# Keyphrases Concentrated Area Identification from Academic Articles as Feature of Keyphrase Extraction: A New Unsupervised Approach

**Article** *in* International Journal of Advanced Computer Science and Applications · January 2022

DOI: 10.14569/IJACSA.2022.0130192

Md Badrul Alam Miah
Mawlana Bhashani Science and Technology University

# Keyphrases Concentrated Area Identification from Academic Articles as Feature of Keyphrase Extraction: A New Unsupervised Approach

Mohammad Badrul Alam Miah[1]
Faculty of Computing
Universiti Malaysia Pahang, Pekan, Malaysia
Information and Communication Technology
Mawlana Bhashani Science and Technology University,
Tangail, Bangladesh

Suryanti Awang[2]
Faculty of Computing
Centre for Data Science & Artificial
Intelligence (Data Science Centre)
Soft Computing & Intelligent Systems
Universiti Malaysia Pahang, Pekan, Malaysia

Md. Saiful Azad[3]
Computer Science and Engineering
Green University of Bangladesh
Dhaka, Bangladesh

Md Mustafizur Rahman[4]
Department of Mechanical Engineering
Faculty of Engineering
Universiti Malaysia Pahang, Gambang, Kuantan, Malaysia

***Abstract*—The extraction of high-quality keywords and summarising documents at a high level has become more difficult in current research due to technological advancements and the exponential expansion of textual data and digital sources. Extracting high-quality keywords and summarising the documents at a high-level need to use features for the keyphrase extraction, becoming more popular. A new unsupervised keyphrase concentrated area (KCA) identification approach is proposed in this study as a feature of keyphrase extraction: corpus, domain and language independent; document length-free; utilized by both supervised and unsupervised techniques. In the proposed system, there are three phases: data pre-processing, data processing, and KCA identification. The system employs various text pre-processing methods before transferring the acquired datasets to the data processing step. The pre-processed data is subsequently used during the data processing step. The statistical approaches, curve plotting, and curve fitting technique are applied in the KCA identification step. The proposed system is then tested and evaluated using benchmark datasets collected from various sources. To demonstrate our proposed approach's effectiveness, merits, and significance, we compared it with other proposed techniques. The experimental results on eleven (11) datasets show that the proposed approach effectively recognizes the KCA from articles as well as significantly enhances the current keyphrase extraction methods based on various text sizes, languages, and domains.**

***Keywords*—Keyphrase concentrated area; KCA identification; feature extraction; data processing; keyphrase extraction; curve fitting***

## I. Introduction

The continuous development of the information age and exponential growth of textual information makes it even more challenging to handle this large amount of information [1]. Before the emergence of technology, this information could be processed by humans, which was very time-consuming. Furthermore, due to the inconsistencies between the amount of data and manual data processing skills, it is challenging to complete this vast information, leading to automated keyphrase extraction systems that utilise computers' extensive computational capability to substitute manual labour [2], [3].

The goal of automated keyword/keyphrase extraction techniques is to extract high-quality keys from documents. In general, Keyphrase offers a high level of description, summary, and characterization of documents, which is crucial for many aspects of Natural Language Processing, such as articles categorization, classification, and clustering [3]. They are, nevertheless, used in a wide range of Digital Information Processing applications, including Digital Content Management, Information Retrieval [3], [4], Contextual Advertising [5], and Recommender System [6]. It also offers a wide range of practical uses, including media searches, search engines, digital libraries, legal and geographic information retrieval [7].

Various keyphrase extraction methods have been developed to support the aforementioned applications [8], [9], [7], [10], [11], [12]. Domain-specific strategies [9], for example, need knowledge of the application domain, whereas linguistic approaches [9] demand language proficiency. They cannot solve problems in other disciplines or languages as a result. Supervised techniques need a lot of unusual train data to extract the quality keyphrases. Owing to their vast number of complicated operations, unsupervised machine learning methods are computationally costly, and they perform badly due to their inability to identify cohesiveness among several words that make up a keyword [7], [13], [14], [15]. Feature extraction is essential for those keyphrase extraction methods that want high-quality keyphrases. It's the process of obtaining characteristics (sometimes referred to as features) that distinguish keywords from other terms [16]. These features also impact the performance of various supervised and un-supervised keyword/keyphrase extraction methods. It is demonstrated that from the previous debate, the feature extraction of keyphrases remains an essential research topic for the study.

Therefore, this article proposes an unsupervised new Keyphrase Concentrated Area identification technique with ensuing significant contributions:

- The proposed technique, which is corpus-independent, can be applied to any text and any corpus.
- KCA identification's a domain- and language-agnostic method that relies on little statistical knowledge.
- The proposed method can be used as a keyphrase feature in both supervised and unsupervised approaches.
- It's a document length-free refers to the fact that there are no requirements for the minimum length of a document that a keyphrase must-have.
- Eleven datasets have been used to test and assess the effectiveness of the proposed method.

The remainder of this paper is organised as follows. Section II outlines the various methodologies, including their benefits and drawbacks, and so emphasises the need for a new strategy to be proposed. The suggested technique is then discussed in depth in Section III. The setup of the experiments is detailed in Section IV, which contains corpus data, evaluation measures, and implementation details. In Section V, all of the obtained findings are plotted and analysed, and Section VI brings this article to a close.

