JOSEIT JOURNAL OF SYSTEMS ENGINEERING AND INFORMATION TECHNOLOGY
Published by: Indonesian Informatics Association
e-ISSN:2829-310X

Summarization and Classification of Sports News using Textrank and KNN

Falahah
Telkom University

falahah@telkomuniversity.ac.id

Abstract. The news summary process is critical in the news analysis process. However, there are frequently barriers to the summary process, such as the large number of news articles and the requirement for news classification. The goal of this study is to develop a news summary and categorization model that will be extremely valuable in the news analysis process. Textrank is the suggested summarizing approach, and KNN will be utilized for news classification. The resulting model can be used to automatically summarize and group news, making content analysis easier. Sports news will be used as the study object from July to August 2023, and the supervised category will be used to identify whether the news comprises sports news in three branches, soccer, badminton / tennis, or basketball. Classification is carried out using the KNN algorithm by training the model using 500 categorized news data. Modeling using k = 3 and k = 5 shows that the precision is around 0.9866 and 0.9666 respectively. The model's implementation on unknown text demonstrates that the model can properly predict text categories as long as the news content falls into the three specified categories, but fails for news content that does not fall into these categories.

Keywords: *sport news, summarization, classification, textrank, KNN*

Cite as: Falahah, "Summarization and Classification of Sports News using Textrank and KNN," Journal of Systems Engineering and Information Technology, vol. 3, no. 1, pp. 23–29, Mar. 2024.
DOI: 10.29207/joseit.v3i1.5706

Received by the Editor: 2024-02-03
Final Revision: 2024-03-05
Published: 2024-03-05

1. Introduction

As part of journalism work, news analysis is a crucial task. It is not sufficient to read only information that has been published in various media, including mainstream media. The ability to do news analysis is necessary for communication professionals so that the findings can be turned into a communication plan to present the information they wish to share with the audience. Because there is so much news nowadays, it is impossible to perform a news analysis work quickly unless the news items to be researched and the categories they must fit into are first organized. To aid in this analysis process, the news must be appropriately chosen, its content extracted, and then categorized. In this manner, news analysts can select relevant news more quickly based on their needs.

The availability of news summaries will be very beneficial for work that involves news analysis. Manually summarizing news requires time and is sometimes subjective, as journalists must determine what material is deemed significant. Therefore, AI-based text summarization will be very beneficial [1]. Text summarizing can aid in the creation of concise material, as well as the summation of more news, which can help journalists write news more effectively and increase the amount of knowledge that can be gained. However, the existence of automatic news summarization also raises concerns about affecting journalist integrity, the summarization algorithm being less accurate in capturing the context of the news, and the potential to distort the message the journalist wants to convey [2]. To address this, a news summary model must first be evaluated with sufficient training data, and in order to improve the accuracy of the news summaries produced, human editor validation is still required.

The responsibilities of a journalist extend beyond the mere summary of text, encompassing the ability to rapidly categorize news articles. This enables them to select the primary subject matter of the news, which can then be utilized as a topic for analysis or other forms of news writing content. The subsequent objective pertains to the process of classifying news articles into distinct categories. One of the current obstacles encountered by journalists involves the efficient acquisition of news from online media platforms, the prompt extraction of relevant information, and the adept categorization therein. The handling of this difficulty can be facilitated

through the utilization of natural language libraries, which exhibit a high degree of reliance on the specific language employed. However, the application of natural language processing (NLP) in the Indonesian context encounters various obstacles, with a notable problem being the scarcity of relevant corpus libraries.

This research aims to address the challenges outlined above by proposing a method that uses web scraping, text summarization, and text classification algorithms to solve the issues associated with the extraction and categorization of online news. In considering the problems outlined above, the issues that follow will be addressed by this research:

a. How to quickly acquire news from online news sites
b. How to write a news content summary
c. How to categorize news information based on its content

This issue will be addressed systematically in the article that follows, which will cover a basic understanding of web scraping techniques, summarization and categorization, model building techniques, model execution results, testing findings, analysis, and discussion regarding study findings.

