IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Keyphrases Frequency Analysis from Research Articles: A Region-Based Unsupervised Novel Approach

MOHAMMAD BADRUL ALAM MIAH[1,2], SURYANTI AWANG[1,3,*], MD MUSTAFIZUR RAHMAN[4], A. S. M. SANWAR HOSEN[5], AND IN-HO RA[6,*]

[1]Faculty of Computing, Universiti Malaysia Pahang, 26600 Pekan, Malaysia (e-mail: badrul.ict@gmail.com)
[2]Dept. of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh (e-mail: badrul.ict@mbstu.ac.bd)
[3]Center of Excellence for Artificial Intelligence & Data Science, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300, Gambang, Kuantan, Pahang, Malaysia (e-mail: suryanti@ump.edu.my)
[4]Department of Mechanical Engineering, Faculty of Engineering, Universiti Malaysia Pahang, Kuantan, Malaysia (e-mail:mustafizur@ump.edu.my)
[5]Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, South Korea (e-mail:sanwar@jbnu.ac.kr)
[6]School of Computer, Information and Communication Engineering, Kunsan National University, Gunsan 54150, South Korea (e-mail:ihra@kunsan.ac.kr)

Corresponding author: Suryanti Awang (suryanti@ump.edu.my) and In-Ho Ra (ihra@kunsan.ac.kr)

This work was supported by the Universiti Malaysia Pahang (UMP) through the FLAGSHIP Research Scheme under Grants (RDU192210 and RDU192212), as well as the National Research Foundation of Korea (NRF) grant by the Korean Government through the Ministry of Science and ICT (MSIT) under Grant 2021R1A2C2014333.

ABSTRACT Due to the advancement of technology and the exponential proliferation of digital sources and textual data, the extraction of high-quality keyphrases and the summarizing of content at a high standard has become increasingly difficult in current research. Extracting high-quality keyphrases and summing texts at a high level demands the use of keyphrase frequency as a feature for keyword extraction, which is becoming more popular. This article proposed a novel unsupervised keyphrase frequency analysis (KFA) technique for feature extraction of keyphrases that is corpus-independent, domain-independent, language-agnostic, and length-free documents, and can be used by supervised and unsupervised algorithms. This proposed technique has five essential phases: data acquisition; data pre-processing; statistical methodologies; curve plotting analysis; and curve fitting technique. First, the technique begins by collecting five different datasets from various sources and then feeding those datasets into the data pre-processing phase using text pre-processing techniques. The preprocessed data is then transmitted to the region-based statistical process, followed by the curve plotting phase, and finally, the curve fitting approach. Afterward, the proposed technique is tested and assessed using five (5) standard datasets. Then, the proposed technique is compared with our recommended systems to prove its efficacy, benefits, and significance. Finally, the experimental findings indicate that the proposed technique effectively analyses the keyphrase frequency from articles and delivers the keyphrase frequency of 70.63% in 1st region and 10.74% in 2nd region of the total present keyphrase frequency.

INDEX TERMS Curve fitting technique, Data pre-processing, Feature extraction, Keyphrase extraction, Keyphrase frequency analysis, KFA technique.

I. INTRODUCTION

The continual expansion of the information age, as well as the exponential growth of textual material, makes managing such a large volume of data even more difficult [1]. Prior to the invention of technology, this vast amount of information could only be processed by people, which took a long time. Moreover, due to discrepancies between the quantity of information and manual information processing skills, it is difficult to complete this vast amount of data, leading to the development of automated keyphrase extraction techniques that use computers' comprehensive computational power to replace physical labor [2], [3]. The purpose of automated keyphrase extraction methods is to extract high-level keyphrases from articles. Keyphrase, in general, gives a high degree of document characterization, summary, and description, which is important for numerous aspects of Natural

Language Processing (NLP), including such things as article classification, clustering, and categorization [1]. "Despite this, they are used in a wide range of Digital Information Processing applications, including Information Retrieval, Digital Content Management, Recommender Systems, and Contextual Advertising. It can also be used for search engines, media searches, legal and geographic information retrieval, and digital libraries, among other things" [1], [3], [4].

To accommodate the above-mentioned applications, several keyword extraction techniques have been established [5]–[11]. Among them, some are domain-specific tactics [5], necessitate application domain expertise; linguistic approaches [7], necessitate language proficiency. As a result, they are unable to tackle problems in other subjects/domains or languages. Supervised machine learning (ML) approaches require a large amount of rare training data to extract quality keystones and generalize poorly outside the domain of the training data, according to [12]. It also increased the storage and computation, decreased the comprehensibility, and made the system computationally expensive [2], [13], [14]. Again, because of the huge number of complex processes, statistical unsupervised techniques such as [9], [15], [16] are computationally expensive. And due to the inability to identify cohesiveness amongst numerous words that compose a keyphrase, graph-based unsupervised approaches perform badly [17]–[21]. Finally, TeKET [8] is extremely versatile and acts similarly to TF-IDF for short data lengths.

For those keyphrase extraction techniques that extract high-level keyphrases from articles, keyphrase frequency analysis (KFA) is required. The proposed KFA technique can be utilized as a feature of those keyphrase extractions to distinguish keywords from other terms. The keyword extraction technique can't extract quality keywords without using good quality features. It has been established that, as a result of the prior debate, keyphrase feature extraction remains a critical research area for the survey.

As a result, this paper provides an unsupervised new KFA technique with the following notable contributions: **Contribution 1**: The proposed approach is corpus, domain, and language agnostic. **Contribution 2**: Both supervised and unsupervised techniques can be benefited from the proposed technique. **Contribution 3**: The proposed method is a document-length-agnostic approach. **Contribution 4**: For testing and evaluating the performance of the proposed technique, five datasets have been utilized and **Contribution 5**: The proposed approach can find the best dataset for the analysis as well as effectively analyze the keyphrase frequency based on region.

The rest of this article is laid out as follows. The various techniques are outlined in section II Related works, together with their strengths and limitations, emphasizing the necessity for a new technique to still be offered. Afterwards, in section III Methodology, a novel region-based unsupervised KFA technique is described for determining the keyphrase frequency in each region of an article. After that, the experiments' setup is described in depth in section IV Experimental setup, which includes datasets details, evaluation metrics, and implementation details. Then, the proposed technique has been tested on five (5) different datasets and evaluated for the effectiveness of the system, and then compared to current methods to determine their benefits and drawbacks, which are seen in detail in section V Results and Discussion. Finally, in section VI Conclusion, the study's contributions, future works, and shortcomings would be identified and stated.

II. RELATED WORKS

The proposed strategy is a fresh approach for analyzing keyphrase frequency from articles that can be utilized as a feature of keyphrase extraction. Hence, the section covers comparable techniques. Based on the training datasets, the majority of keyphrase extraction techniques are classified into two types: unsupervised and supervised [1]. Features as well as feature extraction techniques are used by both approaches. We'll go over the key points of both parties' methods in the sections below.