## II. Related Work

This section will discuss similar strategies because the proposed technique is a novel approach for extracting keyphrase features. Most keyphrase extraction techniques are categorized into two groups such as supervised and unsupervised, based on the training datasets [4]. Feature extraction is used in both ways. Below, we'll go over the main points of both of these groups' approaches.

### A. Supervised Methods

The keyphrase extraction technique is counted as a binary classification problem [1] using this method from articles, with a proportion of candidate keyphrases categorised as keyphrases and non-keyphrase. Methods for solving the classification problem include support vector machines, Decision trees, Naive Bayes [3], Neural networks [17], [18], and C4.5 [19]. The prominent techniques are examined in detail in the subsequence that adopts this method.

As a feature, Key Extraction Algorithm (KEA) [20] uses TFxIDF and the first presence location. It utilises descriptive approaches for identifying candidate keypresses, estimating feature values for each candidate and predicting and determining candidates' good keypresses using the Naive Bayes algorithm. However, KEA depends on the training dataset, and if the training dataset does not match the documents, it may produce poor results.

As a feature, Genitor Extractor (GenEx) [1] assigns first occurrence position, term frequency (TF), and keyphrase length. The most well-known key extraction approach is established on a collection of parametrized heuristic rules that employ genetic algorithms to retain their efficacy across diverse domains, and it is based on a C-4.5 decision-making process. It does not use the Term Frequency-Inverse Document Frequency technique (TF-IDF).

Unlike the GenEx and KEA methods, the Hulth system [1] allows the extracted keys to be as long as they want to be. The four characteristics it utilises are part of speech (POS) tag, n-grams, noun phrase (NP) chunks, first occurrence position, and TF. Unfortunately, no association exists between the various POS tag features. The system doesn't test on KEA or GenEx corpus, and the stated recall value is poor.

The Maui Algorithm [21], based on the KEA system, is an automatic generic topical indexing algorithm. It adds data from Wikipedia to expand the KEA system. However, one of this algorithm's flaws is its lack of assessment abilities.

The position of a term, its first occurrence; phrases; informativeness; keywords; and the length of the candidate term as a feature are all used by HUMB [22]. In a variety of data sets, the HUMB system has produced positive results. HUMB, on the other hand, has only used scientific papers.

The Document Phrase Maximality (DPM)-index, first position, TF, TFxIDF, IDF, first sentence, average sentence length, head frequency, substrings frequencies sum, and five other new features are (18 statistical features) used by DPM-index [23]. Without external knowledge or document structural elements, this system's results have improved significantly compared to other keyphrase extraction systems.

Citation-enhanced Keyphrase Extraction (CeKE) [24] utilize the following keyphrase features such as TFxIDF, relative Pos, inCited, POS, first position, inCiting, TF-IDF-Over, firstPosUnder, citation TF-IDF. They can improve keyphrase extraction and add keyphrase features. (CeKE + keys) the model outperforms other systems [1].

Keyphrase Extraction (KeyEx) Method [25] finds a large number of possible candidate keyphrases and build a classification model for key extraction using supervised learning methods. Experiments conducted by the author revealed that the KeyEx system has effectively improved the extracted keyphrase's quality. In addition, their strategy beats existing sequential pattern mining methods.

### B. Unsupervised Methods

The keyphrase extraction scheme is a ranking issue that is solved without prior knowledge. These methods can be classified as statistical or graph-based [1]. The following sections go over the most important techniques used by both groups in sufficient detail.

PageRank [26] is a graph-based algorithm that uses random walks as its foundation. It is, however, appropriate for raking web and social media pages but not for extracting keyphrase from formal documents. PageRank extension known as PositionRank [14] was discovered to improve performance, which scores word by taking into account all of its positions and its frequency, and thus determines its rank. This technique, however, poorly performs because it ignores topical coverage and diversity.

TextRank [27] uses Parts of Speech (POS) as an internal feature, with several limitations, including the inability to capture cohesiveness, resulting in sub-optimal results. TopicRank [28] is another keyphrase extraction technique that overcomes TextRank's limitations. The noun phrases in the document are extracted and clustered into topics by TopicRank. Furthermore, it has an issue with error propagation. The lengthening of TextRank is SingleRank [29]. It correctly pulls only noun phrases from the records, not keyphrases, by collecting ranked words. However, it does not always filter out low-scoring words and gives longer keys higher scores, but non-significant keys are included in the ranking process.

MultipartiteRank [15] is a technique for resolving the TopicRank error propagation problem. However, it suffers from clustering error, making selecting the most representative candidates challenging. Tree-based Keyphrase Extraction Technique (TeKET) [7] is a renowned unsupervised keyphrase extraction method that is language and domain-independent and needs only rudimentary statistical knowledge. Though it outperforms some other keyphrase extraction techniques, it has some disadvantages, such as tremendous flexibility.

The most common statistical method is named TF-IDF [30]. Although TF-IDF is simple to implement, computing Inverse Document Frequency (IDF) takes a long time and requires a lot of computing power when dealing with a large dataset. The KP-Miner [31] program is used to solve the problem of single-term preference. Although KP-Miner exceeds TF-IDF, it still has some drawbacks, including degrading the global ranking performance if the number of records increases. It's also computationally expensive because it relies on TF-IDF.

