Available online at www.garmian.edu.krd

Passer Journal of Basic and Applied Sciences

Passer 8 (Issue 1) (2026) 603-610

Original Article

Content-based Approach for Title Generation of Academic Papers

Mohammed Ali Mohammed[*1], Ahmed Jasim Jabur[2], Noor Maher Naser[2]

[1]*College of Business Informatics, University of Information Technology and Communications (UOITC), Baghdad, Iraq.*
[2]*University of Information Technology and Communications (UOITC), Baghdad, Iraq.*

Received 27 February 2025; revised 1 March 2026;
accepted 20 March 2026; available online 06 April 2026

DOI: 10.24271/PSR.2026.509186.1973

ABSTRACT

In a document, text, or article, the title should encapsulate the whole content with a few possible words. Thus, forming a powerful title that successfully summarizes the core of an academic paper is often a hard challenge. This paper introduced an automated title generation for academic papers to support the scholarly community that includes two types of process: sentence process and word process. The experimental results show that the title generation is close to the original scientific title. The title generated by the Sumy summarization algorithm and pre-model k2t is a better from other ways due to the highest similarity values (0.732). This approach reached close to real-time, reliable, presenting a methodology to improve the effectiveness of the title creation process.

https://creativecommons.org/licenses/by-nc/4.0/

Keywords: Title generation, Content-based approach, Text frequency, Text summarization, Automatic title generation tools.

1 Introduction

The significance of an academic paper title is extracted in two directions. The researchers quickly evaluate the paper based on its title without examining the whole article in the first direction [1,2]. In the second direction, the number of upcoming readers is highly dependent on the article's title, thus, affecting the citation metrics [3,4]. Accordingly, creating an exciting title is critical for researchers, while some others assign minimum time for this significant task [2].

For authors who are looking for well-written titles to achieve better indexing purposes, several automated title-creation techniques were developed [5]. However, title creation can be identified as a summarization task as an extremely small article summary [1]. For news articles, the extractive summarization approach was used by some researchers [5,6,7]. Extracting relevant words in the article without examining the document discourse structure was represented the most previous works. In reality, discourse offers a type of information that denotes the communication aim expressed to the readers by article authors [8].

The perspective of text summarization as a technique in title generating approach includes an article compression process to summarize its content [1]. The main issue in title creation is the inherent information scarcity. More specifically, when the text contains several terms, then creating a comprehensive and concise summary with only a few selected terms that convey the whole information becomes a primary challenge [9]. Thus, such an issue is far from the straightforward direction. In addition, achieving the shortest possible summary of the article is the aim of the title-generating process, and the error tolerance should be low since any factual error or slight grammatical lack can make the title functionally useless [10,11]

* Corresponding author
E-mail address mohammed.ali@uoitc.edu.iq (Instructor).
Peer-reviewed under the responsibility of the University of Garmian.

However, the contribution of this paper addresses the title generation process using a novel approach, which is comprised of three integrated modules, namely preprocessing, information selection, and title generation. The first module performs five tasks, including capitalization, noise removal, tokenization, stop-word, and lemmatization. The second module includes two approaches. The first approach is named (word approach), which consists of three algorithms: Term Frequency (TF), text rank, and Yake [12], to generate the weighted words. The second approach is named (sentence approach), which has three algorithms: text rank, Sumy, and T5, to achieve an abstractive sentence known as sentence extraction, as well. The last module generates the title based on the key-to-text library, laying the foundation for crafting the final paper title. By navigating through these three modules, this work provides a comprehensive and effective methodology for the title generation of academic papers.

The organization of this paper includes an overview of the related work presented in section 2, while section 3 explains the proposed approach. The tentative results and analysis are presented in section 4, and the conclusion in section 5.

2 Related Work

Investigating the previous work in the field of title generation revealed a plentiful textile of approaches involving several algorithms and techniques. A varying degree of success has been achieved by these approaches in summarizing and capturing the complicated content of academic papers. The examination of such varied circumstances offers helpful perceptions through the assessment of title generation methods, permitting a notable understanding of the limitations and strengths of each method.

