior research findings, we suggest that there is potential for future researchers to enhance the current outcomes. This could be achieved by employing the CNN (Convolutional Neural Network) algorithm along with Tf-idf as a feature extraction technique. As well as an expansion of subspecialties to more than ten specializations. Also collecting data from local magazines to solve a societal problem because it is easier to obtain and ensures higher accuracy of the results

References

DOI: https://doi.org/10.25195/ijci.v50i1.475
Print ISSN: 2313-190X, Online ISSN: 2520-4912

[1] W. E. Donald, "Content analysis of metadata, titles, and abstracts (CAMTA): application of the method to business and management research," *Management Research Review*, vol. 45, no. 1, pp. 47-64, 2022.

[2] E. Parry, E. Farndale, C. Brewster, and M. J. Morley, "Balancing rigour and relevance: The case for methodological pragmatism in conducting large-scale, multi-country and comparative management studies," *British Journal of Management*, vol. 32, no. 2, pp. 273-282, 2021.

[3] M. S. Tullu, "Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key," *Saudi journal of anaesthesia*, vol. 13, no. Suppl 1, p. S12, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6398294/pdf/SJA-13-12.pdf.

[4] P. Conceição, I. B. Vázquez, F. B. Alves, A. R. Sousa, J. A. Rios, and D. S. Tavares, "Planning Thoughts–The evolution of themes in Master Dissertations and Doctoral Theses in the field of specialisation in Spatial Planning, FEUP," *U. Porto Journal of Engineering*, vol. 9, no. 2, pp. 100-113, 2023.

[5] R. C. Padgett and W. E. Donald, "Enhancing self-perceived employability via a curriculum intervention: a case of “The global marketing professional” module," *Higher Education, Skills and Work-Based Learning*, vol. 13, no. 1, pp. 22-35, 2023.

[6] W. E. Donald and D. Jackson, "Subjective wellbeing among university students and recent graduates: Evidence from the United Kingdom," *International Journal of Environmental Research and Public Health*, vol. 19, no. 11, p. 6911, 2022. [Online]. Available: https://mdpi-res.com/d_attachment/ijerph/ijerph-19-06911/article_deploy/ijerph-19-06911.pdf?version=1654420890.

[7] A. Gaur and M. Kumar, "A systematic approach to conducting review studies: An assessment of content analysis in 25 years of IB research," *Journal of World Business*, vol. 53, no. 2, pp. 280-289, 2018.

[8] Q. Yang, C. Ma, Q. Zhang, X. Gao, C. Zhang, and X. Zhang, "Interpretable Research Interest Shift Detection with Temporal Heterogeneous Graphs," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 321-329.

[9] N. Beech and F. Anseel, "COVID-19 and its impact on management research and education: Threats, opportunities and a manifesto," *British Journal of Management*, vol. 31, no. 3, p. 447, 2020.

[10] H. Aguinis, R. S. Ramani, and N. Alabduljader, "What you see is what you get? Enhancing methodological transparency in management research," *Academy of Management Annals*, vol. 12, no. 1, pp. 83-110, 2018.

[11] J. Baas, M. Schotten, A. Plume, G. Côté, and R. Karimi, "Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies," *Quantitative science studies*, vol. 1, no. 1, pp. 377-386, 2020.

[12] M. Katsurai, I. Ohmukai, and H. Takeda, "Topic Representation of Researchers' Interests in a Large-Scale Academic Database and Its Application to Author Disambiguation," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 4, pp. 1010-1018, 2016.

[13] F. P. C. da Fonseca and L. A. Digiampietri, "Improving researcher’s area of expertise identification using TF-IDF Characters N-grams," in *XVII Brazilian Symposium on Information Systems*, 2021, pp. 1-7.

[14] N. Inie, J. Frich, and P. Dalsgaard, "How Researchers Manage Ideas," in *Proceedings of the 14th Conference on Creativity and Cognition*, 2022, pp. 83-96.

[15] J. Zheng and Y. Liang, "User Interest Identification with Social Media Information using Natural Language and Meta-Heuristic Technique," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, pp. 1-17, 2024.

[16] V. K. Singh, P. Singh, M. Karmakar, J. Leta, and P. Mayr, "The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis," *Scientometrics*, vol. 126, pp. 5113-5142, 2021.

[17] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780-2788, 2021.

[18] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341-348, 2019.

[19] Y. Zhang, Y. Zhou, and J. Yao, "Feature extraction with TF-IDF and game-theoretic shadowed sets," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 18th International Conference, IPMU 2020, Lisbon, Portugal, June 15–19, 2020, Proceedings, Part I 18*, 2020: Springer, pp. 722-733.

[20] A. Aninditya, M. A. Hasibuan, and E. Sutoyo, "Text mining approach using TF-IDF and naive Bayes for classification of exam questions based on cognitive level of bloom's taxonomy," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 2019: IEEE, pp. 112-117.

[21] X. Hou, J. You, and P. Hu, "Predicting drug-drug interactions using deep neural network," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, 2019, pp. 168-172.

[22] E. I. Obeagu, D. Nwosu, and G. U. Obeagu, "Interleukin-6 (IL-6): A Major target for quick recovery of COVID-19 patients," 2022.

[23] Z. Zhang and H. Ji, "Abstract meaning representation guided graph encoding and decoding for joint information extraction," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 39-49.

[24] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label classifier performance evaluation with confusion matrix," *Computer Science & Information Technology*, vol. 1, 2020.

[25] I. C. Education, "What is an application programming interface (API)," *Online]. Tillgänglig: https://www. ibm. com/cloud/learn/api (hämtad 2022-04-11)*, 2020.