Yet Another Keyword Extractor (YAKE) [10] is another popular technique for removing the IDF problem by calculating the weighting score of a keyphrase using five features/attributes: as term position, casing, term relatedness to context, term frequency normalization, and term distinct sentence. However, because it uses the N-grams technique to generate candidate keys, its computational complexity grows linearly with N-grams.

According to the previous discussions, both supervised and unsupervised keyphrase extraction techniques have several drawbacks that prevent them from achieving better results. Therefore, this paper proposes a new unsupervised KCA identification technique as a keyphrases feature that will significantly decrease the specified flaws as well as extract high-quality keywords from academic articles.

## III. Methodology

The whole approach of keyphrase concentrated area identification utilizing the proposed method is divided into three major stages: $i$) Data preprocessing, $ii$) Data processing, and $iii$) KCA identification (see Fig. 1). In the subsequence sections, the proposed strategy is illustrated in detail.

### A. Data Pre-processing

It is an important stage in the development of our proposed technique. Initially, the proposed approach gathered eleven datasets (having 9006 papers) covering three languages (Portuguese, English, and Spanish), different disciplines (such as chemistry, physics, computer science, and others). Containing four different kinds of papers (news, abstracts, full articles, and M.Sc/Ph.D. Thesis) ranging from 75 tokens to 8000 tokens per document) [32]. Every dataset has two kinds of file names, like keys and docsutf8, including the same articles/documents. Visit Section IV-A for more information.

After that, the suggested method extracts the docsutf8 files (which include various vital articles as text files) as well as the keys files independently (containing different essential keys known as text files). Afterward, read these two files and save them respectively as document ($\delta$) and keys ($\chi$). After receiving the documents and keys, they must normalize the data, which entails four steps: Convert the document to lower case; Eliminate the irrelevant numbers by employing regular expressions); Remove all punctuation marks; Remove blank spaces (using the strip() function to remove leading and to end spaces) [33]. After that, The splitting technique is applied on keys files to compute the keyphrase learned as GoldKey ($\gamma$) founded on Newline (\n) method. At that moment, in our proposed approach, the length of text or document is split into ten (10) and twenty (20) regions.

### B. Data Processing

This is a crucial step after pre-processing the data. During this step, the proposed system uses the first appearance to locate (*Loc*) of each ($\gamma$) of ($\chi$) from the ($\delta$). Save the *Loc* of $\gamma$ in the proper region of the $\delta$ if located in the $\delta$. Note that the *Loc* is stored on two-dimensional (2D) array in which column is the ($\delta$) region's number and row is the ($\gamma$)'s number. If the $\gamma$ is not located, research the $\delta$ for the next $\gamma$ of $\chi$. This procedure will repeat until $\gamma$ has completed the $\chi$ file for a single dataset document. The same procedure will continue for all datasets.

### C. KCA Identification

It is an important and final phase after data processing. The output of the data processing phase is applied to this phase to find the concentration area of the keyphrases. This phase consists of the three significant steps: $i$) Average value calculation, $ii$) Curve plotting, and $iii$) Curve fitting technique that describes the following sections.

*a) Average Value Calculation:* To begin, for a single document/text, compute the Average (Avg) value of every region and save it in a new 2D array whose row is the number of records in a particular dataset and column is the text/document regions like as before. Afterwards, the process will resume until every document for a specific dataset has been completed. Calculate the average value of every region/portion for every record in a particular dataset and save this average value in another new 2D array whose row is the entire dataset and column is the same as before. After that, the Avg calculation will resume until every dataset has been completed [3]. Definitely, for all datasets, compute the Avg value of all regions.

*b) Curve Plotting (CP):* CP is a graphical presentation approach for a dataset. It's possible to read plotted values as known functions of unknown variables using this method. In data analysis and statistics it is pretty useful. CP is used to understand our proposed method's keyphrases concentration region/area. Because of this, the Avg value of each dataset is plotted alongside the Avg value of the whole dataset.