Putra and Khodra [8] (2017) introduced an automated title generation system for academic papers, considering rhetorical categories, especially the information types of textual units. According to the paper's abstract processing, the system generates several title candidates. The carrying out of tests with integrated rhetorical sentence categories was considered as their notable contribution.

Putra and Fujita [13] (2015) noted that the paper's abstract was selected as a foundation for the task of title generation, due to its briefness, ease of accessibility, and adequate information representing the paper idea.

Chen and Lee [5] (2003) proposed an innovative approach to automate title generation by adapting the title of another paper using terms from a given input paper. They develop a model for Chinese spoken papers, involving key phrase extraction, topic classification, and an adaptive K nearest-neighbor concept. The model was tested on 210 broadcast news stories, while the training set comprised 151,537 news stories with human-generated titles in text form.

In related work, the authors [14] (2020) realized a method set of title generation. They trained the system with 21,190 news stories, while the testing set consists of 1,006 broadcast news. They compared the results of automatically recognized speech to the results of manual transcription. The evaluation metrics included F1 and the average number of correct title words in the correct order. For speech-recognized news stories, the findings of title generation are achievable close to the generated title accuracy for precise text transcripts.

A method called DTATG was introduced by Chao and Wang [15] (2016) for automated title generation. Initially, dominant sentences expressing the crucial meanings of the text and proper for adaptation into a title are extracted by DTATG. Next, a dependency tree is constructed for each sentence. In contrast, the dependency tree compression model removed certain branches. The findings showed that the proposed method could generate sufficient titles with a high F1 score compared to previous methods.

Conversely, to generate a title for an unstructured article, D Bajaj et al. [16] (2022) proposed a novel approach, using deep natural language processing. The problem was reframed as a sequential question-answer task. Note that the title vocabulary is a subset of the article vocabulary.

Lastly, TiZen was a designed model by Jain et al. [17] as a neural title generator by generating the title exclusively based on the paper abstract to enhance the author's work to be noticeable in the community. The proposed model creates a catchy title using specific ground rules, along with extractive and abstractive text summarization. TiZen highlighted that the catchy title motivates the readers to investigate the whole content, as well as to capture their attention. In addition, the findings were that the generated titles were meaningfully close to the expected titles that an author might do in advance

3 Proposed Approach

The intended system achieved a generated article title as a task rooted in summarization. It consists of three integrated modules, where each module performs essential tasks inside the basic pipeline of summarization. These modules are denoted as preprocessing, information selection, and title generation. Figure 1 illustrates the block diagram of the proposed system. The functionality of each module is explained in the following sub-sections.

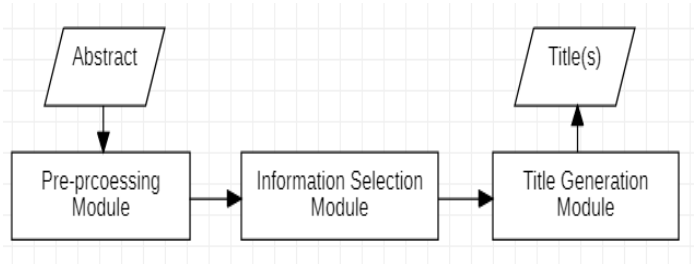

Figure 1: The block diagram of the intended system.

3.1 Pre-processing module

In general, most documents, texts, and articles have several unnecessary (redundant) words like stop-words, slang, misspellings, etc. However, some techniques for text cleaning and preprocessing text datasets are explained in the following (see Figure. 2). Capitalization handling: converting all words to upper or lower case letters is the general approach to overcome such variability [18]. Noise removal redundant characters, special characters will be eliminate unnecessary characters [19]. Tokenization: text breakdown into individual words, symbols, and other entities, according to space character [17,18]. Stop-word removal: removing the stop words (such as "is", "the", "that", etc.) from the text, increasing the relevance of the remaining content [20]. Lemmatization: It is a natural language processing (NLP) procedure aimed at standardizing words by removing or replacing suffixes. The words that transform into their root or base forms are known as lemmas [21].

