Keyphrase Distance Analysis Technique from News Articles as a Feature for Keyphrase Extraction: An Unsupervised Approach

Mohammad Badrul Alam Miah[1], Suryanti Awang[2]*, Md Mustafizur Rahman[3], A. S. M. Sanwar Hosen[4]*
Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah,
26600, Pekan, Pahang, Malaysia[1,2]
Information and Communication Technology,
Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh[1]
Center of Excellence for Artificial Intelligence & Data Science,
Universiti Malaysia Pahang Al-Sultan Abdullah, 26300, Gambang, Pahang, Malaysia[2]
Faculty of Mechanical and Automotive Engineering Technology,
Universiti Malaysia Pahang Al-Sultan Abdullah, 26600, Pekan, Pahang, Malaysia[3]
Department of Artificial Intelligence and Big Data, Woosong University, Daejeon 34606, South Korea[4]

***Abstract*—Due to the rapid expansion of information and online sources, automatic keyphrase extraction remains an important and challenging problem in the field of current study. The use of keyphrases is extremely beneficial for many tasks, including information retrieval (IR) systems and natural language processing (NLP). It is essential to extract the features of those keyphrases for extracting the most significant keyphrases as well as summarizing the texts to the highest standard. In order to analyze the distance between keyphrases in news articles as a feature of keyphrases, this research proposed a region-based unsupervised keyphrase distance analysis (KDA) technique. The proposed method is broken down into eight steps: gathering data, data preprocessing, data processing, searching keyphrases, distance calculation, averaging distance, curve plotting, and lastly, the curve fitting technique. The proposed approach begins by gathering two distinct datasets containing the news items, which are then used in the data preprocessing step, which makes use of a few preprocessing techniques. This preprocessed data is then employed in the data processing phase, where it is routed to the keyphrase searching, distance computation, and distance averaging phases. Finally, the curve fitting method is used after applying a curve plotting analysis. These two benchmark datasets are then used to evaluate and test the performance of the proposed approach. The proposed approach is then contrasted with different approaches to show how effective, advantageous, and significant it is. The results of the evaluation also proved that the proposed technique considerably improved the efficiency of keyphrase extraction techniques. It produces an F1-score value of 96.91% whereas its present keyphrases are 94.55%.**

Keywords*—Curve fitting technique; data pre-processing; data processing; feature extraction; KDA technique; keyphrase extraction

I. Introduction

In the past fifteen years, the paradigm for consuming news has changed from traditional print newspapers to individualized online news aggregation platforms like Google News, News360, and Yahoo! News. These systems gather a lot of news from many sources, aggregating it and presenting it on their respective mobile apps and websites [1], [2]. But now, the dramatic increase in textual news and the continual development of technology make it far more difficult to manage such a vast volume of news. People could just handle this enormous amount of news manually, which took a lot of time before technology was developed [3]. Developing automated keyword extraction techniques that replace manual tasks by utilizing the extraordinary computing power of computers is due to the difficulty of handling this huge amount of news [4], [5]. High-level keyphrases are extracted from news articles using automatic keyphrase extraction techniques. The keyphrase often offers a high level of document characterization, summarization, and description, which is important for numerous NLP features like content categorization, clustering, and segmentation [3]. However, they are utilized in a variety of online information processing applications, including contextual advertising, recommended systems, digital content management, and information retrieval. It is suitable for use in media searches, legal information retrieval, geographic information retrieval, search engines, and digital libraries [5].

To meet the aforementioned applications, a wide range of keyphrase extraction techniques have already been created, including [6], [7], [8], [9], [10]. Some of them, like domain-specific strategies [6], call for application domain knowledge; others, like linguistic techniques [9], [10], demand language competence. As a result, they are unable to solve issues in different fields, languages, or disciplines. According to [11], supervised machine learning algorithms need a large portion of datasets for training to extract high-quality keyphrases, and they generalize ineffectively beyond the range of trained data. Additionally, it made the system less understandable, required more storage and calculation, and was computationally expensive [12], [4], [13]. Due to the enormous number of complex processes, statistical unsupervised methods like [14], [15], [16] seem to be computationally very expensive. The graph-based unsupervised techniques underperform due to their inability to identify coherence between the many words that make up a keyword [17], [18], [19], [20], [21]. Last but not least,

*Corresponding authors. Email: suryanti@ump.edu.my, sanwar@wsu.ac.kr

TeKET [22] is incredibly flexible and behaves similarly to the TF-IDF if the data length is short.