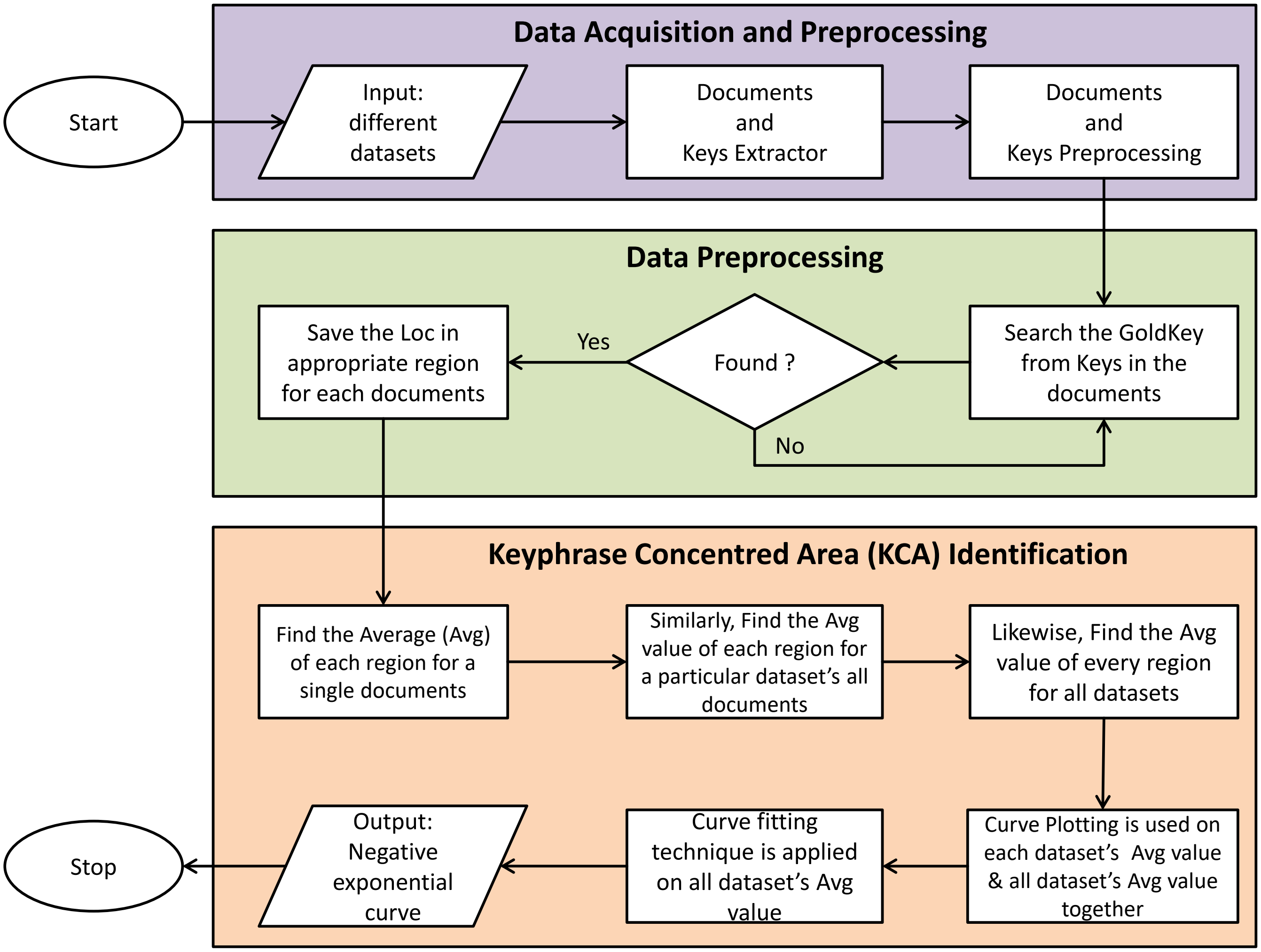
Data Acquisition and Preprocessing
Start
Input: different datasets
Documents and Keys Extractor
Documents and Keys Preprocessing
Data Preprocessing
Save the Loc in appropriate region for each documents
Yes
Found ?
No
Search the GoldKey from Keys in the documents
Keyphrase Concentred Area (KCA) Identification
Find the Average (Avg) of each region for a single documents
Similarly, Find the Avg value of each region for a particular dataset's all documents
Likewise, Find the Avg value of every region for all datasets
Curve Plotting is used on each dataset's Avg value & all dataset's Avg value together
Curve fitting technique is applied on all dataset's Avg value
Output: Negative exponential curve
Stop

Fig. 1. The Proposed Technique's Flow Diagram for KCA Identification.

*c) Curve Fitting Technique (CFT):* It is a helpful method for analysing linear, polynomial, and nonlinear curves. It is most likely the process of producing the best-fitting curve or mathematical function for a constrained set of data points. CFT is used to identify the critical concentration region/area in our proposed approach. As a result, CFT is applied on the Avg value of all datasets, resulting in a negative exponential curve for the proposed approach.

## IV. Experimental Setup

Our proposed method clearly stated that the experimental setting introduces corpus/dataset details, implementation details, and evaluation metrics, presented in the following section. Afterwards, the outcomes are explained in Section V.

### A. Corpus Details

our proposed approach has tested on 11 datasets/corpuses to evaluate the performance. How the proposed approach behaves under many datasets was our another ambition to understand. Standard gatherings such as Inspec [32], SemEval2010 [34], 110-PT-BN-KP [35], Nguyen2007 [36], PubMed [32], Schutz2008 [37], cacic [38], kdd [39], wicc [38], www [39], and theses100 [32] are used in our proposed approach. A quick summary is given in the preceding section III-A, and a statistical review of all datasets is given in Table I. Every corpus is explained in detail in the following sections.

**Inspec** [32] contains 2000 abstracts and 28220 gold keys from computer science articles published from 1998 to 2002. There are two sets of keywords in each document: controlled keywords, selected manually from the Inspec vocabulary, and uncontrolled keywords, which the editors liberally allocate.

