Sriwijaya Journal of Informatic and Applications
Vol. 5, No. 1, February 2024, pp. 17-22
ISSN 2807-2391

Keyphrase Extraction Using TextRank for Indonesian Text

Fadel Muhammad [a,1], Novi Yusliani [b,2*], Muhammad Naufal Rachmatullah [b,3]

[a,b] Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia
[1] fadellmhmad@gmail.com; [2] novi_yusliani@unsri.ac.id; [3] naufalrachmatullah@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received 31 May 2023
Revised 11 February 2024
Accepted 20 February 2024

Keywords
Keyphrase Extraction
TextRank
Indonesian Text

ABSTRACT

Keywords are commonly used as a form of summary from scientific publications. But in determining keywords, it requires expertise in the related field and a long amount of time because you have to read and understand the entire contents of scientific publications. Keyphrase Extraction can be a solution to get relevant keywords in a short time based on titles and abstracts from scientific publications. TextRank method is used to extract keywords. This research will perform keyword extraction using the TextRank method for Indonesian text. The evaluation results of this study showed an accuracy value of 95.53% and an f1-score of 59.32% with a threshold configuration of 80% and using all keyword candidates.

1. Introduction

Keywords are usually used as a form of conclusion from scientific publications. Keywords can help in understanding the contents of scientific publications without the need to read them all. Keywords can represent information contained in text documents, such as scientific publications. Writers generally provide keywords that can describe the contents of their writing [1]. Keywords contained in scientific publications are made manually by reading and understanding the entire contents to draw a conclusion. This will take time if you want to create keywords in a large number of scientific publications. for this reason a keyword extraction system was built.

In building a keyword extraction system, there are approaches that can be used, including Supervised and Unsupervised approaches. The advantage of the Unsupervised approach is that this method does not require training, so this system can work well on various scientific publication topics [2].

In this study a keyword extraction system was developed using the TextRank algorithm. The advantage of this algorithm compared to other algorithms is that it uses an Unsupervised approach and only relies on sentences contained in scientific publications, so no in-depth knowledge of a language is required [3]. TextRank uses an Unsupervised approach so no training data is required and no input is required to process the actual data. TextRank is based on words only and requires no grammar knowledge. Currently, there are many implementations that use TextRank, because it is easy for developers to use it [4].

An automatic keyword extraction system is built with three stages, namely pre-processing, translation, and matching keyword candidates with a list of keywords. This experiment uses 33 article data taken from the PDII LIPI collection of journals. This experiment uses 3 weighting methods, namely TF, TFxIDF and WIDF. The best results were obtained from TFxIDF weighting. To improve the experimental results, researchers use Levensthein's algorithm to improve keyword results [5].

2. Literature Study

a. POS Tag

Part of Speech is a process of automatically assigning word class labels to a word in a sentence [6]. POS Tag is used to find out the noun of each word, such as a noun, verb, or adjective [7].

In improving the accuracy of POS Tagging, a study was conducted on several methods based on the hidden markov model (HMM) for Indonesian language texts. The first way is to use an affix tree which includes word endings and prefixes. The second way is to use HMM as a feature for POS Tagging. The third way is to use an additional lexicon (KBBI-Kateglo) to limit the tag candidates generated by the affix tree. The test results show that the best accuracy is 96.50% with 99.4% for words in the vocabulary and 80.4% for Out-of-Vocabulary (OOV) words. Experiments show that affix trees and additional lexicons are effective in improving POS Tagging accuracy, whereas successful use of POS Tagging does not provide much improvement in OOV handling [8].

b. TextRank

TextRank is a graph-based ranking algorithm. This algorithm is basically a way to determine the importance of vertices in a graph. Based on the information taken repeatedly from all graphs. The basic idea implemented by the graph-based ranking model is "votes" or "recommendations". When one node is connected to another, it is essentially providing support for the other node [9]. The textrank architecture can be seen in Fig. 1.

Fig. 1. TextRank Architecture

1. Pre-processing

   The pre-processing stages in this system consist of tokenization processes, stopwords filtering, POS tag and extract phrases.