A. SUPERVISED TECHNIQUES

Using this technique from the article, the keyword extraction methodology is considered a binary category problem, with a fraction of candidate keys categorized as keyphrases and non-keyphrases. Some of the techniques which can be utilized to address the classification problem including Neural networks (NN) [22], Support vector machines (SVM), Naïve Bayes, decision trees (DT), and C4.5 [1], [3]. The most important techniques are reviewed in depth in the following that uses this approach. Keyphrase Extraction Algorithm (KEA) makes advantage of TFxIDF and first occurrence position as a feature [23], [24]. It employs illustrative methodologies for detecting candidate keyphrases, for each candidate estimating feature values, and utilizing the Naive Bayes algorithm to predict and determine candidates' good keyphrases. However, because KEA is dependent on the training dataset, it could give poor results if the training dataset doesn't match the document.

Genitor Extractor (GenEx) automatically takes first occurrence position, keyphrase length, and term frequency (TF) as a feature [25], [26]. The most extensively used keyword extraction method is based on a C4.5 decision-making approach that involves genetic algorithms to maintain its efficiency across domains. This system doesn't employ the TFxIDF technique.

The Hulth system permits the retrieved keyphrases to be as lengthy as they wish to be, in contrast to the KEA and GenEx approaches [26], [27]. "Part of speech (POS) tag, n-grams, first occurrence position, noun phrase (NP) chunks, and TF" [1] are the four properties it uses. Regrettably, there is no correlation between the diverse POS tag attributes. This system does not try against the GenEx/KEA criteria, and it reports that the value of recall is low.

The Maui Algorithm is an automated generalized topical indexing method which is based upon that KEA system [26], [28]. It expands the KEA system by including data from

Wikipedia. However, shortcomings of this algorithm's is its lack of assessment capabilities.

HUMB system uses the location of a term with its initial occurrence; phraseness; informativeness; keywordness; and the candidate term's length as a feature [29]. The HUMB system has shown positive outcomes in a range of data sets. HUMB, on the other hand, relied on external knowledge sources (GRISP, GROBID/TEI, and HAL) that are linked to scientific disciplines.

The Document Phrase Maximality (DPM)-index uses a total of eighteen (18) statistical factors [30]. The DPM index and an additional five (5) are unique features amongst them. Compared to other keyphrase extraction methods, this system's outcomes have improved dramatically without utilizing outer knowledge or manuscript structural elements.

Citation-enhanced keyphrase extraction is a supervised model known as CeKE [31]. The following essential features are used by the CeKE: Relative position, TFxIDF, POS tag, inCited and inCiting, citation TF-IDF, TF-IDF-Over, first position, firstPosUnder, . They have the ability to improve keywords extraction and add important features. In comparison to previous systems, the CeKE+ keyness model produces noteworthy results [27].

Using supervised learning approaches, the Keyphrase Extraction (KeyEx) Method identifies a huge number of probable candidate keys and builds a classifier standard for keyphrase extraction [32]. Experiments by the author revealed that the KeyEx method considerably enhanced the quality of the retrieved key. Additionally, their strategy beats current sequential pattern mining methods.

B. UNSUPERVISED TECHNIQUES

The keyphrase extraction approach using this technique is a ranked problem that can be addressed without any prior experience. These methods are categorized as graph-based or statistical-based, according to [26]. The parts that follow go into great detail about the most essential techniques utilized by both groups.

PageRank [33] is indeed a graph-based method that is built on random walks. It's fine for sifting through web pages and social media pages, but it can't extract crucial information from authorized manuscripts. The PositionRank [20] is the extension of PageRank that has been established to improve performance, and it evaluates words by considering all of their placements and frequency, determining their rank. However, because it overlooks thematic coverage and diversity, this method performs badly.

TextRank [34], [35] employs POS tag like an intrinsic feature, but it has a number of drawbacks, including the difficulty to capture cohesion, which leads to sub-optimal outcomes. Another major extraction technique that overcomes TextRank's restrictions is TopicRank [18]. TopicRank extracts noun phrases from the document and groups them into subjects. It also has a problem with error propagation. TextRank's lengthening is SingleRank [17], [36]. By acquiring ranked words, it accurately extracts just noun phrases from datasets, not keyphrases. In ranking phase, unimportant keywords are used, although this does not always screen out small scoring terms, providing longer keywords greater scores.

The TopicRank propagation matter is resolved using the MultipartiteRank technique [21]. However, it has a clustering inaccuracy, making it difficult to choose the most representative candidates. The well-known unsupervised graph-based keyphrase extraction technique is called TeKET (Tree-based Keyphrase Extraction Technique) [8] which is domain-independent, language agnostic, and requires fundamental statistical understanding. Although this approach beats the several important keyphrase extraction strategies, it includes some drawbacks, like as provides extensive flexibility.

TF-IDF [37] is the most often used statistical approach. Though TF-IDF is straightforward to build, determining Inverse Document Frequency (IDF) with a big dataset needs a lengthy time as well as lots of computer resources. The KP-Miner [38] algorithm is also employed to resolve the matter of single-term preferences. Though KP-Miner outperforms TF-IDF, it's have several disadvantages, such as worsening global ranking performance as the amount of data grows. Since it depends on TF-IDF, it also is computationally costly.

Yet Another Keyword Extractor (YAKE) calculates the weighting scores of a keyword utilizing five attributes: "as term position, casing, term relatedness to context, term frequency normalization, and term distinct sentence" [9]. Furthermore, since it generates candidate keys using the N-grams approach, its computational cost effect increases with this N-grams technique.

The prior discussions demonstrate that the many opposing characteristics of unsupervised and supervised keyphrase extraction technique prevent them from achieving the better performance. Thus, this research offers a new unsupervised KFA technique as feature of keyphrases that will considerably reduce the described weaknesses and also will help to extract high-quality keyphrases from research articles.

III. METHODOLOGY

The keyphrases frequency analysis technique has five important phases (shown in Fig. 1): i) Data acquisition, ii) Data pre-processing, iii) Statistical methodologies, iv) Curve plotting analysis, and then v) Curve fitting technique. The following sections provide a more detailed explanation of the proposed system.