Feature extraction is an essential technique in keyphrase extraction for those who want high-quality keyphrases. It is the process of acquiring characteristics (also referred to as features) that distinguish keyphrases or keywords from other terms [23]. These features also affect the effectiveness of various supervised and unsupervised keyphrase extraction methods. Keyphrase distance analysis (KDA) is important for all of those keyphrase extraction methods as a feature that helps to take top-level keyphrases from any article. The KDA technique is used as a feature of such keyphrase extraction techniques to extract keyphrases. Without applying high-quality features, the keyphrase extraction method cannot extract high-quality keyphrases [5]. As an outcome of the previous discussion, it has been determined that keyphrase feature extraction continues to be a crucial research field for the study. So, a region-based unsupervised KDA technique is proposed in this paper for news articles, which led to the following important achievements:

- The proposed technique called keyphrase distance analysis (KDA) introduced new features of keyphrases to calculate the distance of keyphrases from the center point of the news article.
- The proposed technique is corpus-independent, domain-independent, and language-independent.
- The proposed technique can be used by both unsupervised and supervised techniques.
- The proposed technique doesn't depend on document length means that it is a length-free technique.
- Two (2) standard news datasets are utilized to test as well as evaluate the performance of our proposed technique.

The remainder of this article is organized as follows: The various strategies are discussed in Section II along with their advantages and disadvantages, highlighting the demand for a fresh approach. Following that, a region-based unsupervised KDA technique is provided in Section III for figuring out the distance of keyphrases in each region of a news article. The phase of experiments is then thoroughly discussed in Section IV, including information about the datasets, evaluation metrics, and implementation details. The efficiency of the system was then assessed on two (2) standard datasets, and the suggested strategy was compared to existing methods to ascertain its advantages and disadvantages, which are discussed briefly in Section V. Lastly, in Section VI, the research's contributions, follow-up research, as well as flaws, were noted.

II. Related Works

The proposed keyphrase distance analysis (KDA) method from news articles can be used as an attribute or characteristic for keyphrase extraction methods [24], [25]. Therefore, similar strategies are covered in this section. There are two common types of keyphrase extraction techniques, depending on the training dataset. One is unsupervised, and the other is supervised. Both techniques can make use of feature extraction methods [3]. The essential components of both techniques will be covered in the following subsections:

A. Unsupervised Techniques

These techniques, which are categorized as statistical or graph-based techniques, are used to extract keyphrases from documents without any prior knowledge. They are thought to be a ranking issue [26]. PositionRank is a PageRank enhancement that enhances performance by combining word locations and frequency. However, it performs quite poorly, as evidenced by [20]. Another keyphrase extraction method that outperforms TextRank's constraints is TopicRank. Additionally, it has an issue of error propagation, according to [18]. TextRank's extension is SingleRank. Only noun phrases can be correctly extracted from a document. However, it is unable to accurately extract the keywords from the compiled ranked phrases [17]. A method called MultipartiteRank addresses the issue of topic rank, such as error propagation. There is an error in clustering [21]. TeKET is a more famous key extraction method that does not depend on a language or a domain. It requires only basic statistical knowledge. It offers a great deal of versatility, even though it performs better than others [22]. The KP-Miner is used to overcome the problem of preferring single phrases. Despite outperforming TF-IDF, it suffers a decline in performance on a worldwide scale. It is also expensive computationally [14], [27]. Another better technology that can solve the IDF issue is YAKE. For N-grams, however, the computing complexity rises linearly [16].