2. Related Studies

2.1. Text Summarizing and Textrank

In general, the field of text summarizing can be classified into two main categories: extractive summarization and abstractive summary. Extractive summarization is a technique in which a subset of sentences is chosen from the original text to form the summary. Abstractive summarization involves the reorganization of language within a given text and, if required, the incorporation of new words or phrases in the resulting summary. The extractive summarization process is considered to be comparatively simpler than abstractive summarization. Furthermore, extractive summary evaluation can be performed using ROUGE metrics [3].

Previous works on extractive text summarization in news articles have been conducted by many researchers. Summarization is carried out using the following algorithms: Bert [4], Textrank [5], Fuzzy logic [6], LSA [7], TF-IDF [8], and many others. The summarization result is evaluated using ROUGE or bleu [9]. The textrank algorithm is frequently preferred among the available summarization algorithms due to its relatively lower computational complexity, enabling faster execution compared to BERT. Multiple previous research studies have demonstrated that the Textrank algorithm performs better compared to alternative algorithms [10] [11] [12] [13] [14]. Some results of textrank performance show the result about 17% in ROUGE-1 and 30% in ROUGE-2 [15].

TextRank is a supervised algorithm that is used to automatically summarize natural language text. It falls within the group of extractive summarization strategies, which are capable of taking the most crucial passages from the source text and using them to create a summary. Textrank uses a graph-based method in which words and sentences are viewed as graph vertices [16]. The graph-based ranking algorithm is a method to determine relevant nodes in a graph by recursively depicting global information from the entire network. The fundamental concept is to integrate voting or suggestions from the graph's nodes. When node A is connected to node B, node A votes on node B. The higher the rating number of a node, the more important the node. As a result, a node's score is decided by the votes it receives from other nodes, as well as the scores of the nodes that receive the votes.

Let **G** = (**V,E**) be a token graph formed from tokens (words), where **V** is a collection of vertices, and **E** is a set of (undirected) edges. Each word w_i is mapped to a vertex v_i, so the edge between two vertices has weight[16]:

$$e_{ij} = <\phi(w_i), \phi(w_j)> \qquad (1)$$

Where $<\phi(w_i), \phi(w_j)>$ denotes the cosine similarity between the (word) embedding of w_i and w_j (denotes by $\phi(w_i)$, and $\phi(w_j)$. $eij\ is$ usually set to zero if it is below a similarity threshold.

Unbiased Texrank, is the original textrank method, compute the score of v_i iteratively, for each vertex $i \in V$ as:

$$v_i = (1-d) + d \times \left(\sum_{j \in N(i)} \frac{e_{ij}}{\sum_{k \in N(j)} e_{kj}} \times v_j \right) \qquad (2)$$

with:

d : damping factor (value ranging from 0 to 1), but typically 0.85.
$N_{(i)}$: set of vertices which share and edge (with non-zero edge weight) with vertex

The value of ***d*** is used to integrate the probability of traversing from one random node in the network to another. This program applies the "random surfer model" to web browsing, in which the user randomly presses a link with probability ***d*** and jumps to a completely new page with probability 1-***d***.

2.2. Automatic Categorization

Automatically categorizing news articles can employ many classification approaches, among which the K-Nearest Neighbors (KNN) algorithm is an option [17]. Several previous studies have indicated that the precision of the K nearest neighbors (KNN) classification exceeds that of alternative classification techniques, including support vector machines (SVM), logistic regression, and random forests (RF) [18] [19]. The k-nearest neighbor (k-NN or KNN) technique is a method of classifying things based on learning data that are closest to the item [20].

The principle of learning by analogy underpins K-Nearest Neighbor. N-dimensional numeric attributes are used to characterize learning data. Each learning data set represents a point in the n-dimensional space, indicated by c. When a data query with an unknown label is entered, K-Nearest Neighbor will search for k pieces of learning data that are nearest to the query data in n-dimensional space. The distance between the query data and the learning data is obtained by measuring the distance between the query data point and all the learning data points using the Euclidean distance formula. The Euclidean distance is most commonly used to calculate the distance. Euclidean distance functions are used to test a measure that can be used to interpret the closeness of two objects, which is represented as follows [19]:

$$dist = \sum_{i=1}^{p} \sqrt{(x2 - x1)\text{^}2} \qquad (3)$$

With:

dist	:	Distance
*x*1	:	Training Data
*x*2	:	Testing Data
i	:	Variable data
p	:	Number of attributes

3. Method

The data used in this research source are sports news data gathered from ***detik.sport*** media (https://sport.detik.com/). This medium was chosen since the material provided has been labeled and there is a lot of language in the news. Several other media outlets frequently broadcast news that is too brief; therefore, the benefits of the summarizing process are insignificant.