A. DATA ACQUISITION

First, the proposed method acquired five standard datasets from Github (https://github.com/LIAAD/KeywordExtractor-Datasets) to evaluate our proposed method. The five datasets are SemEval2010, citeulike180, fao780, Krapivin2009, and Nguyen2007, which contain a total of 3718 documents, with English language, paper-type documents, and various domains covered (such as computer science, agriculture, and misc.) [39]. There are two types of files in every dataset named: keys files that contain the goldkey and documents

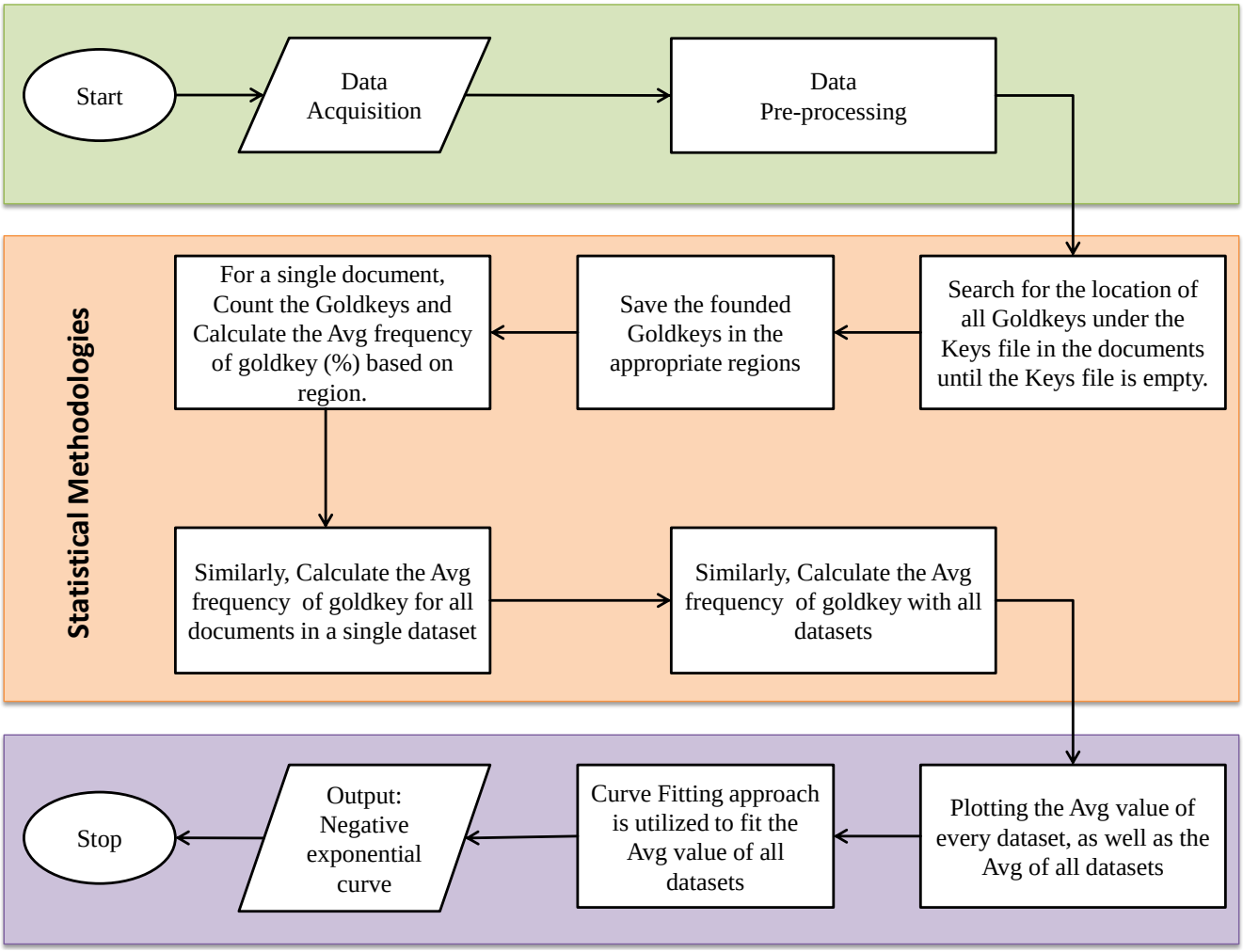


FIGURE 1. The proposed architectural flow diagram for the KFA technique

files (named as docsutf8) that contain the articles. More information on the dataset can be found at section IV-A.

B. DATA PRE-PROCESSING

Data needs to be highly polished as well as of high quality in order to produce strong analytical results when employing machine learning techniques. A variety of preprocessing techniques were used in order to prepare the original dataset for a high-quality analysis [40]. For this purpose, at first the proposed approach extracts both the "docsutf8" files (which contain multiple vital documents as a text files) and the "keys" files types (which contain multiple vital keyphrases/goldkey as a text files). Sometimes, datasets contain several issues like punctuation, accent marks and other special characters, numbers, white spaces, abbreviations, and uppercase etc. In our proposed technique, Various pre-processing techniques have been used, including lowercasing; removing numbers; eliminating punctuation; and removing whitespaces, to obtain the desired level of data quality [41]. After that, the splitting approach is performed to the Keys files to calculate the number of keyphrase gained as goldkey based on the Newline ($backslash$n) function. After that, the proposed technique consider the length of document is in eight (8) regions as well as consider the first appearance keyphrase to analysis the keyphrases frequency. The text pre-processing techniques are explained in details in the following sub-subsections.

1) Lowercasing

The conversion of the text into lower case is one of the most frequent preprocessing steps. Documents in the dataset normally contain both uppercase and lowercase text. When this kind of data is utilized for classification, classifiers identify several variations of the same input class. Being case sensitive, classifiers classify "cast" and "CAST" as 2 distinct inputs. This issue is resolved by converting the entire dataset to lowercase [40]. The lowercasing output is show in Table 1.

TABLE 1. Data preprocessing output for lowercasing; removing number, punctuation, and whitespaces

Input Text:	**" \t Box &A Contains 3 Red [and] 5 White Balls!!!, While Box B Contains #4 Red and 2% Blue* Balls. The 5 biggest countries {by} population (in) 2017 are China!, India!, ?United States; Indonesia, and ~Brazil. \t "**
Output after Lowercasing:	box &a contains 3 red [and] 5 white balls!!!, while box b contains #4 red and 2% blue* balls. the 5 biggest countries {by} population (in) 2017 are china!, india!, ?united states; indonesia, and ~brazil.
Output after Removing Numbers:	box &a contains red [and] white balls!!!, while box b contains # red and % blue* balls. the biggest countries {by} population (in) are china!, india!, ?united states; indonesia, and ~brazil.
Output after Removing Punctuation:	box a contains red and white balls while box b contains red and blue balls the biggest countries by population in are china india united states indonesia and brazil
Output after Removing Whitespaces:	box a contains red and white balls while box b contains red and blue balls the biggest countries by population in are china india united states indonesia and brazil

2) Removing Numbers

Sometimes, the datasets contain the numbers in the documents which is irrelevant. For this reason that irrelevant numbers needs to remove from the documents for our analysis. The proposed technique uses regular expressions to eliminate numbers. The output of this process is shown in Table 1.