B. Supervised Techniques

The extracted keyphrases from any articles using supervised techniques are categorized in a binary fashion, with some candidate keyphrases being labeled as keyphrases and others as non-keyphrases [26]. The classification problem can be solved using a number of well-known methods, including support vector machines (SVM) [28], decision trees (DT), naive bayes [29], neural networks (NN) [23], [30], and so forth. The KEA utilizes the Naive Bayes technique, which uses TFxIDF with the first occurrence location as a feature. However, it may yield subpar results and rely on the training dataset [31]. First appearance position, length of keyphrase, and term frequency (TF) are all automatically taken into account by the Genitor Extractor (GenEx) as features. This system does not make use of the TFxIDF method [26]. The Maui algorithm extended the KEA technique to combine data from Wikipedia. Its primary flaw is a lack of assessment abilities [32]. Informingness, keyphraseness, length of candidate terms, beginning occurrence position, and term position are among the characteristics used by HUMB. Even though it has only used academic papers, it has had positive results across a variety of data sets [33], [25]. The first position of words, relative Pos, POS, keyphraseness, and TFxIDF are all features used by CeKE. It could make keyphrase extraction better [34]. The KeyEx Method significantly improves the quality of the retrieved key. Additionally, this method works better than other consecutive pattern mining methods [35].

The aforementioned discussions demonstrate that supervised and unsupervised keyphrase extraction approaches have a number of distinct disadvantages that prevent them from performing as well as they could. As a result, the proposed KDA technique will assist in minimizing the observed shortcomings and extracting high-quality keyphrases from news articles.