The data on the media is then processed using Python programming, which makes use of a number of libraries such as:

a. Beautiful soup: webs crapping library for python
b. PySastrawi: stopwords library for Bahasa Indonesia
c. NLTK: NLP library for tokenizing, parsing, stemming and tagging.
d. Request-html: web scrapping library for Python

All of the tools listed above are operational in the Google Colab environment. As seen in Figure 1, there are several steps involved in creating and implementing a model. The first action is web scraping. We will perform some preprocessing on the text, such as tokenizing and eliminating stopwords, and then apply textrank for text summarization after extracting some content from the news. After summarizing, an output for text classification is produced. The text classification model that was produced was then used to test news articles from unclassified sources.

3.1 Web Scrapping

The dataset used to train the model includes news text data summarized from the sport.detik.com website and labeled in three categories: soccer, basketball, and rackets. Each category received 200 news links. The links were then extracted for text, and news that did not have significant text elements (for example, photos/videos) were deleted, yielding a total of 529 connections. 500 news stories were chosen at random from 529 links as input for the model learning procedure. In general, the web scraping process is shown in the picture below (Figure 2).

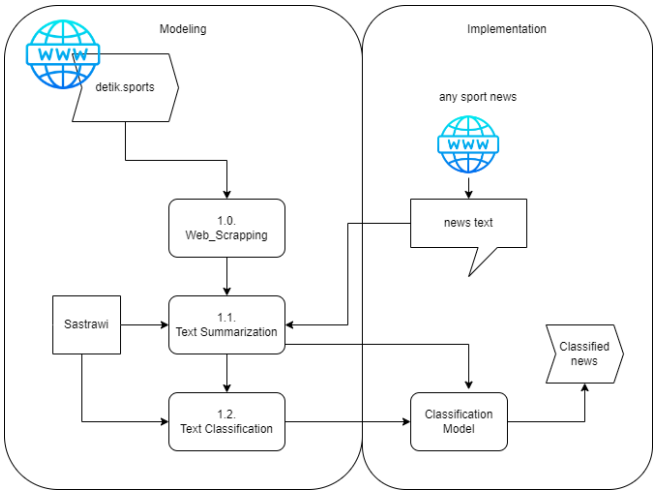

Figure 1. Step in News Summarization and Classification

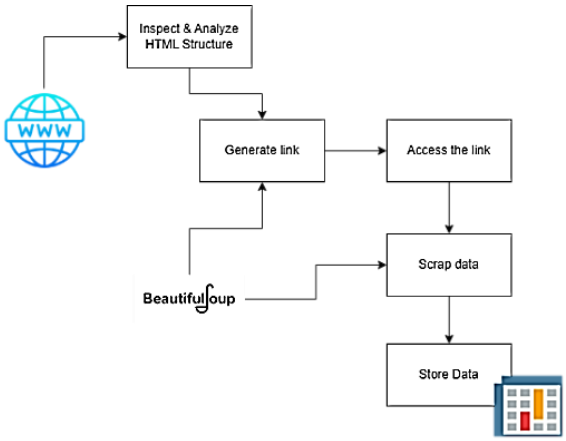

Figure 2. Web Scrapping Process

3.2. Preprocessing and text Summarization

To use textrank for text summarization, take the following step:

a. Take-Out Articles.
b. Convert articles into phrases: Eliminating stopwords, cleaning, and tokenizing.
c. Produce word embeddings or vector representations.
d. Compute the vector similarity matrix.
e. Use the similarity matrix to create a graph with sentences as vertices and similarity scores as edges.
f. The ranking of sentences.
g. Create a synopsis