3) Removing Punctuation

Since the documents contain punctuation, accent marks, and other special characters, they need to be removed from the documents in this step. Python's string library has a predefined list of punctuation, including $!"\#\%\&'()*+-./:;?@[]\{\}, \$$. The output for this technique is also given in same Table 1.

4) Removing Whitespaces

Sometimes, documents contain leading and trailing spaces, which are unnecessary for the analysis. For this reason, it needs to be removed from the documents. To remove leading and ending spaces, the proposed system utilizes the strip() function. To see the output of this process, visit the same Table 1.

C. STATISTICAL METHODOLOGIES

This is a crucial step after the data pre-processing phase. This step uses the output from the data pre-processing step to analyze the keyphrase frequency based on article regions. This phase consists of two essential processes, such as goldkey/keyphrase searching and saving, and keyphrase frequency counting and averaging, described in the following sections.

1) Goldkey Searching and Saving

During this stage, the proposed approach searches for the location of the goldkey under the keys file in the document. If the goldkey is found in the document, save the location of that goldkey in the appropriate region for further processing. After that, the proposed technique searches for the next goldkey from the keys file in the documents. It's important to note that the location of the goldkey is saved in a two-dimensional (2D) array, with the document region number in the column and the goldkey's number in the row. If that goldkey is not found in the documents, the proposed technique is to look for the next goldkey in the documents. Even If in any document, there is no goldkey/key, the proposed technique provides the location value is -1 for those goldkeys

and starts search for the next documents as well as for the next keys file. Mention here that all datasets contain two types of files: "documents" files (there is no blank document and also every document has the goldkey) and "keys" files (which contain the goldkey). For more details, see section III-B. This operation will resume unless goldkey has finished reading the keys file for one document as well as for a particular dataset. Every dataset will be processed in the same way.

2) Frequency Counting and Averaging

To begin, count the number of goldkey and compute the Average (Avg) frequency of goldkey (%) value based on articles regions for a given document and save this frequency Avg value in a separate two - dimensional array where the row indicates the document's number in a dataset are exist and the column represents the number of document regions, as described earlier [3]. "The process will then repeat until all of the documents for a given dataset have been completed. Similarly, again calculate the frequency Avg value of each region for all documents in a dataset and save it in another 2D array with the row as the dataset's number and column as the same as the previous array. The average (Avg) calculation procedure will then continue until all datasets have been conducted" [1]. Finally, Calculate the Avg for all the regions for all datasets without fail. The Average calculation process of our proposed methodology employed the following equation (1). Where, N,D_N, and R_N represents the number of datasets, documents, and regions respectively. X denotes the keyphrase or goldkey frequency based on region.

$$Avg = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{D_N}\sum_{j=1}^{D_N}\left(\frac{1}{R_N}\sum_{k=1}^{R_N}X_{k,j,i}\right)\right) \quad (1)$$

D. CURVE PLOTTING (CP)

After the frequency counting and averaging processes, CP is a very vital step for our proposed system. It's a graphical representation technique for all types of values as well as a dataset, and it's pretty beneficial in data statistics and analysis. Our proposed methodology of keyphrases frequency analysis based on article region is explained using CP. As a result, the average (Avg) value of every dataset is represented separately from the Avg value of all datasets.

E. CURVE FITTING TECHNIQUE (CFT)

"CFT can be utilized to analyze linear, polynomial, and nonlinear curves after the CP process. The method of finding the best-fitted curve or mathematical operation is likely given a group of data points that's limited" [1]. In our proposed methodology, CFT is employed to find the keyphrase frequency based on the article's region as well as to show the number of keyphrases found in each region. As an outcome, CFT is applied to the average value across all datasets, leading to a negative exponential curve using our proposed method.

IV. EXPERIMENTAL SETUP

Our recommended technique explicitly states the experimental setup contains the dataset details, evaluation measures, and implementation details presented in the following sections. Later on, in section V, the outcomes are briefly presented.

A. DATASET DETAILS

To demonstrate the effectiveness of the proposed technique, it was tested on five (5) datasets such as SemEval2010, citeulike180, fao780, Krapivin2009, and Nguyen2007 [39]. Another goal was to figure out how the proposed method behaved across a variety of datasets. The preceding section III-A has a concise summary, while Table 2 contains a statistical assessment of all datasets. This table includes language names, document types, names of domains, number of documents, total goldkeys, present & absent goldkeys, and the processing time for all datasets are explained. The next paragraph go through each corpus in great depth. **citeulike180** [39] based on CiteULike.org, this dataset covers full-text paper type documents. It is also covered the miscellaneous domain, 183 documents, 3187 goldkeys in which 2071 goldkeys are present and the rest are absent, and the processing time of 0.531 sec.

fao780 [39] With 780 documents, the dataset is based on agricultural papers gathered from two databases based on the United Nations' Food and Agriculture Organization (FAO). It consists of 780 full-text papers chosen at random from the FAO's collection, where the 6215 goldkeys (3702 are present and 2513 are absent) were manually tagged with keywords from of the Agrovoc vocabulary by professional FAO staffs.

The biggest collection in terms of quantity of documents is **Krapivin2009** [42], which contains 2304 full papers from the Computer Science discipline released in ACM between 2003 and 2005. The publications were obtained from the CiteSeerX Autonomous Digital Library, and the authors assigned keywords to each one, which were then confirmed by the reviewers. It's included 12296 goldkeys in which 9933 are present and 2363 are absent, and processing time of 0.984 seconds.

SemEval2010 [43] is the most well standard datasets, with 244 complete scientific papers taken from the ACM Library. The articles are 6 to 8 pages long and address four dimensions of computer science: distributed artificial intelligence, information search and retrieval, social and behavioral sciences, and distributed systems. The author as well as professional editors designate a set of keywords to each article. It is also covered 3129 goldkeys are present and 656 are absent of total goldkey as well as the processing time of 1.078 sec.

Nguyen2007 [44]: This dataset contains 209 scientific documents and 2507 goldkeys in which 2008 are present and the rest are absent. Three papers were provided to student volunteers to read before goldkeys were handed manually. On average, every document contains twelve (12) goldkeys.The processing time of 0.578 sec.