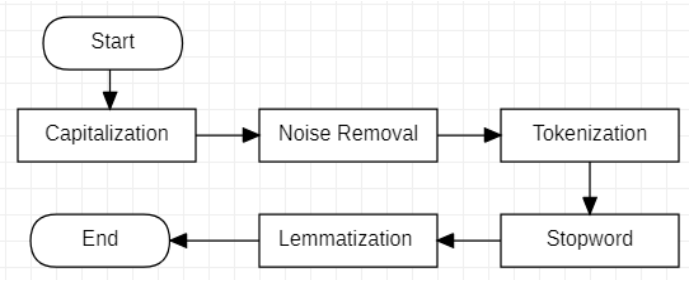

Figure 2: The utilized techniques of the preprocessing module.

3.2 Information selection module

The output of the first module comprises weighted and unweighted words. The second module known as information selection is employed to isolate the unweighted words effectively. The isolation process is performed through two different approaches: word and sentence. Each approach embraces several algorithms that are designed to generate a set of candidate titles. These approaches are explained in the following:

3.2.1 Word approach

This approach comprises three algorithms: text frequency, text rank, and yake to perform additional processing to the output of the first module. Each algorithm extracts further relevant information that is added to the module output and serves as input to the last module (title generation). This accurate approach certified that each candidate title is derived from a robust selection of weighted words, improving the overall quality and relevance of the generated titles. Additional details of these algorithms are mentioned in the following:

3.2.1.1 Text frequency (TF)

It is also known as the bag-of-words algorithm [22]. It represents the most straightforward technique for text feature extraction. It involves the counting of words in each document and their assignment to the feature space. Methods extending TF often utilize word frequency with Boolean or logarithmically scaled weighting. The TF generates a vector containing the word and its occurrence frequency throughout the entire text, ultimately identifying the word with the maximum number of occurrences [21,23,24]. The calculation of TF is expressed as follows [25]:

$$TF(i,j) = \frac{n(i,j)}{\sum_k n(i,j)}$$

Where $\sum_k n(i,j)$ is the frequency sum of all words in the text, k denotes the total number of words, $n(i,j)$ represents the frequency of that word in the text, and $TF(i,j)$ signifies the Term Frequency of the word i in the document.

3.2.1.2 TextRank

It is also recognized as a graph-based algorithm. It conceptualizes words as nodes in the graph, while the connections between words represent the graph's edges. Additionally, the relationships among words are indicated by the occurrence number of the words within a fixed-sized sliding window [7].

3.2.1.3 Yake

For multi-lingual keyword extraction from single documents, accommodating texts of diverse sizes, domains, or languages, Yake is a groundbreaking feature-based system. In contrast, most systems depend on dictionaries, thesauri, or training against corpora, while Yake does not. As an alternative, it follows an unsupervised approach, building upon features extracted from the text. This unique characteristic renders Yake appropriate to papers written in various languages lacking the necessary for outside information [12].

3.2.2 Sentence approach

After pre-processing, the abstract becomes an input to summarization algorithms (Text Rank, Sumy, and T5 algorithms) employing both Extractive and Abstractive approaches. Subsequently, the output is fed back into the word approach.

Text summarization addresses the task of lowering the number of words and sentences in a document without altering its meaning. However, there are different methods for extracting information from raw text data for summarization models, broadly categorized as Extractive and Abstractive. Extractive methods identify crucial sentences within a text without necessarily comprehending their meaning, resulting in a summary that is a subset of the full text. In contrast, Abstractive models employ advanced NLP techniques, such as word embedding [26], to grasp the semantics of the text and generate a meaningful summary. Consequently, Abstractive techniques are more challenging to train from scratch, requiring a substantial number of parameters and extensive data.

3.3 Title generation module

The title generation represents the concluding phase of the proposed methodology. In this step, the key-to-text library is utilized to generate a title using the weighted words derived from the initial approach as input values. Figure 3 provides a more detailed illustration of the steps involved in the proposed methodology.

The Key-to-text library operates on the principle of constructing a model capable of taking keywords as inputs and producing sentences as outputs. Potential use cases encompass areas such as Marketing, Search Engine Optimization, Topic Generation, and the fine-tuning of topic modeling models. Key-to-text leverages the capabilities of the Amazing T5 Model, k2t Model, and k2t-base Model [27]. Algorithm 2 shows the main steps of the proposed system.