3.3. Text Classification

Classification of news is done using KNN as follow:

a. Input: text summary
b. Label: category (with soccer, rackets, and basketball were then assigned numbers one through three)
c. Process:
    1) Convert category to number
    2) Vectorizer using TfIdf (using stopwords = Sastrawi)
    3) Split (train, test)
    4) KNN (k, text)

4. Result and Discussion

4.1. Acquire News through Web Scrapping

To solve the first issue and acquire a lot of news content in a short amount of time, web scraping techniques will be used. However, the structure of the targeted web page must be taken into consideration when implementing web scraping. A web page can be easily scraped if it follows a consistent structure, makes use of obvious tags or classifications, and does not obstruct the scraping process. Because of the well-organized structure of the Web pages, it is easy to extract content from the (sport.detik.com) website.

Two steps are involved in web scraping, specifically:

a. Search for news links within a specified time frame, then select three categories, football, rackets, and basketball, based on tags.
b. Remove any links that do not have text news (such as those that only have brief descriptions and images or videos). Links are not included because the news is contained in the picture or video rather than the text.

Here are some sample links that include:

a. Video:https://20.detik.com/detikupdate/20231012-231012048/wayne-rooney-resmi-jadi-pelatih-birmingham-city
b. Photo: https://sport.detik.com/sepakbola/foto-sepakbola/d-6976162/foto-jude-bellingham-seperti-michael-jordan-lihat-saja-lidahnya

Figure 3 show the contents from the sample links above.

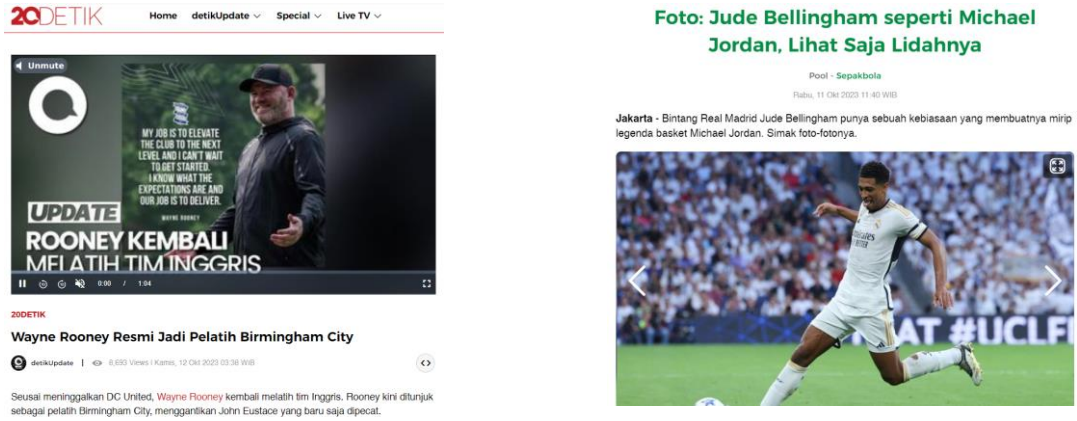

Figure 3. Content of Photo/Video Links

c. Extract the news content and remove any unnecessary text (news content could be repeated links that are read as text, for instance).

The Beautiful Soup library is used to perform web scraping and the output is saved as a CSV file. It has four columns: link, title, category, and content. Through web scraping, the journalist can easily and quickly obtain content.

4.2. Text Summarization

After 500 news stories are collected, the dataset becomes an input for text summary using textrank. Table 1 shows a sample of extract and summary results. As shown by the sample result, the summarize procedure can decrease the total amount of text and select the most important terms without impacting the main point of the news. We implement ROUGE-1 and ROUGE-2, for 10 data tests, and the result indicates the average values for ROUGE-1 and ROUGE-2 are 0.792 and 0.67 respectively.