TABLE 2. Summary of the dataset for the study of absent and present goldkeys

Dataset	Language	Type of Doc	Domain	#Docs	#GoldKeys	#Present Goldkey	#Absent Goldkey	Processing Time (sec)
citeulike180	EN	Paper	Misc.	183	3187	2071	1116	0.531
fao780	EN	Paper	Agriculture	779	6215	3702	2513	1.860
Krapivin2009	EN	Paper	Comp. Science	2304	12296	9933	2363	0.984
SemEval2010	EN	Paper	Comp. Science	243	3785	3129	656	1.078
Nguyen2007	EN	Paper	Comp. Science	209	2507	2008	499	0.578

B. EVALUATION METRICS

Accuracy (Acc), precision, recall, and F1-score are the most important and relevant metrics that are frequently used to evaluate/assess a system's performance. Accuracy is the key classification/prediction evaluation metrics since it shows how effectively a classification/prediction model predicts the class label for unidentified samples [4]. The accuracy measure is the proportion of correct predictions produced out of it's whole number of patterns investigated. The accuracy is represented by the equation (2).

$$Acc = \frac{T_{Pos} + T_{Neg}}{T_{Pos} + F_{Pos} + T_{Neg} + F_{Neg}} \tag{2}$$

Where, True Positive (T_{Pos}) and True Negative (T_{Neg}) represents the number of correctly recognized positive and negative keywords, respectively. On the contrary, False Positive (F_{Pos}) as well as False Negative (F_{Neg}), reflect the number of positive and negative keyphrases that were incorrectly recognized.

Again, *Precision* is the ratio of properly expected positive values with respect to the total expected values. Another word, it is employed to calculate the positive patterns that are correctly predicted from the total predicted patterns in a positive class. It can be calculated using the following equation 3:

$$Precision = \frac{T_{Pos}}{T_{Pos} + F_{Pos}} \tag{3}$$

On the other hand, *Recall* is the ratio of accurately expected positive values with respect to the actual positive values; and can be calculated using the following equation 4:

$$Recall = \frac{T_{Pos}}{T_{Pos} + F_{Neg}} \tag{4}$$

Again, *F1-score* is the weighted average of Precision and Recall, which can be calculated using the following equation 5.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

The *F1-score* metric is much more sophisticated than conventional accuracy metric since it takes both false positives and false negatives into consideration.

C. IMPLEMENTATION DETAILS

Python 3.6 and the Spyder-IDE are used to implement the proposed method. It is a high-level, object-oriented programming language that is straightforward to learn and use. It has a user-friendly, adaptable data structure that is supported by a variety of libraries. It is open-source and free, increases productivity, and is interpretative and dynamically typed. It's employed in a variety of domains, including big data, machine learning, and cloud computing. Following that, the laptop is outfitted with an Intel Core i7 CPU, 12GB of RAM, a 256GB SSD drive, and Windows 10 OS [1], [3].

V. RESULTS AND DISCUSSION

In this section, the experiment's outcomes are thoroughly examined. The proposed technique splits the document length into eight (8) regions to analyze the keyphrase frequency. When the number of regions is increased by more than eight, the first portion has a lower keyphrase frequency than the eight-region system. Likewise, if there are less than eight regions, 1st portion seems to have a higher keyphrase frequency than the eight-region system. The proposed method aims to display the frequency of articles based on each region. As a result, the model considers the article length in eight regions instead of expanding or reducing the areas. The two main phases of this section are result analyses and comparisons of proposed method, which are explained in the next subsection.

A. RESULTS ANALYSIS

In this step, the proposed system's performance is assessed by utilizing the following three forms of analysis: dataset analysis, plotting analysis, and curve-fitting analysis.

1) Dataset Analysis

To evaluate the effectiveness of the recommended strategy, the proposed method is tested using five (5) distinct datasets (see in details in section IV-A). Following that, based on the examination of the datasets, the proposed scheme estimates the document's (doc) number, the total goldkeys, total absent goldkeys and total present goldkeys, and processing time (sec) in per datasets presented in Table 2. As demonstrated in Fig. 2, the average number of keywords/goldkeys present and absent each doc is studied for each dataset. Likewise, analyze the average number of absent and present goldkeys in

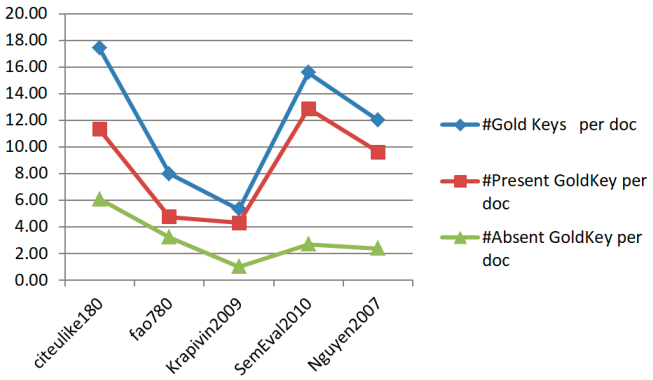


FIGURE 2. Analyze the goldkeys per doc as well as the presence and absence of goldkeys

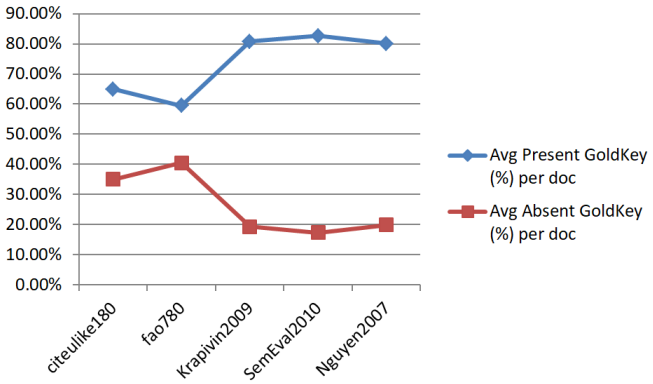


FIGURE 3. Analyze the Avg percentage (%) of present and absent goldkeys per doc for all datasets

percentage (%) per doc for all datasets is shown in Fig. 3. The predicted results/outcomes are summarized in a confusion matrix for all datasets is shown in Fig. 4. Since the proposed technique is a keyphrase frequency analysis technique and not a keyphrase extraction technique, it has no Actual Negative value for the confusion matrix for the goldkey/keyphrase. If the goldkey is present in the documents, the proposed technique can find it easily. Similarly, if the goldkey is absent from the documents, the technique can't find that goldkey. For this reason, the confusion matrix has only Actual Positive value for the goldkey/keyphrase. So, F_{Pos} and T_{Neg} are always zero, as well as T_{Pos} and F_{Neg} have a value for every dataset. Our findings show that on average, 73.59% of keyphrases are present per doc throughout all datasets, whereas 26.41% are missing/absent.

2) Plotting Analysis

According with earlier discussion, since an average of 73.59% of keyphrases/goldkeys are available in each doc throughout the whole dataset, all of the outcomes inside this study are focused on the 73.59% of goldkeys that are present. In our proposed method, the first occurrence keyphrases in a text are considered, and the text length is broken into eight portions. The proposed technique then displays the Avg values of the five (5) datasets together, and then the Avg values of the all datasets relying on each article region. Whenever the document length is segmented into eight portions, Fig. 5 illustrates the analysis of keyphrase frequency in percent (%) based on every region by considering the 1st appearance keyphrases. Likewise, Fig. 6 depicts the assessment of Avg keyphrase frequency in percent (%) based on each portion/region by same consideration as previous for KFA technique. Since all the datasets curve are negatively exponential together, it is proven that the maximum keyphrase frequencies are discovered in 1st region/portion of articles, then in 2nd portion, and so forth, illustrated in Fig. 5 and Fig. 6.