2. Word Weighting

   PageRank will be used to calculate the weight of each word. In its application, sentences will be considered as graphs and words will be considered as nodes. The following is the PageRank formula.

$$S(V_i) = (1 - d) + d * \sum_{j \in In(v_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad (1)$$

3. Rank Keyphrase

   This process is carried out to rank the results of keyword candidates based on the weight of each value. Later the top keywords will be taken in accordance with the conditions of retrieval.

c. Confusion Matrix

Confusion Matrix is used to measure performance in the classification model. Perform calculations on the level of accuracy in data mining. The Confusion Matrix contains correctly predictable information on the classification system [10].

A confusion Matrix is used to measure the performance of the system that has been created. Several measurement units are used, namely Accuracy, Precision, Recall, and F1-score.

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN} \text{ x } 100\% \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \text{ x } 100\% \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \text{ x } 100\% \quad (4)$$

$$F1\text{-}score = \frac{2\,x\,Recall\,x\,Precision}{Recall+Precision} \quad (5)$$

3. Methodology

a. Data Collection Method

The type of data used in this study is primary data. The data used are 100 Indonesian language scientific publications with topics specifically in the field of computer science. The scientific publications used are publications that have been accredited by Sinta 2 and Sinta 3. Data were retrieved manually by downloading from 3 websites, namely jtiik.ub.ac.id, journal.mdp.ac.id, and journal.untan.ac. id. In the 100 publications, the title, abstract, and keywords will be taken which will be used as a dataset.

b. System Architecture

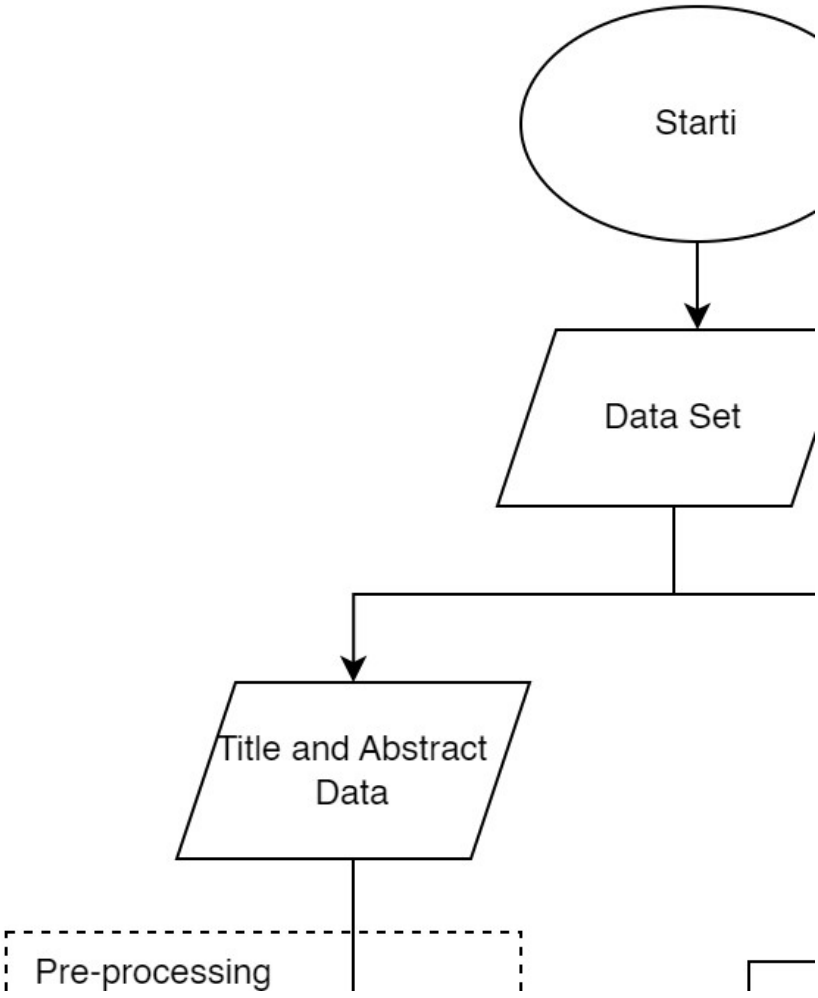

Fig. 2. Framework

Fig. 2. describes the system architecture which will be explained below.