Table 1. Sample of extract and summary result

Attribute	Data
Link	https://sport.detik.com/sepakbola/liga-inggris/d-6973922/mac-allister-blunder-van-dijk-salahku-juga
Title	Mac Allister Blunder, Van Dijk: Salahku Juga
Category	Sepakbola
Content	Virgil van Dijk pasang badan untuk Alexis Mac Allister usai blunder saat Liverpool ditahan Brighton & Hove Albion. Van Dijk mengaku salah juga. Mac Allister bertanggung jawab pada gol pertama Brighton di Amex Stadium, Minggu (8/10/2023) kemarin, oleh Simon Adingra. Gelandang internasional Argentina itu lengah saat menerima umpan dari Van Dijk sehingga direbut lawan. Liverpool sendiri sempat membalas dengan dua gol Mohamed Salah. Namun gol Lewis Dunk pada babak kedua memastikan kedua tim berbagi poin. Mac Allister telah mengakui kesalahannya, tak cermat

Attribute	Data
	dalam mengamati posisi pemain. Namun Kapten Liverpool Virgil van Dijk juga mengakui bahwa umpannya terlalu berisiko."Pastinya itu bukan cuma kesalahannya. Itu kesalahan saya juga," kata Van Dijk dikutip Metro."Itu bola berisiko dan kami mencoba bermain dari belakang. Hal-hal semacam ini terjadi dan Anda dihukum. Cara kami bangki itu reaksi yang bagus," imbuhnya. Dengan hasil imbang itu, Liverpool gagal menang di dua partai terakhir. Sebelumnya mereka kalah 1-2 dari Tottenham Hotspur, dalam laga yang diwarnai pembatalan gol Luis Diaz secara keliru.
Summary	Virgil van Dijk pasang badan untuk Alexis Mac Allister usai blunder saat Liverpool ditahan Brighton & Hove Albion. Van Dijk mengaku salah juga. Mac Allister bertanggung jawab pada gol pertama Brighton di Amex Stadium, Minggu (8/10/2023) kemarin, oleh Simon Adingra. Gelandang internasional Argentina itu lengah saat menerima umpan dari Van Dijk sehingga direbut lawan. Liverpool sendiri sempat membalas dengan dua gol Mohamed Salah. Mac Allister telah mengakui kesalahannya, tak cermat dalam mengamati posisi pemain. Namun Kapten Liverpool Virgil van Dijk juga mengakui bahwa umpannya terlalu berisiko.

4.2. Text Classification

The classification algorithm will then be put to use. To accomplish this purpose, the results and summary of the dataset were generated at random, resulting in a list of random categories in the dataset. The confusion matrix and the accuracy score are used to evaluate the model's performance. In this study, trials were conducted using two KNN values, 3 and 5. At k = 3, the accuracy level was = 0.98666, while at k = 5, the accuracy level was = 0.9666. As a result, k = 3 will be used as the starting point for the modeling. The generated model is then used to test or predict text with an unknown category. As test data, the following scenarios will be used:

a. comes from the same source (detik.sports);
b. comes from different sources (Kompas and CNN).

Table 2 shows the outcome of the test scenario. As we can see from the results above, the model can properly predict three specified categories, but it does not forecast news that was not included in the data set collection and labeling. As a result, all news that does not fit into one of the three categories listed above will be assigned to the last category (basket). Because of this weakness in supervised learning, we must create new categories or prepare a large number of existing ones to contain all text that was not included in the previous category. This will increase accuracy. Since it is possible that the text in the new category in the above dataset is sports news, but from a different sport entirely, it cannot be labeled "not sports news." This suggests that if the variety of knowledge areas is very large, categorization in one knowledge domain will be challenging. Therefore, a binary model—appropriate and improper, for instance, hoaxes and non-hoaxes, sports and non-sports—is typically the categorization that is applied to KNN the most.

The study's findings demonstrate how strongly the news context influences the text extraction and classification process. In one news setting, a word will mean one thing and something else, and also depends on the language used. For example, the word "*mencukur*" in the Bahasa Indonesia soccer sports news will signify something different in the crime news. Similar terms that appear frequently in sports news include "*merumput*", *"servis", "offside", "tarkam", "nyambi", "buka-bukaan*", etc. Thus, it is crucial that professionals define these key terms as words that have a significant meaning. This has the potential to improve the accuracy of news classification and summarizing results.