3) Curve Fitting Analysis

Following the plotting analysis, we utilize the average value from all datasets to assess our proposed approach. The system then tries to discover the negative exponential equation by finding the fitted curve for every region's average value. The assessment for the curve fitting approach of the proposed KFA technique for each portion/region is exhibited in Fig. 7, with the document/text length partitioned across eight (8) portions, providing the equation (6) of negative exponential, where $m = 4.04$, $n = 2.09$, and $r = 0.02$.

$$y = m * e^{-nx} + r \tag{6}$$

Since the fitted curve from the curve fitting analysis is negative exponential, it is also proved that the majority of the keyphrase frequencies are located in 1st region of articles, then in 2nd area/region, and so on, as shown in Fig. 7. Finally, the proposed approach can write from Fig. 6 and Fig. 7 that the keyphrase frequency of 51.98% in the 1st region, then 7.90% in the 2nd region, then 4.12% in the 3rd region, and so forth are found from the total of 73.59% of present keyphrases.

B. COMPARISON OF PROPOSED TECHNIQUE

In this phase, firstly, the proposed technique compares all the datasets' performance to find a better dataset. Secondly, the proposed approach compares our two recommended approaches to find the best model or approach. Therefore, this phase consists of two types of comparison: comparisons for finding a better dataset and comparisons for finding a better model/approach that are described in the following sub-subsection.

1) Comparison for Finding a Better Dataset

The proposed technique uses the evaluation metrics (such as accuracy, precision, recall, and f1-score) to measure the performance of each dataset to find a better one by using the confusion matrix shown in Fig. 4. In our proposed approach, *Precision* is always 100%, and *Accuracy* and *Recall* values are always the same because F_{Pos} and T_{Neg} are always zero. The performance comparison of all datasets to find a better dataset is shown in Table 3. From this table, we can

	Actual Positive	Actual Negative
Predicted Positive	2071	0
Predicted Negative	1116	0

Confusion Matrix of citeulike180

	Actual Positive	Actual Negative
Predicted Positive	9933	0
Predicted Negative	2363	0

Confusion Matrix of Krapivin2009

	Actual Positive	Actual Negative
Predicted Positive	3702	0
Predicted Negative	2513	0

Confusion Matrix of fao780

	Actual Positive	Actual Negative
Predicted Positive	3129	0
Predicted Negative	656	0

Confusion Matrix of SemEval2010

	Actual Positive	Actual Negative
Predicted Positive	2008	0
Predicted Negative	499	0

Confusion Matrix of Nguyen2007

FIGURE 4. The confusion matrix of all dataset for the KFA technique

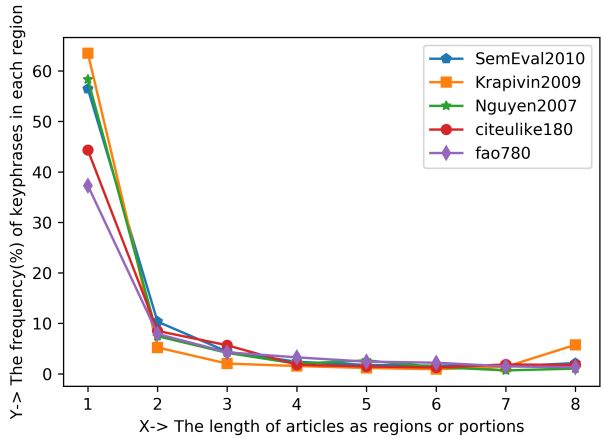


FIGURE 5. The analysis of keyphrase frequency in percent(%) using first occurrence keyphrase based on eight regions for all datasets

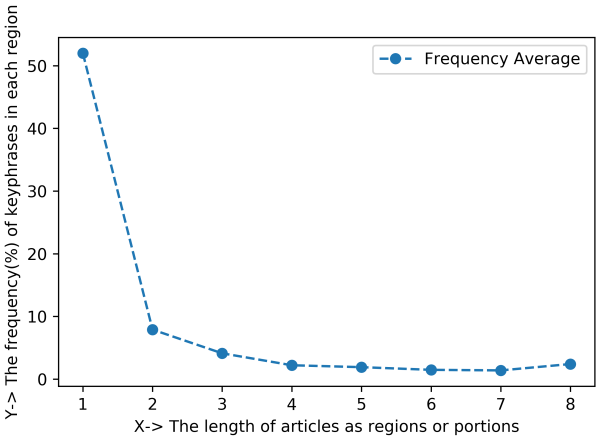


FIGURE 6. The analysis of Avg keyphrase frequency in percent(%) using same consideration for KFA technique

write that the "SemEval2010" dataset provides the highest accuracy of 82.67% and the highest F1-score of 90.51%, and the "Krapivin2009" dataset provides the 2nd highest accuracy of 80.71% and the 2nd highest F1-score of 89.33%. Finally, it is demonstrated that the "SemEval2010" dataset is better than other datasets in our proposed approach.

2) Comparison for Finding a Better Approach

Since the proposed KFA is a novel approach with really no established procedures, it cannot be compared to other techniques. In this section, the proposed approach compares our two recommended approaches, such as eight (8)regions and sixteen (16) regions for the KFA technique, as shown in Table 4. For both approaches, the proposed method uses five (5) datasets. According to Table 4, 51.98% and 44.11% of keyphrase frequency in the first region, 7.90% and 7.87% of keyphrase frequency in the second region, and 4.12% and 4.67% of keyphrase frequency in the third region are found for eight-regions, and sixteen-regions approached, respectively. The eight-region strategy delivers more keyphrase frequencies in the first, second, and subsequent regions than the 16-region approach. Finally, these two approaches support the proposed KFA technique in the article.