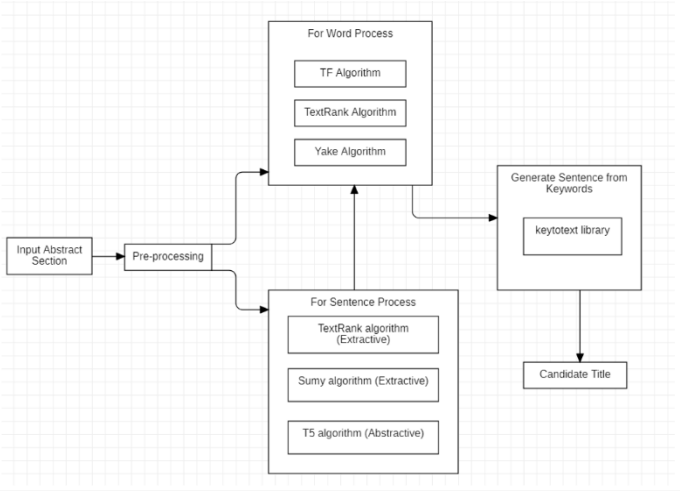

Figure 3: Detailed steps of the Proposed Methodology.

Algorithm 2: Proposed methodology	
Input:	New Abstract
Output:	Set of titles
Steps:	
1.	Apply the pre-processing step.
2.	List1: Get weighted word:
2.1	Apply the TF algorithm.
2.2	Apply Text Rank algorithm
2.3	Apply Yake algorithm
3	List 2: Get weighted words after applying text summarization:
3.1	Apply Text Rank algorithm.
3.2	Apply Sumy algorithm
3.3	Apply T5 algorithm
4.	Generate Titles using a key-to-text library with List 1 and List 2 as input values.

3.4 Evaluation metric

Among different evaluation metrics, the similarity method is the mostly closest metric, and it is obtained by comparing word vectors (so-called "word embeddings"). These vectors are multiple dimensions of meaning representations of a word. An algorithm like word2vec can be used to generate these vectors [27]. Small pipeline packages known as "spaCy's" are employed to make these vectors fast and compact. It can compare two objects and generate a prediction of how much they are similar. Predicting similarity is very helpful for flagging duplicates and building recommendation systems. For instance, it is possible to label a support ticket as a duplicate if it is extremely close to a previously existing one or suggest user content, which is identical to what they are now searching for. Implementation of spaCy's similarity commonly supposes a well enough similarity general-purpose definition.

4 Results and Discussions

Sun, et al. (2014) [28] published an article entitled "Data security and privacy in cloud computing". It is employed as a step-by-step example to explain and certify the proposed system. The abstract of the paper is started with "Data security has consistently been a major issue" and ended with "techniques used in cloud computing".

As mentioned earlier, the intended system consists of three modules. The first module (named preprocessing) includes cleaning and removing unwanted words like "in, this, we, of, the, has, etc.". In addition, a lemmatization step is then applied, where the suffix is removed, for example, the word "techniques" is converted to "technique". Thus, the text under test beyond the first module becomes: "data security consistently major issue information technology. cloud computing environment, becomes particularly serious data located different place even globe. data security privacy protection main factor user's concern cloud technology. though many techniques topic cloud computing investigated academic industries, data security privacy protection becoming more important future development cloud computing technology government, industry, business. data security privacy protection issues relevant hardware software cloud architecture. study review different security technique challenge software hardware aspect protecting data cloud aim enhancing data security privacy protection trustworthy cloud environment. paper, make comparative research analysis existing research work regarding data security privacy protection technique used cloud computing".

The second module (named information selection) consists of two approaches: word and sentence. The word approach has three individual algorithms. Table 1 lists the findings of these algorithms.

Table 1: the findings of the word approach.