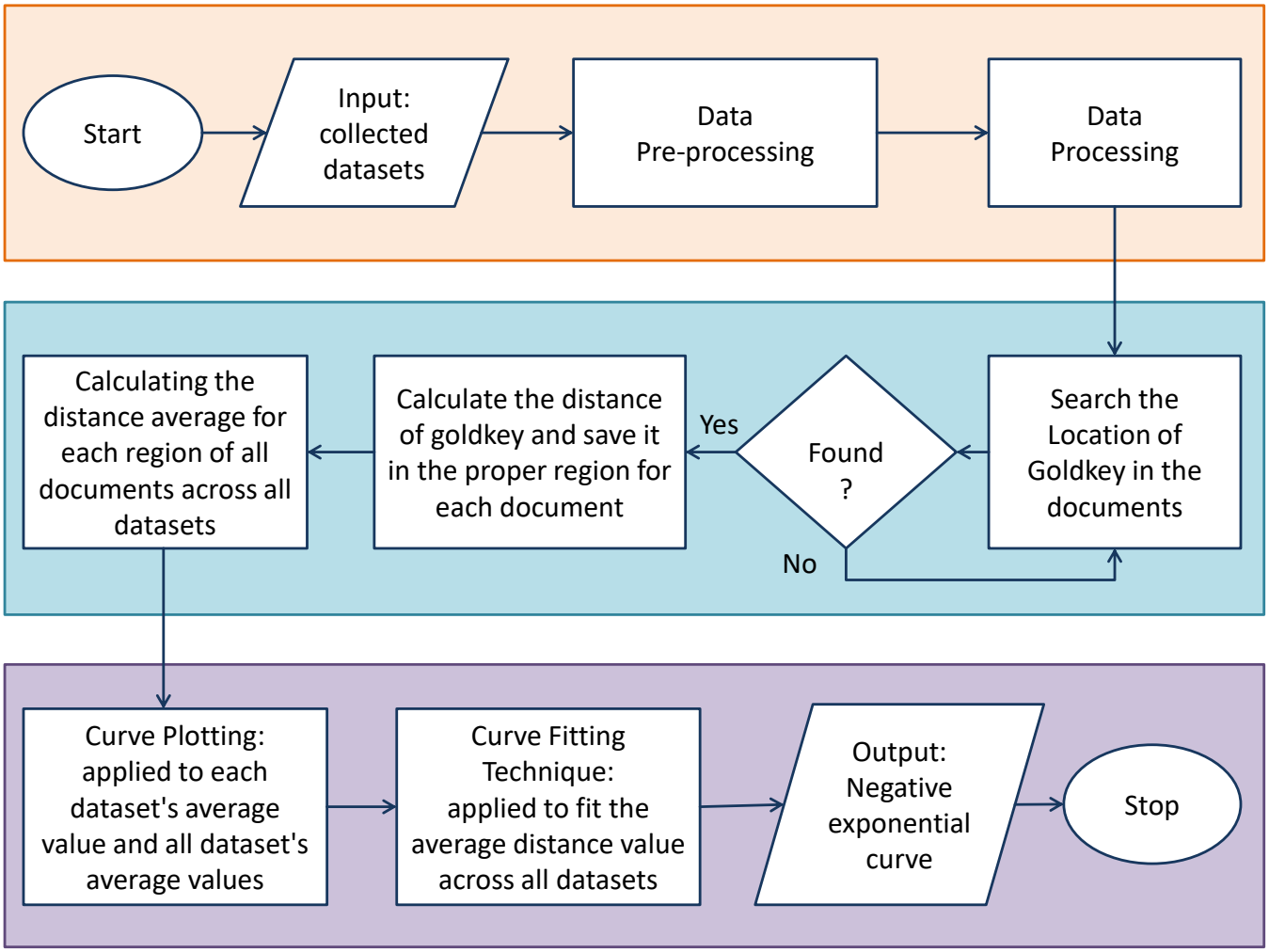

Fig. 1. The architectural flow diagram of proposed KDA technique.

III. Material and Methods

The eight steps of the proposed KDA technique are depicted in Fig. 1: data collection, data preprocessing, data processing, searching keyphrases or goldkeys, distance computation, averaging distance, curve plotting, and the last, the curve fitting technique. The next subsections explore the proposed approach in more detail.

A. Data Collection

The suggested technique gathered a total of two datasets that contained news in this phase. The two datasets are 110-PT-BN-KP and 500N-KPCrowd, which contain 610 news articles that cover the languages (English and Portuguese) [36]. Every dataset consists of two different types of files: the "keys file" (which contains keyphrases with the name "goldkeys") and the "documents file" (which has the name "docsutf8" and contains the articles/papers). In Section IV-A, the dataset is described in depth.

B. Data Pre-processing

After that, the data pre-processing phase receives the collected datasets. The suggested system then separates the key files and the docsutf8 files. The key files contain several critical keys, and the docsutf8 files contain numerous crucial documents or articles. The docsutf8 files and key files are then read, and they are saved as text files (named text and key, respectively). Afterward, normalization is required to apply to these files, which entails four procedures, such as text conversion to lowercase, removing all the punctuation marks from the text, eliminating empty or white spaces by using the strip() function (delaying all the leading and ending spaces), and removing digits (using regular expressions, non-relevant numbers are removed).

C. Data Processing

After the data pre-processing phase, the proposed technique tries to determine the total goldkeys depending on the Newline (\n) method from the key files. After that, the proposed approach uses the first appearance keyphrase, assumes the length of the document has eight regions, and then determines the midpoint of each document to calculate the distance of keyphrases from the midpoint.

D. Goldkey Searching

The proposed method then attempts to determine each goldkeys location *(Loc)* inside the document by taking into account their first occurrence. If the document has the location of Goldkey, it advances to the next phase. If the location of the goldkey cannot be found, try looking for the next goldkey in the document's key file. This procedure will continue until the key file (goldkeys) has been finished for a single document as well as for a single dataset. The exact same process will go on for all datasets.

E. Distance Calculation

In this phase, the proposed technique must compute the distance *(Dist)* between the document's midpoint and the goldkey location after the data searching step and preserve this *Dist* value in the relevant region of the document. Be aware that this value is preserved in a two-dimensional *(2-D*) array, where each column shows the number of document regions and each row shows the number of goldkeys [25].

F. Distance Averaging

After distance calculation, it is an important stage. This step involves calculating the distance average *(AVG)* between every region for a document and storing the results in a new *(2-D)* array with the same columns for article regions and rows for the total number of papers in the dataset. Then, until all documents for a particular dataset are finished, this procedure will continue. Similarly, for all documents in a dataset, calculate the distance average for every region in a manner similar to that described before, and then store the result in another *2-D* array, where the column represents the document region's number and the row indicates the total number of datasets. "Then, until all datasets have been completed, this AVG distance-calculating process will continue. Lastly, determine the AVG distance for each region across all datasets" [25].