1. Pre-processing

Text pre-processing in this study includes the process of Tokenizing, Stopwords Filtering, POS Tagging and Extract Phrase. The following is an overview of the Text Pre-processing process using scientific publication samples.

a. Tokenizing is the process of separating text into several tokens. In this study, the NLTK (Natural Language Toolkit) library was used to help carry out the Tokenizing process.
b. Stopwords Filtering is a process to reduce common words that don't have any important meaning.
c. POS Tagging is done to filter and select words according to certain tags. In this study the tags taken were NN (noun), JJ (adjective), NNP (Proper Noun) and NNG (Genitive Noun) and FW (foreign words).
d. Extract Phrase is a process to form key phrases. This word will be compared with the dataset to see the location of the adjoining words which will later be formed as key phrases.

2. Word Weighting is a process for determining the weight of the value of each word based on the closeness between words.
3. Rank Keyphrase is a process for sorting candidates from the highest score.
4. Keyphrase List is the final stage for getting keywords. Keywords are taken from the Candidate Keyphrase that has the highest value. The following is a sample to get 5 keywords.

4. Result

Tests were carried out using 100 scientific publication data. The process of obtaining keywords using TextRank is carried out according to the architecture that has been made. The process starts with data pre-processing, namely Tokenizing, Stopwords, POS Tags and forming keyword candidates. Furthermore, each keyword candidate will be given a value using TextRank. After the scoring process is complete, the candidates will be sorted based on the highest score. Keyphrases will be selected according to the Top Ranking level. The performance of selected keywords will be evaluated in two ways, namely taking keywords based on Top Ranking and using Cosine Similarity to compare keywords generated by the software with keywords provided in scientific publications.

1. Test Result Data Using Top Ranking

Table 1. Average measurement results for each decision provision

Rank Keyphrase	***Precision***	***Recall***	***F1-Score***	***Accuracy***
Top 5	0,128	0,1366	0,1303	0,9168
Top 10	0,107	0,2266	0,1425	0,8746
Top 50	0,0494	0,5249	0,0901	0,4912
Top 100	0,0355	0,6628	0,0682	0,1233

Table 1 is the average result of measurements using the Confusion Matrix in all top keyword retrieval using the entire dataset. The metrics measured include Precision, Recall, F1-Score and Accuracy.

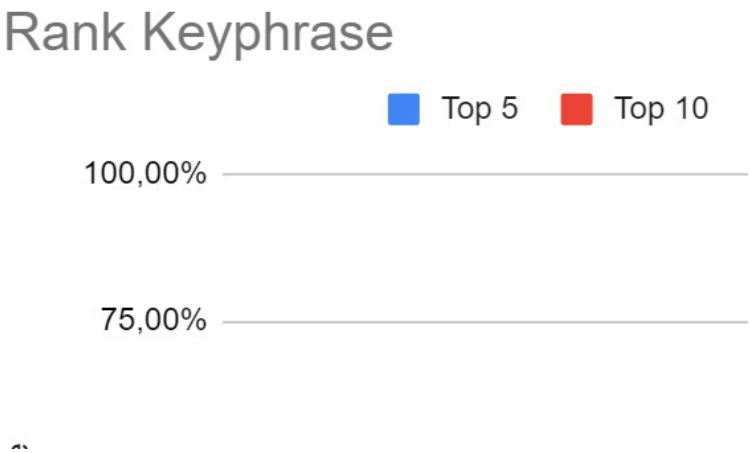

Fig. 3. Top Rank Evaluation

It can be seen in Fig. 3. that the best performance is in the Top 10. This conclusion was drawn because the Top 10 has a high Accuracy value and has a better F1-Score value compared to the others.

2. Test Result Data Using Cosine Similarity

Table 2. Average measurement results for each Threshold configuration

Threshold	*Precision*	*Recall*	*F1-Score*	*Accuracy*
40%	0,1534	0,6988	0,2409	0,7789
60%	0,2548	0,6988	0,3639	0,8854
80%	0,5612	0,6988	0,5932	0,9553

Table 2 is the average result of measurements using the Confusion Matrix for all Threshold configurations using the entire dataset. The metrics measured include Precision, Recall, F1-Score and Accuracy.