The result also implies that we should expand KNN approaches in the future by adding complex algorithms or clustering techniques. Language itself is the source of the difficulty in classifying texts; therefore, language libraries are essential for improving the accuracy of the classification model.

Table 2. Result of Unknown Text Prediction

No	Source	Text	Predicted Category
1	Kompas	Sistem pertandingan kandang dan tandang ini sebenarnya memberikan keunggulan bagi tim tuan rumah. Saat ini, tim-tim IBL hanya menghasilkan pendapatan dari penjualan merchandise di samping sponsor. Junas berharap bahwa dengan pengenalan sistem baru ini, setiap tim akan berusaha untuk menarik pendukung ke pertandingan mereka. Ini dapat menghasilkan pembentukan basis penggemar yang lebih luas, ujarnya	Basket
2	CNN	Pramudya Kusumawardana dan Yeremia Rambitan terlibat dalam perang dingin. Pelatih ganda putra, Aryono Miranat menceritakan momen-momen perselisihan itu muncul. Pemandangan aneh terlihat di Arctic Open tiap kali Pram/Yere bertanding. Pasangan tersebut minim interaksi dan komunikasi. Toast dengan tangan yang lazim terlihat saat menonton ganda bermain juga tidak tampak.	Raket
3	Detik	Shin Tae-yong memastikan bahwa Timnas Indonesia bakal tampil menyerang melawan Brunei Darussalam pada leg kedua Kualifikasi Piala Dunia 2026 zona Asia di Stadion Hassanal Bolkiah, Bandar Seri Begawan, Selasa (17/10/2023). Timnas Indonesia berhasil membuat Brunei tak berdaya di leg pertama Kualifikasi Piala Dunia 2026 zona Asia di Stadion Gelora Bung Karno (SUGBK),	Sepakbola (soccer)
4	CNN	Seorang marshal MotoGP Mandalika Pertamina Grand Prix 2023 Sofyan Sauri punya kenangan	Basket

No	Source	Text	Predicted Category
		berkesan dengan Marc Marquez meski tugasnya berjaga-jaga di lintasan masih hitungan jari. Sofyan yang berasal dari Praya, Lombok Tengah bukanlah marshal yang sudah kenyang pengalaman. Dari pengakuannya, Sofyan yang berusia 22 tahun baru tiga kali bertugas sebagai marshal di Sirkuit Internasional Pertamina Mandalika.	
5	CNN	Dalam video yang dibagikan lewat akun media sosial X (twitter) resmi Netanyahu, terlihat pemimpin Israel itu berbincang dan menyapa dengan sejumlah prajurit di basis militer di luar Gaza. Dia pun mendengarkan penjelasan dari komandan IDF di sana. Sebelumnya, tenggat waktu yang diberikan Israel buat warga meninggalkan Gaza, Palestina selama enam jam telah selesai, Sabtu petang waktu setempat.	Basket

4. Conclusion

Based on the results and discussion above, we can conclude that the model produced can be used to predict sports news into three categories based on the labels supplied to the sample data used to train the model. We also identify cases of inaccurate classification prediction which can be attributed to the model's limited categories; thus alternative strategies must be implemented so that the model can distinguish texts that fall outside of the stated categories. We are able to determine that the process of summarizing and categorizing text in a certain language is highly dependent on the availability of NLP libraries in that language. The results show that our chosen approach, text summarization using textrank, web scraping, and text classification, is capable of immediately addressing the problem of getting news content and category ready.

References

[1] M. Frackiewich, "The Impact of AI Text Summarization on Journalism and Media," *T2S Space*, 2023. https://ts2.space/en/the-impact-of-ai-text-summarization-on-journalism-and-media-2/#gsc.tab=0 (accessed Nov. 20, 2023).

[2] D. Wilding, P. Fray, S. Molitorisz, and E. McKewon, "The Impact of Digital Platforms on News and Journalistic Content," 2018. [Online]. Available: https://www.accc.gov.au/system/files/ACCC+commissioned+report+-+The+impact+of+digital+platforms+on+news+and+journalistic+content,+Centre+for+Media+Transition+(2).pdf.

[3] C. Zhu, "Applications and future of machine reading comprehension," in *Machine Reading Comprehension*, Elsevier, 2021, pp. 185–207.