G. Curve Plotting (CP)

The CP is a graphical method for representing the collected data. It effectively enables the creation of thoughts that do not emerge from a list of values and visually depicts the link between variables. In data statistics and analysis, CP is crucial. This technique is used to get the idea that keyphrase distance from the document's center depends on the article region in our proposed method. This is the justification for plotting the average or mean distance value of every dataset and the mean distance value for all datasets [25].

H. Curve Fitting Technique (CFT)

One of the most effective and often used analysis tools is curve fitting, which may be applied to linear, polygonal, and nonlinear curves. Most frequently, it involves creating a mathematical equation or curve that best fits a set of data points that are oriented toward limitations. The suggested approach uses CFT to generate a mathematical equation as well as a curve to determine the distance of keyphrases from the documents' centers and their density in each area. The CFT is then applied to the entire dataset's average value, and the suggested system gets a negative exponential curve as a result.

IV. Experimental Setup

In this section, the proposed technique explains the experimental setting, the details of the corpus, the evaluation metrics, and the details of how it will be used. The details of how it will be used are explained in more depth in the next subsection. This then anticipates the discussion of the findings in Section V.

A. Corpus Details

The performance of the suggested technique has been tested on two (2) different datasets. In this proposed system, typical collections like 500N-KPCrowd and 110-PT-BN-KP are used [36]. The datasets have been briefly described in the earlier Section III-A. A table that is shown in Table I explains the number of languages, categories of documents, document's number, goldkey's number, present goldkey's number, the number of present-and-absent goldkeys per document (%), and execution/processing time for each datasets [5], [2].

The dataset named *110-PT-BN-KP*, is a television (TV) broadcast news (BN) related dataset. The European Portuguese ALERT BN corpus contains 110 scripts from eight TV broadcast news programs. These programs cover a wide range of topics, such as banking, sports, and politics" [3]. All terms that made up text content summaries were removed using a tagger, yielding 24.44 goldkeys per document. There are 72 keyphrases that are missing and 2616 keyphrases that are present; processing took 0.047 seconds.

A dataset of broadcast news transcriptions is called the *500N-KPCrowd*. This dataset consists of 500 English-language broadcast news articles from ten different categories, each of which has 50 articles (art and culture, crime, fashion, business, health, world politics, politics, sports, science, and technology) [36]. Along with the processing time of 0.203 sec, it also includes the keyphrases 2265, which is absent, and 22345, which is the keyphrase that is present.

B. Evaluation Metrics

In our proposed method, the three most significant and relevant measures are used to contrast performance with alternative methods: *Precision, recall, and F1-score*. Here, *Precision* refers to the proportion of correctly predicted values to all positively predicted values. In other words, it is used to determine the positive patterns in a positive class that are successfully anticipated out of the overall projected patterns [5], [2]. The following equation (1) can be used to calculate it:

$$Precision = \frac{Key_{corrected}}{Key_{predicted}} \tag{1}$$

Where $Key_{corrected}$ is the total correctly predicted keyphrases that are matched with standard keyphrases and $Key_{predicted}$ is the total predicted keyphrases. Similarly, the ratio of precisely expected positive values to actual positive values is known as *Recall*, and it may be calculated using the equation (2):

$$Recall = \frac{Key_{corrected}}{Key_{standard}} \tag{2}$$

TABLE I. AN OVERVIEW OF THE NEWS DATASET USED TO ANALYZE THE PRESENT AND ABSENT KEYPHRASE / GOLDKEY WITH EXECUTION / PROCESSING TIME

News Dataset	Language Types	Total Docs	Total Gold-Keys	Total Present Goldkey	Present Goldkey per doc (%)	Absent Goldkey per doc (%)	Processing Time (sec)
110-PT-BN-KP	PT	110	2688	2616	98.66%	1.34%	0.047
500N-KPCrowd	EN	500	24610	22345	90.45%	9.55%	0.203

Where $Key_{standard}$ is the total keyphrases in the standard keyphrase list. Lastly, the weighted average of Precision and Recall is known as *F1-score*. The *F1-score* is calculated by utilizing the below equation (3).