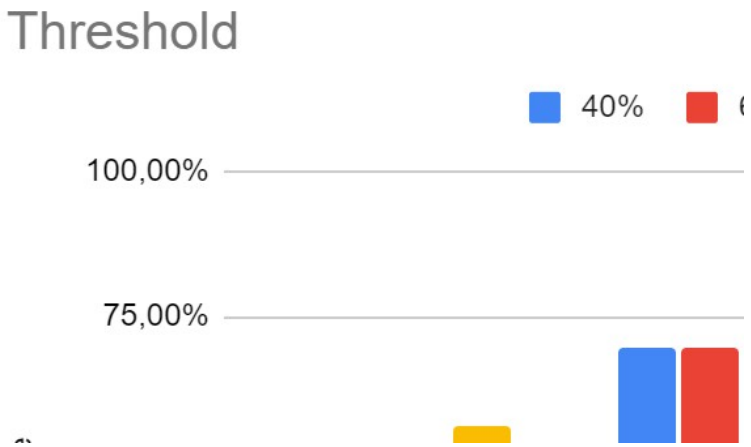

Fig. 4. Evaluation using Cosine Similarity

It can be seen in Fig. 4. that the best results are shown by using a Threshold of 80%. This is obtained because the higher the Threshold that is used will reduce the possibility of getting keywords that pass the Threshold. This will have an impact on decreasing the FP value (false positive) so that the resulting performance will be better.

5. Conclusion

The conclusion that can be drawn from this study is that Keyphrase Extraction using TextRank for Indonesian language text was successfully developed with the results of program testing based on Top Ranking which obtained an accuracy of 87.46% and an f1-score value of 14.25% and programs using Cosine Similarity obtained results accuracy of 95.53% and f1-score value of 59.32%.

References

[1] Asrori, R. B., Setyawan, R., & Muljono, M. (2020). Performance analysis graphbased keyphrase extraction in Indonesia scientific paper. *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, ISemantic 2020*, 185–190. https://doi.org/10.1109/iSemantic50169.2020.9234231.

[2] Mothe, J., Ramiandrisoa, F., Rasolomanana, M., Mothe, J., Ramiandrisoa, F., & Rasolomanana, M. (2020). *Automatic keyphrase extraction using graph-based methods To cite this version : HAL Id : hal-02640988.*

[3] Abdurrohman. (2018). *Evaluasi Algoritma Textrank pada peringkasan teks berbahasa indonesia.* 4–16.

[4] Bałcerzak, B., Jaworski, W., & Wierzbicki, A. (2014). Application of textrank algorithm for credibility assessment. *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014, 1,* 451–454. https://doi.org/10.1109/WI-IAT.2014.70.

[5] Diana Permata Sari; Ayu Purwarianti. (2014). Ekstraksi Kata Kunci Otomatis Untuk Dokumen Bahasa Indonesia Studi Kasus: Artikel Jurnal Ilmiah Koleksi Pdii Lipi. Baca: Jurnal Dokumentasi Dan Informasi, 35(2), 139–147. https://doi.org/http://dx.doi.org/10.14203/j.baca.v35i2.192

[6] Pramudita, H. R., Utami, E., & Amborowati, A. (2016). Pengaruh Part of Speech Tagging Berbasis Aturan dan Distribusi Probabilitas Maximum Entropy untuk Bahasa Jawa Krama. *Jurnal Buana Informatika,* 7(4), 235–244. https://doi.org/10.24002/jbi.v7i4.764.

[7] Dinakaramani, A., Rashel, F., Luthfi, A., & Manurung, R. (2014). Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus. *Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014*, 66–69. https://doi.org/10.1109/IALP.2014.6973519.

[8] Wicaksono, A. F., & Purwarianti, A. (2010). HMM based part-of-speech tagger for Bahasa Indonesia. In Fourth International MALINDO Workshop, Jakarta. Fourth International MALINDO …, June. https://www.researchgate.net/profile/Alfan-Farizki-Wicaksono/publication/209387036_HMM_Based_Part-of-Speech_Tagger_for_Bahasa_Indonesia/links/04d39ed44efdacf36e3af1a2/HMM-Based-Part-of-Speech-Tagger-for-Bahasa-Indonesia.pdf

[9] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004 - A Meeting of SIGDAT, a Special Interest Group of the ACL Held in Conjunction with ACL 2004, 85, 404–411.

[10] Gunawan, B., Pratiwi, H. S., & Pratama, E. E. (2018). Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *4*(2), 113. https://doi.org/10.26418/jp.v4i2.27526.