[4] D. Miller, "Leveraging BERT for Extractive Text Summarization on Lectures," Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.04165.

[5] N. Zhou, W. Shi, R. Liang, and N. Zhong, "TextRank Keyword Extraction Algorithm Using Word Vector Clustering Based on Rough Data-Deduction," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–19, Jan. 2022, doi: 10.1155/2022/5649994.

[6] S. Kemahduta, "Automatic Text Summarization dengan kategorisasi pada berita online mengenai tokoh masyarakat indonesia dengan metode Fuzzy Logic," Universitas Sebelas Maret, 2019.

[7] H. Gupta and M. Patel, "Method Of Text Summarization Using Lsa And Sentence Based Topic Modelling With Bert," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Mar. 2021, pp. 511–517, doi: 10.1109/ICAIS50930.2021.9395976.

[8] K. U. Manjari, S. Rousha, D. Sumanth, and J. Sirisha Devi, "Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, Jun. 2020, pp. 648–652, doi: 10.1109/ICOEI48184.2020.9142938.

[9] P. Modaresi, P. Gross, S. Sefidrodi, M. Eckhof, and S. Conrad, "On (Commercial) Benefits of Automatic Text Summarization Systems in the News Domain: A Case of Media Monitoring and Media Response Analysis," Jan. 2017, [Online]. Available: http://arxiv.org/abs/1701.00728.

[10] K. S. Thakkar, R. V Dharaskar, and M. B. Chandak, "Graph-Based Algorithms for Text Summarization," in *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, Nov. 2010, pp. 516–519, doi: 10.1109/ICETET.2010.104.

[11] A. Abdurrohman, "Evaluasi Algoritma Textrank pada Peringkasan Teks Berbahasa Indonesia," Universitas Sumatera Utara, 2018.

[12] Y. Marsyah and S. H. Wijaya, "Perbandingan Kinerja Algoritme TextRank dengan Algoritme LexRank pada Peringkasan Dokumen Bahasa Indonesia," IPB University, 2013.

[13] S. R. K. Harinatha, B. T. Tasara, and N. N. Qomariyah, "Evaluating Extractive Summarization Techniques on News Articles," in *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Jul. 2021, pp. 88–94, doi: 10.1109/ISITIA52817.2021.9502230.

[14] M. Zhang, X. Li, S. Yue, and L. Yang, "An Empirical Study of TextRank for Keyword Extraction," *IEEE Access*, vol. 8, pp. 178849–178858, 2020, doi: 10.1109/ACCESS.2020.3027567.

[15] E. Yulianti, N. Pangestu, and M. A. Jiwanggi, "Enhanced TextRank using weighted word embedding for text summarization," *Int. J. Electr. Comput. Eng.*, vol. 13, no. 5, p. 5472, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5472-5482.

[16] S. Mishra, M. Kuznetsov, G. Srivastava, and M. Sviridenko, "VisualTextRank: Unsupervised Graph-based Content Extraction for Automating Ad Text to Image Search," Aug. 2021, doi: 10.1145/1122445.1122456.

[17] J. Ahmed and M. Ahmed, "ONLINE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES," *IIUM Eng. J.*, vol. 22, no. 2, pp. 210–225, Jul. 2021, doi: 10.31436/iiumej.v22i2.1662.

[18] Nur Ghaniaviyanto Ramadhan, "Indonesian Online News Topics Classification using Word2Vec and K-Nearest Neighbor," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 6, pp. 1083–1089, Dec. 2021, doi: 10.29207/resti.v5i6.3547.

[19] K. Munawaroh and A. Alamsyah, "Performance Comparison of SVM, Naïve Bayes, and KNN Algorithms for Analysis of Public Opinion Sentiment Against COVID-19 Vaccination on Twitter," *J. Adv. Inf. Syst. Technol.*, vol. 4, no. 2, pp. 113–125, Mar. 2023, doi: 10.15294/jaist.v4i2.59493.

[20] Z. Wang and Z. Liu, "Graph-based KNN text classification," *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*. IEEE, 2010, doi: 10.1109/fskd.2010.5569866.