**WWW** [39] and **KDD** [39] are the tiniest datasets (on an Avg of 84 and 75 tokens per document). The collection of those datasets (like Inspec) is based on abstracts of papers published between 2004 and 2014 at the ACM Conference and the World Wide Web(WWW) Conference on Knowledge Discovery in Databases (KDD). There are 1,330 and 755 documents in each and 6405 and 3093 goldkeys.

**SemEval2010** [34] is one of the famous standard datasets, which contains 244 whole scientific articles extracted from the ACM Library. The papers range in length from 6 to 8 pages

TABLE I. A STATISTICAL DATASETS SUMMARY FOR THE ANALYSIS OF PRESENT AND ABSENT GOLDKEYS

| Dataset | Language | Type of Doc | Domain | #Docs | #Gold Keys | #Present Goldkey | #Absent Goldkey | Absent Goldkey per doc(%) | Present Goldkey per doc(%) |
|---|---|---|---|---|---|---|---|---|---|
| 110-PT-BN-KP | PT | News | Misc. | 110 | 2688 | 2616 | 72 | 1.34% | 98.66% |
| Cacic | ES | Paper | Comp. Science | 888 | 3396 | 3057 | 339 | 10.44% | 89.56% |
| Inspec | EN | Abstract | Comp. Science | 2000 | 28220 | 12007 | 16213 | 55.98% | 44.02% |
| Kdd | EN | Paper | Comp. Science | 755 | 3093 | 1031 | 2062 | 65.78% | 34.22% |
| Nguyen2007 | EN | Paper | Comp. Science | 209 | 2507 | 2008 | 499 | 18.96% | 81.04% |
| PubMed | EN | Paper | Comp. Science | 500 | 7120 | 2513 | 4607 | 63.91% | 36.09% |
| Schutz2008 | EN | Paper | Comp. Science | 1231 | 55718 | 47387 | 8331 | 14.79% | 85.21% |
| SemEval2010 | EN | Paper | Comp. Science | 243 | 3785 | 3129 | 656 | 17.12% | 82.88% |
| Theses100 | EN | MSc/PhD Thesis | Misc. | 100 | 667 | 302 | 365 | 55.14% | 44.86% |
| Wicc | ES | Paper | Comp. Science | 1640 | 5860 | 5275 | 585 | 9.16% | 90.84% |
| WWW | EN | Paper | Comp. Science | 1330 | 6405 | 2122 | 4283 | 64.68% | 35.32% |

and cover four distinct areas of computer science: information search and retrieval, Distributed artificial intelligence, Distributed Systems, and Social and behavioural sciences. Every paper has a set of keyphrases assigned by the author as well as by professional editors.

**Nguyen2007** [36]: There are 209 scientific conference papers and 2507 gold keys in this dataset. Three articles were provided to student volunteers to read, and the goldkeys were handed out manually. Each document has twelve(12) goldkeys on Avg.

Both **Schutz2008** [37] and **PubMed** [32] are corpuses compiled from a PubMed Central full-text paper that cites over 26 million online books of life science journals from MIDLINE. Schutz2008 is made up of 1,231 articles chosen from PubMed Central, whereas PubMed is made up of 500 articles chosen from identical sources. The authors' Schutz2008 keyword is hidden in the paper and employed as goldkeys, yielding 45.26 goldkeys per document. The gold keyword in PubMed is Medical Subject Headings (MeSH), which is a controlled vocabulary glossary utilised to index articles, occurring in 14.24 goldkeys in each document.

**Theses100** [32] corpus comprises of hundred(100) complete Masters and PhD thesis from University of Waikato, New Zealand. These domains are relatively dissimilar, departing from computer science, chemistry, economics, philosophy, psychology, history, etc. It has 6.67 goldkeys per document, on Avg.

**110-PT-BN-KP** [35] is a Television(TV) Broadcast News(BN) corpus including 110 transcripts from eight(8) broadcast news programmes from the European Portuguese ALERT BN corpus, including finance, sports, politics, and others theme. Goldkeys were created by having a tagger remove all keywords that contained document content summaries, yielding 24.44 goldkeys per document.

**Cacic** [38] consists of 888 scientific publications from 2005 to 2013. It also comprises the minor number 3.82 goldkeys in each document, on Avg. The **Wicc** [38] dataset made up of 1,640 scientific papers published from 1999 to 2012, with an Avg of 3.57 goldkeys in each document.

TABLE II. CONFUSION MATRIX

| | Actual Positive Class | Actual Negative Class |
|---|---|---|
| Predicted Positive Class | $T_P$ | $F_N$ |
| Predicted Negative Class | $F_P$ | $T_N$ |

### B. Evaluation Metrics

*Accuracy*, *error rate*, *recall*, *precision*, $F_1$*-score*, and other significant and relevant metrics are routinely used to measure the performance of a system. To evaluate the performance of our proposed approach, we employ accuracy data and a confusion matrix (shown in Table II). The accuracy measure is generally defined as the percentage of correct predictions out of the total number of patterns analysed. The following equation (1) represents *accuracy*.

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{1}$$

Here, True Positive ($T_P$) and True Negative ($T_N$) denote the number of positive and negative keyphrases accurately classified, respectively. On the other hand, False Positive ($F_P$) and False Negative ($F_N$) represent the number of positive and negative keywords that were wrongly classified.