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

C. Implementation Details

The suggested approach is put into practice using Python 3.6 and the Spyder IDE. Python is very simple and easy to use as well as learn. It's an object-oriented and high-level programming language. It offers a flexible data format that is user-friendly and backed by a number of libraries. It boosts productivity, is interpretive and dynamically typed, and is open-source and free. Python is used in many different fields, such as machine learning, big data, and cloud computing. The computer is outfitted with a 256GB SSD drive, 12GB of RAM, an Intel Core i7 processor, and Win-10 OS [2].

V. RESULTS AND DISCUSSION

This section discusses the in-depth study of the outcomes of the experiments. If the region's number is raised to more than eight in our proposed technique, the first region is found to have a relatively shorter AVG distance and fewer goldkeys. Likewise, if the region's number is decreased to less than eight, the first region has a bigger AVG distance and more goldkeys. The proposed technique therefore considers the text length to be eight regions rather than varying the number of regions because its goal is to analyze the keyphrase distance from news articles. In this section, the performance of the suggested method is examined under various headings (such as the Results of Dataset Analysis, the Results of Curve Plotting (CP) Analysis, the Results of Curve Fitting Technique, and finally Comparisons of the Suggested Method) that are described in the following subsections.

A. Result of Dataset Analysis

Two independent datasets (described in IV-A) were used to examine and assess the performance of the proposed approach. The proposed technique attempts to determine the processing time for each dataset from the dataset analysis. It also determines the total of documents, the total of goldkeys, the number of present-absent goldkeys, and the total present-absent keyphrases or goldkeys per document in percentage that are existent in each dataset, as shown in Table I. Based on this investigation, the proposed method takes an average of 0.13 seconds to process, has an AVG presence rate of 94.55% for keyphrases, and an AVG absence rate of 5.45% per document.

B. Result of Curve Plotting (CP) Analysis

According to the prior discussion, since the datasets contain an average of 94.55% of keyphrases/goldkeys per document that are actually present, all results in this work have been conducted based on that percentage of goldkey. The suggested method then tries to plot the average distance of each dataset and all datasets' average values together to represent the distance between each region by considering the first occurrence keyphrase and length of documents as eight (8) regions. The proposed KDA technique analyzes keyphrase distance and is depicted in Fig. 2. This study confirms that the first region of the document, followed by the second region, and so on, has the highest average distance and most frequent keyphrases. Because the plotted curves are negatively exponential, which can be seen in Fig. 2.

C. Result of Curve Fitting Technique (CFT)

Following the inspection of the plots, the CFT is adjusted to take into account the value of the average distance across all datasets. The suggested approach then seeks to identify the fitting curve as well as an exponential equation that is negative for that AVG distance value. Fig. 3 illustrates the analysis of the curve fitting technique in every region for the proposed KDA process while taking into account the eight regions for the documents and offers the negatively exponential curve and equation denoted in the following (4). Since this fitted curve is likewise negative exponential, as shown in Fig. 3, it is proven from this study that the highest number of keyphrases and the greatest distance are located in the 1st region of the document, then the second portion or region, and so forth.

$$y = b * e^{-cx} + d \tag{4}$$

where, $b = 12834.22$*,* $c = 1.59$*, and* $d = 339.30$. Lastly, the proposed system also attempted to demonstrate the region-based keyphrase distance analysis (KDA) from news articles using curve plotting analysis as well as curve-fitting analysis.

D. Comparison of Proposed Technique

This step involves two different kinds of comparisons, which are detailed in the next sub-subsection: comparisons to identify a superior dataset as well as comparisons to identify a superior model or technique.