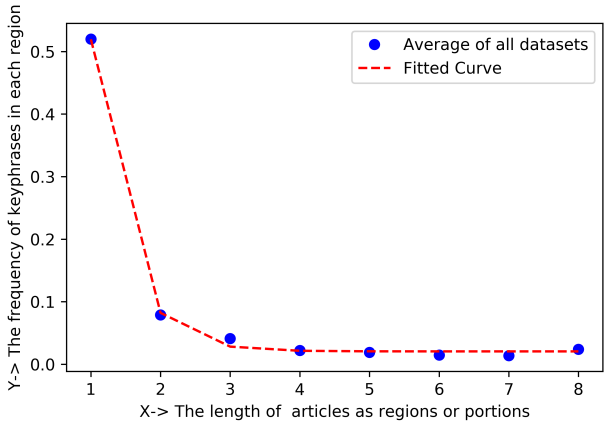


FIGURE 7. The analysis of curve fitting technique for the Avg value of all datasets based upon eight regions

TABLE 3. Performance comparison of all datasets for finding a better one

Dataset	Performance measurements (%)			
	Accuracy	Precision	Recall	F1-Score
citeulike180	64.98	100	64.98	78.78
fao780	59.52	100	59.52	74.63
Krapivin2009	80.71	100	80.71	89.33
SemEval2010	82.67	100	82.67	90.51
Nguyen2007	80.08	100	80.08	88.94

VI. CONCLUSION

This work introduces a novel unsupervised approach called KFA technique to analyze the keyphrase frequency from research articles: a region-based method. It is domain and language agnostic, requires little statistical knowledge, and does not rely on training data. The proposed approach begins with data acquisition and then pre-processing, then moves on to statistical methodologies, curve plotting analyses, and lastly the curve fitting procedure. The proposed techniques effectively analyze the keyphrase frequency of the articles based on region and produce a negative exponential formula shown in equation (6), indicating that most of the frequency of keyphrases is located in 1st region of articles, then 2nd region, and then so forth. Afterwards, the proposed technique was tested and evaluated on five different datasets and delivering 51.98% of the keyphrase frequency in 1st region, 7.90% in 2nd region, and so on, where, a total keyphrase frequency of 73.59% are present. The proposed approach also find the best dataset named "SemEval2010" with highest accuracy of 82.67% and the highest F1-score of 90.51% as well as it will improve the effectiveness of present keyphrase extraction methods significantly. We have a plan to design a robust key extraction algorithm with in future using the more statistical features introduced throughout this research. We're also working on a solution for the problem of missing keywords, which occurs when several manually assigned keywords aren't discovered in the text.

ACKNOWLEDGMENT

The authors are grateful to the Universiti Malaysia Pahang (UMP) for providing laboratory workspace and funding, as well as to the National Research Foundation of Korea (NRF).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] M. B. A. Miah, S. Awang, M. S. Azad, and M. M. Rahman, "Keyphrases concentrated area identification from academic articles as feature of keyphrase extraction: A new unsupervised approach," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[2] C. Sun, L. Hu, S. Li, T. Li, H. Li, and L. Chi, "A review of unsupervised keyphrase extraction methods using within-collection resources," *Symmetry*, vol. 12, no. 11, p. 1864, 2020.

[3] M. B. A. Miah, S. Awang, and M. S. Azad, "Region-based distance analysis of keyphrases: A new unsupervised method for extracting keyphrases feature from articles," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*. IEEE, 2021, pp. 124–129.

[4] N. S. M. Nafis and S. Awang, "An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification," *IEEE Access*, vol. 9, pp. 52 177–52 192, 2021.

[5] Y.-f. B. Wu, Q. Li, R. S. Bot, and X. Chen, "Domain-specific keyphrase extraction," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 283–284.

[6] U. Parida, M. Nayak, and A. K. Nayak, "Insight into diverse keyphrase extraction techniques from text documents," *Intelligent and cloud computing*, pp. 405–413, 2021.

[7] T. Tomokiyo and M. Hurst, "A language model approach to keyphrase extraction," in *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, 2003, pp. 33–40.

[8] G. Rabby, S. Azad, M. Mahmud, K. Z. Zamli, and M. M. Rahman, "Teket: a tree-based unsupervised keyphrase extraction technique," *Cognitive Computation*, pp. 1–23, 2020.

[9] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.

[10] K. Aggarwal, M. M. Mijwil, Sonia, A.-H. Al-Mistarehi, S. Alomari, M. Gök, A. M. Z. Alaabdin, and S. H. Abdulrhman, "Has the future started? the current growth of artificial intelligence, machine learning, and deep learning," *Iraqi Journal For Computer Science and Mathematics*, vol. 3, no. 1, p. 115–123, Jan. 2022.

[11] R. Qamar, N. Bajao, I. Suwarno, and F. A. Jokhio, "Survey on generative adversarial behavior in artificial neural tasks," *Iraqi Journal For Computer Science and Mathematics*, vol. 3, no. 2, p. 83–94, Mar. 2022.

[12] K. S. Hasan and V. Ng, "Automatic keyphrase extraction: A survey of the state of the art," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1262–1273.

[13] K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, and M. Jaggi, "Simple unsupervised keyphrase extraction using sentence embeddings," *arXiv preprint arXiv:1801.04470*, 2018.

[14] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[15] S. R. El-Beltagy and A. Rafea, "Kp-miner: A keyphrase extraction system for english and arabic documents," *Information Systems*, vol. 34, no. 1, pp. 132–144, 2009.

TABLE 4. Compare our proposed two approaches for KFA technique.

Articles Regions	Keyphrase frequency in 1st region (%)	Keyphrase frequency in 2nd region (%)	Keyphrase frequency in 3rd region (%)	Co-efficient of Negative Exponential ($m*e^{-nx}+r$)
Eight (8) Regions	51.98%	7.90%	4.12%	$m = 4.04$, $n = 2.09$, and $r = 0.021$
Sixteen (16) Regions	44.11%	7.87%	4.67%	$m = 2.47$, $n = 1.75$, and $r = 0.013$

[16] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "Yake! collection-independent automatic keyword extractor," in *European Conference on Information Retrieval*. Springer, 2018, pp. 806–810.

[17] X. Wan and J. Xiao, "Collabrank: towards a collaborative approach to single-document keyphrase extraction," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 2008, pp. 969–976.

[18] A. Bougouin, F. Boudin, and B. Daille, "Topicrank: Graph-based topic ranking for keyphrase extraction," in *International joint conference on natural language processing (IJCNLP)*, 2013, pp. 543–551.

[19] L. Sterckx, T. Demeester, J. Deleu, and C. Develder, "Topical word importance for fast keyphrase extraction," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 121–122.

[20] C. Florescu and C. Caragea, "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1105–1115.

[21] F. Boudin, "Unsupervised keyphrase extraction with multipartite graphs," *arXiv preprint arXiv:1803.08721*, 2018.

[22] M. B. A. Miah and M. A. Yousuf, "Detection of lung cancer from ct image using image processing and neural network," in *2015 International conference on electrical engineering and information communication technology (ICEEICT)*. ieee, 2015, pp. 1–6.

[23] J. Li, "A comparative study of keyword extraction algorithms for english texts," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 808–815, 2021.

[24] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "Kea: Practical automated keyphrase extraction," in *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*. IGI global, 2005, pp. 129–152.