### C. Implementation Details

Python 3.6 and the Spyder-IDE are used to implement the proposed method. It is a high-level and object-oriented programming language that is easy to learn and utilise. It has a data structure that is user-friendly, versatile, and supported by numerous libraries. It increases productivity, is interpreted, dynamically typed, and is free and open-source. It is applied in big data, Cloud Computing, and Machine Learning, etc.

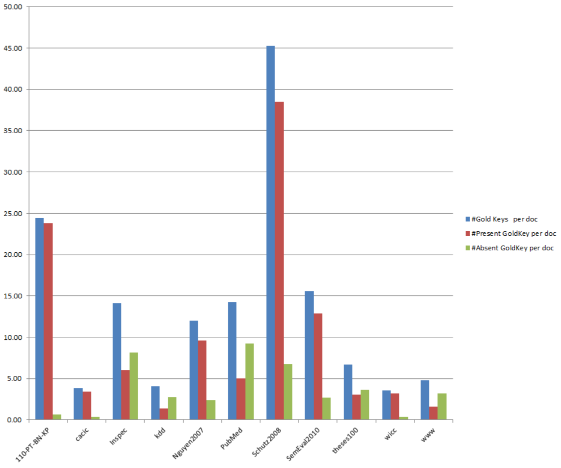

Fig. 2. The Avg Number of Goldkeys are Present and Absent in Each Document for All Datasets.

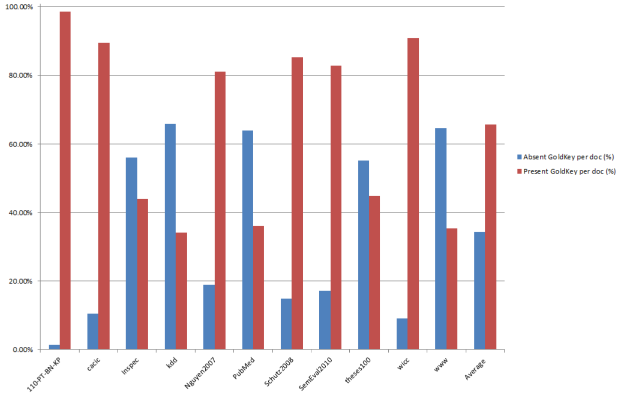

Fig. 3. The Avg Number of Goldkeys are Absent and Present in Percentage per Document for all Datasets.

Following that, the machine is outfitted with an Intel Core i7 processor, RAM-12GB, a SATA-connected solid state drive (SSD), and the Windows 10 operating system [3].

## V. Results and Discussion

This section includes a full examination of the experiment outcomes. The proposed system divides the text or documents length into twenty (20) and ten (10) regions to identify the Keyphrases Concentrated Area (KCA). When more than twenty regions are raised, the first region produces significantly less goldkey than twenty regions. Similarly, if the number of regions is lowered to less than 10, the first region has significantly more goldkey than ten regions. Our proposed technique aims to locate the KCA in documents/articles; thus, instead of expanding or lowering the regions, the system is examined for all types of text lengths as ten and twenty regions. This section is divided into two phases described in the following section: $i$) Result Analyses, and $ii$) Comparison of Proposed Systems.

### A. Results Analysis

The proposed system's performance is evaluated in this phase using the following criteria: i) Dataset Analysis, ii) Plotting Analysis, and iii) Curve Fitting Analysis, are the three types of results analysis.

*a) Dataset Analysis:* The proposed system has been tested on eleven (11) datasets (detail in section IV-A) to judge the performance of the proposed technique. Afterwards, the proposed system determines how many documents, number of goldkeys, present and absent goldkeys, as well as present and absent goldkeys in each article in (%) exist in every dataset provided in Table I based on the analysis of the datasets. The Avg number of goldkeys present and absent per document are examined for each dataset, exhibited in Fig. 2. Likewise, the Avg number of goldkeys absent and present in percentage(%) of each document for all datasets is displayed in Fig. 3. According to our findings, 65.70% of goldkeys per document are present on Avg across all datasets, while 34.30% are absent.

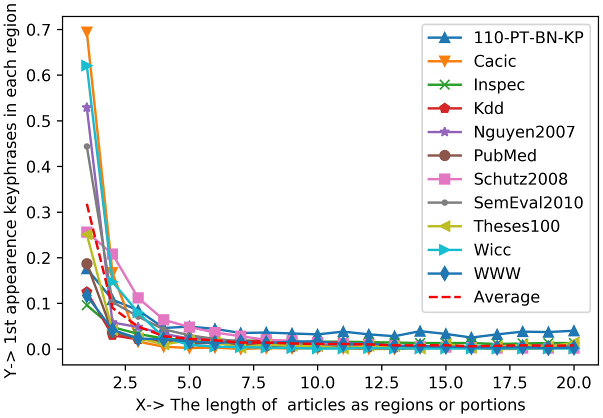

Fig. 4. The Plotting Analysis of KCA Identification by Considering 1st Appearance Keyphrases for 20 Regions.