1) Comparison to Identify a Superior Dataset: The suggested method measures the effectiveness of each dataset and identifies the best one, as shown in Table II, using evaluation criteria (such as precision, recall, and f1-score). From this

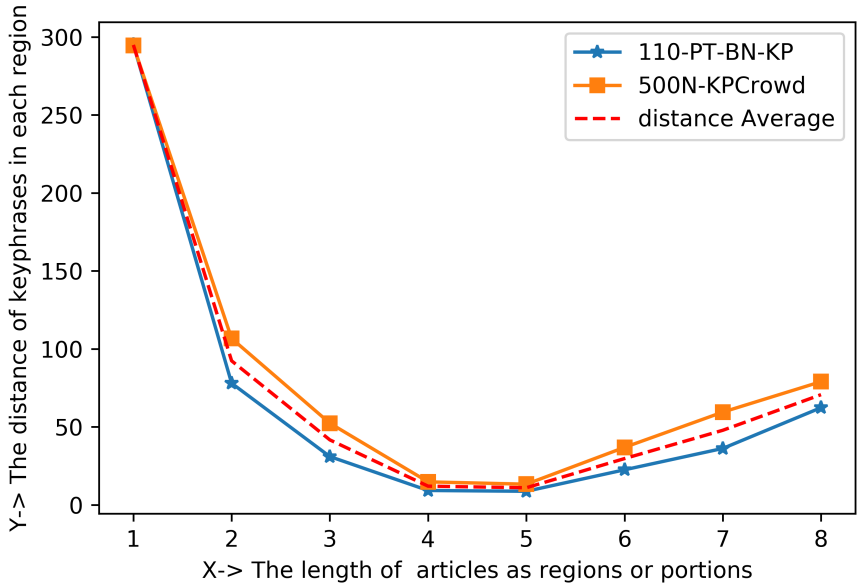

Fig. 2. The analysis of keyphrases distance by considering 1st occurrence and eight regions.

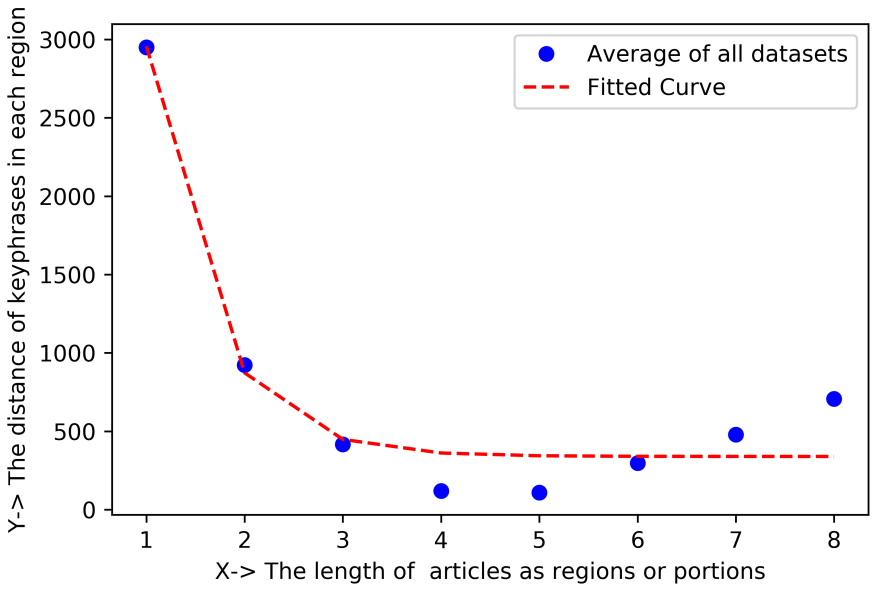

Fig. 3. The curve fitting analysis for the proposed KDA technique based on 8 regions.

table, the dataset named "110-PT-BN-KP" has a higher F1-score of 98.64% and a recall value of 97.32% than the "500N-KP-Crowd" dataset. It is indicated that the dataset "110-PT-BN-KP" performs better than the "500N-KP-Crowd" dataset.

2) Comparison to Identify a Superior Technique: As there is only one current method and the proposed KDA method is a novel method, it can be contrasted with the existing technique in this section. The comparison of our proposed KDA technique is exhibited in Table III. Based on Table III, the proposed KDA technique provides a higher recall value of 94.06% and a higher F1-score of 96.91% than the existing technique. Here, the proposed approach considers average performance measurements.