TF Algorithm		**TextRank algorithm**	**Yake algorithm**	
Word	**Weights**		**Word**	**Weights**
data	0.08	Protecting	Security	0.052
cloud	0.08	data security	Data	0.062
security	0.07	technology cloud computing	Privacy	0.065
privacy	0.05	Research	Cloud	0.072
protection	0.05	Different	Protection	0.098
computing	0.03	Issue	Computing	0.124
technique	0.03	Industry	Technology	0.178
issue	0.02		Technique	0.212

technology	0.02		Consistently	0.257
different	0.02		Major	0.257

This approach can proceed alone to the two modules without the sentence approach. Table 2 shows the candidate titles using the pre-trained models k2t and k2t-base.

Table 2: The candidate titles with only the word approach.

Pre-trained model	Algorithm	Candidate Title	Similarity value
k2t	TF	The protection of the data with the protection of computing technology is privacy. The name of the difference between the two teams is different	0.631
	Text Rank	The study of protection| data security technology is different from the industry.	0.689
	Yake	The security protection of the 'technical' technology' was a major.	0.490
k2t-base	TF	The protection of data in the cloud with privacy of privacy is a type of technology. The different names for people are different.	0.641
	Text Rank	The protection of data security is a type of research that is part of the industry.	0.589
	Yake	The major component of security systems is encryption, which is a type of protection.	0.584

The second approach is intended to enhance the accuracy of the proposed system. It also has three different algorithms. Its output is feedback on the word approach to improve the selection accuracy. Tables 3, 4, and 5, list the output of the word approach for the second module using the TextRank, Sumy, and T5 algorithms, respectively.

Table 3: The findings of the second module using the TextRank algorithm of the sentence approach.

TF Algorithm		TextRank algorithm	Yake algorithm	
Word	Weights		Word	Weights
Data	0.081	Cloud	Security	0.093
Security	0.081	Protection main factor	Privacy	0.093
Privacy	0.081	Research analysis existing	Data	0.177
Protection	0.081	Architecture	protection	0.144
Cloud	0.081	Hardware	technology	0.177
Research	0.054	Data	Cloud	0.1863
Main	0.027		Main	0.252
Factor	0.027		Factor	0.252
User's'	0.027		User	0.252
concern	0.027		Concern	0.252

Table 4: The findings of the second module using the Sumy algorithm.

TF Algorithm		TextRank algorithm	Yake algorithm	
Word	Weights		Word	Weights
Cloud	0.093	cloud computing	Security	0.045
Data	0.069	protecting	Privacy	0.049
Security	0.069	data security	Cloud	0.049
Technique	0.046	industry	Data	0.055
computing	0.046	technique	industry	0.060
privacy	0.046	software hardware	computing	0.064
protection	0.046		protection	0.064
though	0.023		technique	0.081
many	0.023		industries	0.091

topic	0.023		government	0.091

Table 5: The findings of the second module using the T5 algorithm.

TF Algorithm		TextRank algorithm	Yake algorithm	
Word	**Weights**		**Word**	**Weights**
Data	0.074	research	Data	0.069
research	0.074	data security	Security	0.077
Security	0.061	technology cloud computing	research	0.080
Cloud	0.061	protection main factor	privacy	0.100
privacy	0.049	different	cloud	0.101
protection	0.049		protection	0.114
paper	0.037		paper	0.116
make	0.037		make	0.140
comparative	0.037		comparative	0.162
analysis	0.037		analysis	0.162

Finally, the developed candidate titles from integrating the sentence approach (TextRank, Sumy, and T5) with the word approach (TF, TextRank, and Yake) are listed in Table 6 using the pre-trained models k2t and k2t-base.

Table 6: The developed candidate titles from integrating word and sentence approaches.

Pre-trained model	Sentence approach algorithm	Word approach algorithm	Candidate Title	Similarity value
k2t	TextRank	TF	The main ingredient in the protection of the cloud is the user's.	0.591
		TextRank	The cloud| protection's main factor in a hardware store is the hardware.	0.591
		Yake	The main security of the system, which has a protection of the data, is privacy.	0.643
	Sumy	TF	The web browser, with protection of many aspects, is known as a cloud.	0.568
		TextRank	The technology for cloud computing is software hardware.	0.630
		Yake	The security protection for cloud services in the industry of the company "Technical Engineering Industry" is cloud.	0.732
	T5	TF	The journal "Spare " is also the protection for the paper.	0.459
		TextRank	The main factor of research in the field of data security is different.	0.676
		Yake	The protection of the data with a paper composition of analysis is "guardians".	0.600
k2t-base	TextRank	TF	The main ingredient of data protection is the cloud	0.639
		TextRank	The main factor of cloud research analysis is research analysis.	0.583
		Yake	The main ingredient of security systems is security.	0.602
	Sumy	TF	cloud contains many different types of security, such as many types of encryption, one of which is a number of types of protection.	0.591
		Text Rank	Cloud Computing is a type of software used in the industry.	0.671