[25] Ö. Ünlü and A. Çetin, "A survey on keyword and key phrase extraction with deep learning," in *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, 2019, pp. 1–6.

[26] Z. Alami Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: a survey and trends," *Journal of Intelligent Information Systems*, vol. 54, no. 2, pp. 391–424, 2020.

[27] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 216–223.

[28] E. Gopan, S. Rajesh, G. Vishnu, M. Thushara *et al.*, "Comparative study on different approaches in keyword extraction," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2020, pp. 70–74.

[29] P. L. L. Romary, "Automatic key term extraction from scientific articles in grobid," in *SemEval 2010 Workshop*, 2010, p. 4.

[30] M. Haddoud and S. Abdeddaïm, "Accurate keyphrase extraction by discriminating overlapping phrases," *Journal of Information Science*, vol. 40, no. 4, pp. 488–500, 2014.

[31] F. Bulgarov and C. Caragea, "A comparison of supervised keyphrase extraction models," in *Proceedings of the 24th international conference on World Wide Web*, 2015, pp. 13–14.

[32] F. Xie, X. Wu, and X. Zhu, "Efficient sequential pattern mining with wildcards for keyphrase extraction," *Knowledge-Based Systems*, vol. 115, pp. 27–39, 2017.

[33] J. Wang, J. Liu, and C. Wang, "Keyword extraction based on pagerank," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2007, pp. 857–864.

[34] M. Zhang, X. Li, S. Yue, and L. Yang, "An empirical study of textrank for keyword extraction," *IEEE Access*, vol. 8, pp. 178 849–178 858, 2020.

[35] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.

[36] N. Giarelis, N. Kanakaris, and N. Karacapilidis, "A comparative assessment of state-of-the-art methods for multilingual unsupervised keyphrase extraction," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2021, pp. 635–645.

[37] W. Zhuohao, W. Dong, and L. Qing, "Keyword extraction from scientific research projects based on srp-tf-idf," *Chinese Journal of Electronics*, vol. 30, no. 4, pp. 652–657, 2021.

[38] S. R. El-Beltagy and A. Rafea, "Kp-miner: Participation in semeval-2," in *Proceedings of the 5th international workshop on semantic evaluation*, 2010, pp. 190–193.

[39] R. Campos and V. Mangaravite, "Datasets of automatic keyphrase extraction," 2020. [Online]. Available: https://github.com/LIAAD/KeywordExtractor-Datasets

[40] M. A. Qureshi, M. Asif, M. F. Hassan, A. Abid, A. Kamal, S. Safdar, and R. Akber, "Sentiment analysis of reviews in natural language: Roman urdu as a case study," *IEEE Access*, vol. 10, pp. 24 945–24 954, 2022.

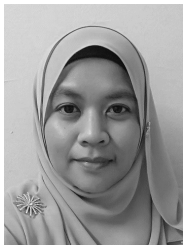
[41] O. Davydova, "Text preprocessing in python: Steps, tools, and examples," *Data Monsters https://es. wikipedia. org/wiki/Expresi% C3% B3n_regular*, 2019.

[42] M. Krapivin, A. Autaeu, and M. Marchese, "Large dataset for keyphrases extraction," 2009.

[43] S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin, "Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010, pp. 21–26.

[44] T. D. Nguyen and M.-Y. Kan, "Keyphrase extraction in scientific publications," in *International conference on Asian digital libraries*. Springer, 2007, pp. 317–326.

MOHAMMAD BADRUL ALAM MIAH received the B.Sc.(Engg.) and M.Sc.(Engg.) degree in Information and Communication Technology (ICT) from Mawlana Bhasani Science and Technology University (MBSTU), Bangladesh in 2007 and 2012 respectively. He has been serving as an Associate Professor in the Dept. of ICT, MBSTU. He is currently pursuing the Doctor of Philosophy (Ph.D.) degree in System Network and Security with the Faculty of Computing, Universiti Malaysia Pahang (UMP). He is the coauthor of more than 40 journals and conference papers. His research interests include Machine Learning, Data Mining, Text Mining, Neural Network, Computer Interfacing and Automation, Digital Signal Processing, Bio-Informatics, Network security, Computer Vision, Human Computer Interaction (HCI), Pattern Recognition and Image Processing.



SURYANTI AWANG received the Ph.D. degree in electrical engineering from Universiti Teknologi Malaysia, Johor Bahru, Malaysia, in 2014. She worked as a Research Officer with the Centre of Artificial Intelligence and Robotic (CAIRO), Universiti Teknologi Malaysia, from 2002 to 2005. She has been a Senior Lecturer (equivalent to an Assistant Professor) with the Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang, Malaysia, since 2005 until now. She is the coauthor for more than 30 journals and conference papers. Her research interests include pattern recognition, machine learning, and soft computing. She has collaborating with many industries in developing artificial intelligence systems. She receives numerous research grants from agencies, including a grant from Ministry of Higher Education of Malaysia under the Fundamental Research Grant Scheme with title "A New Feature Selection Technique for Text Classification." She had been awarded with Gold Medal in MTE'19 for Vehicle Type Recognition System, Silver Award in MTE'18, ITEX'18, and ITEX'17, for other projects.

MD. MUSTAFIZUR RAHMAN received the Ph.D. degree from the Department of Mechanical and Materials Engineering, Universiti Kebangsaan Malaysia, Malaysia. He served as the Dean of research with the Research and Innovation Department, Universiti Malaysia Pahang, Malaysia, where he has been working with the College of Engineering since April 2007. His research interests include artificial intelligence in mechanical systems, applied mechanics (fatigue and fracture), computational mechanics, advanced machining, optimization, finite element analysis, modeling of modern materials, internal combustion engine, and alternative fuels.

A. S. M. SANWAR HOSEN (M'22) received the M.S. Ph.D. degrees in Computer Science and Engineering, from Jeonbuk National University (JBNU), Jeonju, South Korea, in 2013 and 2017, respectively. He is currently an Assistant Professor with the division at JBNU. He has published several articles in journals and international conferences. He has been an expert reviewer for IEEE Transactions, Elsevier, Springer, and MDPI journals and magazines. He has also been invited to serve as the technical programme committee member in several reputed international conferences such as IEEE ACM. His recent research interests are focused on wireless sensor networks, internet of things, fog-cloud computing, cyber security, data distribution services, artificial intelligence, blockchain, and green IT.

IN-HO RA received the Ph.D. degree in computer engineering from Chung-Ang University, Seoul, South Korea, in 1995. He has been with the School of Computer, Information and Communication, Kunsan National University, where he is currently a Professor. From February 2007 to August 2008, he was a Visiting Scholar with the University of South Florida, Tampa, FL, USA. His major research interests include wireless ad hoc and sensor network, blockchain, IoT, PS-LTE, and microgrid.