*b) Plotting Analysis:* According to the previous discussion, Since the Avg of 65.70% of goldkeys is present per document for each dataset, all the results in this work have been predicated on 65.70% of present goldkeys. The first appearance keyphrases in a document are considered in our proposed method, and the text length is divided into twenty(20) and ten(10) regions. The proposed method then plots the eleven (11) dataset's values and Avg value of all datasets together based on each region of articles. Fig. 4 shows the analysis of first appearance keyphrases in each region for KCA identification when the text length is divided into twenty(20) regions. Similarly, Fig. 5 shows the analysis of first appearance keyphrases in each region for KCA identification when the text length is divided into ten(10) areas/regions. Since all dataset curves together are negative exponential, it is confirmed that the maximum goldkeys/keyphrases are found in 1st region, then 2nd region of the articles, and so forth, as shown in Fig. 4 and Fig. 5.

*c) Curve Fitting Analysis:* After completing the plotting analysis, the Avg value of entire datasets is applied in this analysis of our proposed system. Afterwards, the system attempts to discover the first fitted curve and then the negative exponential equation for each region's Avg value. In Fig. 6, the

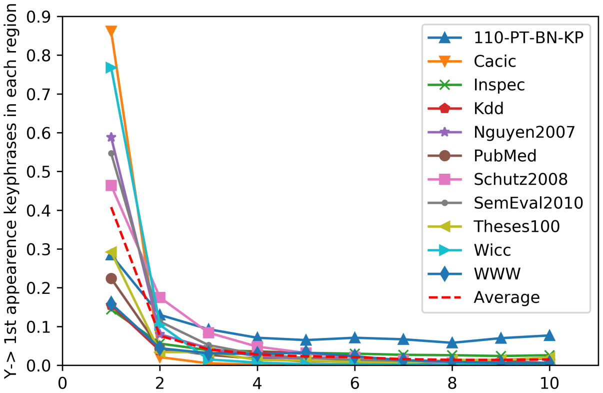

Fig. 5. The Plotting Analysis of KCA Identification by Considering 1st Appearance Keyphrases for 10 Regions.

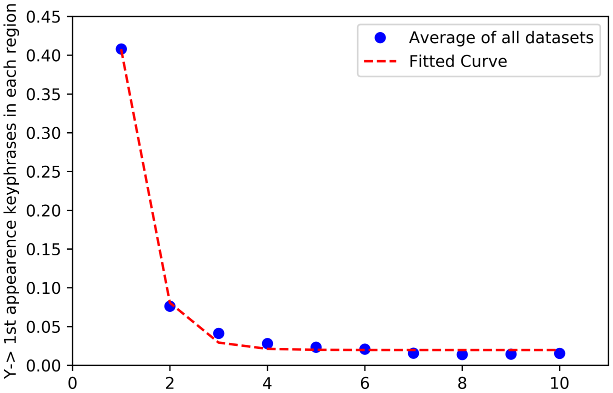

Fig. 7. The Curve Fitting Analysis of KCA Identification by Considering the Text Length as 10 Regions.

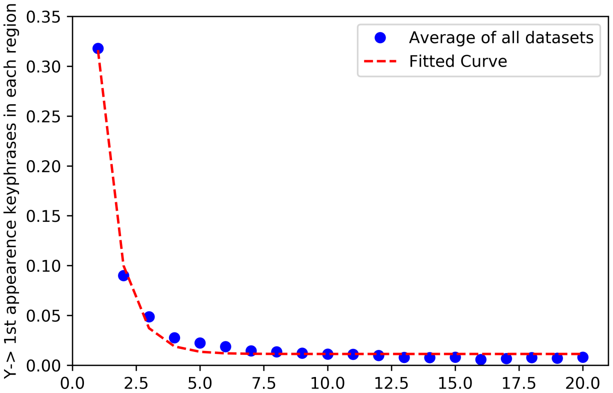

Fig. 6. The Curve Fitting Analysis of KCA Identification by Considering the Text Length as 20 Regions.

analysis of the curve fitting technique for KCA identification in each region is shown, with the text length divided into twenty (20) parts/regions, yielding the negative exponential equation expressed as follows (2) where $p = 1.05$, $q = 1.25$, and $r = 0.01$. Similarly, KCA identification from this analysis for the length of text as ten (10) regions or portion is displayed in Fig. 7 and also gives the similar equation which is negative exponential in where $p = 2.47$, $q = 1.85$, and $r = 0.02$. Since the fitted curves are found in negative exponential from the curve fitting analysis, It is demonstrated that most of the keyphrases are concentrated in the 1st portion of the documents, and next to the 2nd region of documents and so on, that are exhibited in Fig. 6 and Fig. 7.

$$y = p * e^{-qx} + r \tag{2}$$

### B. Comparison of Proposed Systems

Since KCA is a new technique with no existing policies, the proposed method does not compare with other techniques. The proposed system compares our two proposed approaches considering the length of the documents as ten (10) regions and twenty (20) regions for KCA identification shown in the following Table III. Both proposed systems are employed 11 datasets for comparison. From Table III, in ten (10) regions, more keyphrases concentrated in 1st region (62.09%) than twenty (20) regions (48.37%) of the documents/articles. Similarly, in ten (10) regions, more keyphrases concentrated in 1st two regions combine (73.70%) than twenty (20) regions (62.08%) of the documents/articles. Afterwards, the ten(10) regions approach provides more keyphrases concentration in the 1st three regions combined (79.97%) than twenty (20) regions (69.48%). Finally, we can say that our proposed technique for ten (10) regions provide more keyphrase concentration than twenty (20) regions in 1st regions, then 2nd region, and so on. The KCA in an article is proven from these two approaches.