		Yake	The government of the United States is the source of the encryption of the "cloud".	0.585
	T5	TF	"Spare | analysis" is a "for privacy|protected" paper.	0.245
		Text Rank	The main ingredient of a data security system is different.	0.590
		Yake	The study of the cloud with a comparative analysis is privacy.	0.614

5 Discussion

Figuring similarity scores is very valuable in various cases. In contrast, it is also crucial to have realistic expectations concerning what information dose similarity can offer. More deeply, there are several ways to correlate words with each other, and getting only one similarity score often means a mix of various signals. In addition, training vectors on diverse data will deliver out-of-the-ordinary results, which may be worthless for the required purpose. Thus, one can keep in mind the following important considerations:

a- Similarity has no objective definition. For instance, “I like pasta” and “I like burgers” are the same based on the application, since both talk about food preferences. In contrast, if the analysis is about the mentions of food, then both sentences are completely not the same.

b- The Span and Doc object have similarities that default to the average of the token vectors. More specifically, the vector “fast food” means the average of the vectors for “fast” and “food”. Indeed, it is not necessarily illustrative of the phrase “fast food”.

c- Vector averaging can be defined as a vector of multi-tokens that is insensible to the sequence of words. For example, a lower similarity score can be achieved if two documents with unlike wording express the same meaning than other two documents expressing different meanings but contain the same words.

Conclusion

After comparing and discussing these approaches, the sentence approach showed better than the word approach, while the proposed system has the best performance when using the summarization step with the sumy summarization algorithm and pre-model k2t due to the highest similarity values (0.732). The approach reached close to real-time, efficient, more reliable, and highly accurate.

In considering an academic paper, the generation of a powerful title is an essential work, which affects the primary overview of a scholarly task. Constructing an exciting title that successfully summarizes the core of the author's research is often a hard challenge. The field of automatic title generation is investigated in this paper, especially the employment of the content-based approach. Moreover, the paper explores a novel technique to enhance and modernize the title creation process by leveraging the inherent paper content itself. The methodology includes using two processes: the sentence process and the word process, to extract crucial salient information, concepts, and themes from the paper. In contrast, the systematic content-based approach offers a solution to the difficult task of talking to the authors by automating the process of title generation. The findings showed a helpful contribution to scholarly communication with a comprehensive discourse, presenting a technique to improve the effectiveness of the title creation process and to ensure that the created title matches strongly with elementary paper content. This approach offers a valuable tool for authors to construct titles that efficiently express the importance of their research, and it has the potential to optimize the workflow of scholarly publications.