VI. Conclusion

In many computer science applications, feature extraction of keyphrases has now become crucial. This article proposes

TABLE II. THE DATASET'S PERFORMANCE COMPARISON FOR FINDING A BETTER ONE

Dataset	Average Performance Measurements		
Name	Precision	Recall	F1_Score
110-PT-BN-KP	100%	97.32%	98.64%
500N-KPCrowd	100%	90.80%	95.18%

TABLE III. COMPARISON OF OUR PROPOSED KDA TECHNIQUE

Existing	Average Performance Measurements		
Techniques	Precision	Recall	F1_Score
RDAK Technique [25]	100%	69.31%	80.09%
Proposed KDA Technique	100%	94.06%	96.91%

a region-based keyphrase distance analysis (KDA), a technique for automatic unsupervised feature extraction that is independent of domain and language and necessitates little statistical expertise and training data. Data collection, data preprocessing, data processing, searching keyphrases, distance computation, averaging distance, curve plotting, and lastly, the curve fitting technique, are the steps that make up the process of our proposed approach. After that, the KDA technique was tested and evaluated on two benchmark datasets to determine its effectiveness. Then it has produced a negative exponential equation and curve for the distance value, indicating that greater distance as well as more gold keys appear in the first region of documents. The proposed technique finally produced an accuracy/recall value of 94.06%, an F1-score of 96.91%, and a total present keyphrase of 94.55%. The proposed method also identifies *110-PT-BN-KP* as a better dataset, with a maximum recall value of 97.32% and the highest F1-score value of 98.64% than any other dataset. With the more statistical aspects given in this work, we intend to create a powerful keyphrase extraction technique in the future. The limitation of our research is that for the absent or missing keyphrase, the distance value is zero. For this reason, we are currently addressing the issue of missing goldkeys or keyphrases, such as when a large number of explicitly provided keyphrases are lacking from the document itself.

ACKNOWLEDGMENT

The authors are grateful to University Malaysia Pahang for giving laboratory space and funding under the University FLAGSHIP Research Grants program (Project number RDU192210 and RDU192212). The APC was fully funded by the Woosong University Academic Research Fund 2023, South Korea.

REFERENCES

[1] 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.

[2] M. B. A. Miah, S. Awang, M. M. Rahman, A. S. Hosen, and I.-H. Ra, "A new unsupervised technique to analyze the centroid and frequency of keyphrases from academic articles," *Electronics*, vol. 11, no. 17, p. 2773, 2022.

[3] 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.

[4] 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.

[5] M. B. A. Miah, S. Awang, M. M. Rahman, A. S. M. Sanwar Hosen, and I.-H. Ra, "Keyphrases frequency analysis from research articles: A region-based unsupervised novel approach," *IEEE Access*, vol. 10, pp. 1–1, 2022.

[6] 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.

[7] 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.

[8] F.-S. Alotaibi, S. Sharma, V. Gupta, and S. Gupta, "Keyphrase extraction using enhanced word and document embedding," *IETE Journal of Research*, pp. 1–13, 2022.

[9] 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.

[10] A. Dima and A. Massey, "Keyphrase extraction for technical language processing," *UMBC Faculty Collection*, 2021.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] L. Ajallouda, F. Z. Fagroud, A. Zellou, and E. B. Lahmar, "Kp-use: An unsupervised approach for key-phrases extraction from documents," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, 2022.

[16] 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.

[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] M. Garg and M. Kumar, "Kest: A graph-based keyphrase extraction technique for tweets summarization using markov decision process," *Expert Systems with Applications*, vol. 209, p. 118110, 2022.

[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] 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.

[23] 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.

[24] M. B. A. Miah and S. Awang, "Kda: An unsupervised approach for analyzing keyphrases distance from news articles as a feature of keyphrase extraction." in *The 6th National Conference for Postgraduate Research (NCON-PGR 2022)*. Universiti Malaysia Pahang, 2022, p. 83.

[25] 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.

[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] 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.

[28] 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.

[29] K. Sarkar, M. Nasipuri, and S. Ghose, "Machine learning based keyphrase extraction: Comparing decision trees, naïve bayes, and artificial neural networks." *JIPS*, vol. 8, no. 4, pp. 693–712, 2012.

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

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

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