## VI. Conclusion

The extraction of features for the keyphrase extraction approach has evolved into a critical component in a wide range of computer science applications. A new unsupervised approach termed Keyphrases Concentrated Area identification as feature of keyphrase extraction is presented in this paper. It is domain and language independent, needs little statistical expertise, and does not need the use of train data. The proposed technique starts with data pre-processing, processing, and KCA identification (average calculation, plotting analysis, and curve-fitting analysis).The proposed approach effectively recognises the KCA from texts/articles and creates a negative exponential equation, showing that the first region of the document/article contains more keyphrases than the rest of the articles.

In comparison to the suggested two techniques, the system tested on 11 datasets and produced a superior result based on the 65.70 per cent existing goldkey. Taking use of the more statistical elements discussed in this research, we want to develop a strong keyphrase extraction approach in the future. Moreover, when multiple manually specified keywords are not found in the page, there are some limitations in resolving the missing goldkeys/keywords issue.

TABLE III. COMPARE OUR PROPOSED TWO APPROACHES FOR KCA IDENTIFICATION

| Articles Regions | Keyphrase concentrated in 1st region(%) | Keyphrase concentrated in 1st two regions combine (%) | Keyphrase concentrated in 1st three regions combine (%) | Negative Exponential $(p{*}e^{-qx}{+}r)$ |
|---|---|---|---|---|
| Ten (10) Regions | 62.09% | 73.70% | 79.97% | p=2.47, q=1.85, r=0.02 |
| Twenty (20) Regions | 48.37% | 62.08% | 69.48% | p=1.05, q=1.25, r=0.01 |

## ACKNOWLEDGMENT

The authors are grateful to University Malaysia Pahang for giving laboratory space and funding under the University FLAGSHIP Research Grants programme (Project number RDU192210 and RDU192212).

## REFERENCES

[1] Z. A. 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.

[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] Z. A. Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: An overview of the state of the art," in *2016 4th IEEE international colloquium on information science and technology (CiSt)*. IEEE, 2016, pp. 306–313.

[5] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel, "A semantic approach to contextual advertising," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 559–566.

[6] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*. Springer, 2011, pp. 1–35.

[7] 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.

[8] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, and C. Zhang, "Sifrank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model," *IEEE Access*, vol. 8, pp. 10 896–10 906, 2020.

[9] E. Papagiannopoulou and G. Tsoumakas, "A review of keyphrase extraction," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 2, p. e1339, 2020.

[10] 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.

[11] T. Li, L. Hu, H. Li, C. Sun, S. Li, and L. Chi, "Triplerank: An unsupervised keyphrase extraction algorithm," *Knowledge-Based Systems*, vol. 219, p. 106846, 2021.

[12] T. Haarman, B. Zijlema, and M. Wiering, "Unsupervised keyphrase extraction for web pages," *Multimodal Technologies and Interaction*, vol. 3, no. 3, p. 58, 2019.

[13] 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.

[14] 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.

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

[16] 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.

[17] M. Al-Amin, M. B. Alam, and M. R. Mia, "Detection of cancerous and non-cancerous skin by using glcm matrix and neural network classifier," *International Journal of Computer Applications*, vol. 132, no. 8, p. 44, 2015.

[18] M. B. A. Miah, "A real time road sign recognition using neural network," *International Journal of Computer Applications*, vol. 114, no. 13, 2015.

[19] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on c4. 5 algorithm for online voltage stability assessment," *International Journal of Electrical Power & Energy Systems*, vol. 118, p. 105793, 2020.

[20] 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.

[21] O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 1318–1327.

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

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

[24] 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.

[25] 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.

[26] W. Souma, I. Vodenska, and L. Chitkushev, "Classification of paper values based on citation rank and pagerank," *Journal of Data and Information Science*, vol. 5, no. 3, p. 57, 2020.

[27] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, "Graph-based text summarization using modified textrank," in *Soft computing in data analytics*. Springer, 2019, pp. 137–146.

[28] 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.

[29] 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.

[30] L. Yao, Z. Pengzhou, and Z. Chi, "Research on news keyword extraction technology based on tf-idf and textrank," in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. IEEE, 2019, pp. 452–455.

[31] 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.

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

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

[34] 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.

[35] L. Marujo, A. Gershman, J. Carbonell, R. Frederking, and J. P. Neto, "Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization," *arXiv preprint arXiv:1306.4886*, 2013.

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

[37] A. T. Schutz *et al.*, "Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods," *M. App. Sc Thesis*, 2008.

[38] G. O. Aquino and L. C. Lanzarini, "Keyword identification in spanish documents using neural networks," *Journal of Computer Science & Technology*, vol. 15, 2015.

[39] S. D. Gollapalli and C. Caragea, "Extracting keyphrases from research papers using citation networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, 2014.