References

[1] Jamali, H. R. & Nikzad, M. Article title type and its relation with the number of downloads and citations. *Scientometrics* **88**(4), 653–661, doi:10.1007/s11192-011-0412-z (2011).
[2] Xu, H., Martin, E. & Mahidadia, A. Extractive summarisation based on keyword profile and language model. *NAACL HLT Proc.* 123–132, doi:10.3115/v1/n15-1013 (2015).
[3] Paiva, C. E., Lima, J. P. S. N. & Paiva, B. S. R. Articles with short titles describing the results are cited more often. *Clinics* **67**(5), 509–513, doi:10.6061/clinics/2012(05)17 (2012).
[4] Letchford, A., Moat, H. S. & Preis, T. The advantage of short paper titles. *R. Soc. Open Sci.* **2**(10), 10–15, doi:10.1098/rsos.150266 (2015).
[5] Chen, S. C. & Lee, L. S. Automatic title generation for Chinese spoken documents using an adaptive k nearest-neighbor approach. *Proc. Eurospeech* 2813–2816, doi:10.21437/eurospeech.2003-749 (2003).
[6] Teufel, S. Argumentative zoning: Information extraction from scientific text. (University of Edinburgh, 1999).
[7] Li, W. & Zhao, J. TextRank algorithm by exploiting Wikipedia for short text keywords extraction. *Proc. ICISCE* 683–686, doi:10.1109/ICISCE.2016.151 (2016).
[8] Putra, J. W. G. & Khodra, M. L. Automatic title generation in scientific articles for authorship assistance: A summarization approach. *J. ICT Res. Appl.* **11**(3), 253–267, doi: 10.5614/itbj.ict.res.appl.2017.11.3.3 (2017).
[9] Lin, H. et al. Gen-FL: Quality prediction-based filter for automated issue title generation. *J. Syst. Softw.* **195**, 111513 (2023).
[10] Diao, J. A lexical and syntactic study of research article titles in Library Science and Scientometrics. *Scientometrics* **126**, 6041–6058 (2021).
[11] Dua’a, A., Alyasiri, O. M., Allogmani, E., Amer, M. & Salman, T. M. S. Unlocking ChatGPT’s title generation potential. *J. Theor. Appl. Inf. Technol.* **101**, 22 (2023).
[12] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C. & Jatowt, A. YAKE! keyword extractor. *Lect. Notes Comput. Sci.* **10772**, 806–810, doi:10.1007/978-3-319-76941-7_80 (2018).
[13] Wira, J. & Putra, G. Scientific paper title validity checker utilizing vector space model. *Proc.* 88–93 (2015).
[14] Gu, X. et al. Generating representative headlines for news stories. *Proc. Web Conf.* 1773–1784 (2020).

[15] Shao, L. & Wang, J. DTATG: Automatic title generator based on dependency trees. *IC3K Proc.* **1**, 166–173, doi:10.5220/0006035101660173 (2016).
[16] Bajaj, D. et al. TiGen–Title generator based on deep NLP transformer model. *Proc. Int. Conf. Commun. Netw. Comput.* 297–309 (2022).
[17] Jain, H. et al. TiZen: Neural title generation for scientific papers (n.d.).
[18] Tated, R. R. & Ghonge, M. M. A survey on text mining techniques and application. **3**, 380–385 (2015).
[19] Pahwa, B., Taruna, S. & Kasliwal, N. Sentiment analysis strategy for text pre-processing. *Int. J. Comput. Appl.* **180**, 15–18, doi:10.5120/ijca2018916865 (2018).
[20] Saif, H., Fernandez, M., He, Y. & Alani, H. On stopwords and filtering for sentiment analysis. *Proc. LREC* 810–817 (2014).
[21] Kowsari, K. et al. Text classification algorithms: A survey. *Information* **10**, 1–68, doi:10.3390/info10040150 (2019).
[22] Github. TF-IDF tutorial. Available at: https://ethen8181.github.io/machine-learning/clustering_old/tf_idf/tf_idf.html (Accessed 11 January 2025).
[23] Salh, D. A. & Nabi, R. M. Kurdish fake news detection based on machine learning approaches. *Passer J. Basic Appl. Sci.* **5**, 262–271, doi:10.24271/psr.2023.380132.1226 (2023).
[24] Abdulhammed, O. Y. & Karim, P. J. Sentiment analysis using SVM-based SSO intelligence algorithm. *Passer J. Basic Appl. Sci.* **4**, 17–39, doi:10.24271/psr.2022.160801 (2022).
[25] Erra, U., Senatore, S., Minnella, F. & Caggianese, G. Approximate TF–IDF based on topic extraction using GPU. *Inf. Sci.* **292**, 143–161 (2015).
[26] Hussein, D. J. et al. Machine learning approach to sentiment analysis. *Passer J. Basic Appl. Sci.* **4**, 71–77, doi:10.24271/psr.2022.312664.1101 (2022).
[27] keytotext. Python package. Available at: https://pypi.org/project/keytotext/ (Accessed 17 January 2025).
[28] Sun, Y., Zhang, J., Xiong, Y. & Zhu, G. Data security and privacy in cloud computing. *Int. J. Distrib. Sens. Netw.* **10**, 190903 (2014).