

Automatic summarization of YouTube video transcription text using term frequency-inverse document frequency

Rand Abdulwahid Albeer, Huda F. Al-Shahad, Hiba J. Aleqabie, Noor D. Al-shakarchy

Department of Computer Science, Faculty of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq

Article Info

Article history:

Received Sep 2, 2021

Revised Mar 12, 2022

Accepted Apr 1, 2022

Keywords:

Automatic text summarization

Stop words

Term frequency-inverse

document frequency

Video transcript

Word frequency

ABSTRACT

Automatic summarization is a technique for quickly introducing key information by abbreviating large sections of material. Summarization may apply to text and video with a different method to display the abstract of the subject. Natural language processing is employed in automated text summarization in this research, which applies to YouTube videos by transcribing and applying the summary stages in this study. Based on the number of words and sentences in the text, the method term frequency-inverse document frequency (TF-IDF) was used to extract the important keywords for the summary. Some videos are long and boring or take more time to display the information that sometimes finds in a few minutes. Therefore, the essence of the proposed system is to find the way to summarize the long video and introduce the important information to the user as a text with few numbers of lines to benefit the students or the researchers that have no time to spend with long videos for extract the useful data. The results have been evaluated using Rouge method on the convolutional neural network (CNN)-dailymail-master data set.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Rand Abdulwahid Albeer

Department of Computer Science, Faculty of Computer Science and Information Technology

University of Kerbala

Karbala, Iraq

Email: rand.a@uokerbala.edu.iq

1. INTRODUCTION

Individuals are intimidated by the vast number of web papers and information, as well as the emotional development of the internet. Ajmal and Haroon [1], the development of document usability has necessitated comprehensive study in the field of automated text summarization. A summary is a text derived from one or more texts that contains just the most relevant information from the original text(s) and is no more than half the length of the original text(s), if not significantly less. A summary's major objective is to convey the essential concepts of a text in as little space as possible. Generating a summary would be futile if all sentences in the text document were similar, because any reduction in the size of the document would result in an equivalent reduction in the content of the information [1].

The challenge of creating a short and smooth summary while retaining important information content and overall meaning is known as automatic text summarization several techniques for automatic text summarization have been developed in recent years and utilised widely in a variety of fields. Automatic text synthetization is a particularly difficult task since we, as humans, generally read a text piece to fully comprehend it and then create a summary that shows its major ideas. Automated text summarization is a hard and time-consuming procedure since computers lack human language and knowledge [2].

The automated text summarization makes consumers less complex with various implementations for the natural processing language, close to data recovery, answering questions, and content diminishing. Programmed text summarization requires an eventual portion of this data by presenting notable and relevant data [2], [3]. In the last several years, there have been a number of notable studies in the field of text summarization. Earlier efforts focused mostly on single document text summarization. With the increasing technology and computer capacity, this has led the way for quicker, more efficient and more precise processing of documents compared to previous ways [4].

Programmed text summarization is a bit of the common dialect preparation area that helps personal computers (PCs) to distinguish and taking importance from the human tongue as, Mâaloul *et al.* [5] present an automated Arabic text summary strategy based on the rhetorical structure theory (RST). Keskes *et al.* [6] use another method to summarize Arabic text that is the latent semantic analysis (LSA) technique model. Froud *et al.* [7] represent their system that summarizes the Arabic text using natural language processing. This proposal uses two algorithms of the steaming algorithms are Lovins steamer algorithm that use to remove the longest and used n-grams algorithm to make the similarity between the words down to the root account and repeat the words in the text to get the summary. Generally, two automatic text summary strategies exist: extractive and abstract, as discussed in the research of Mahdipour [8] discussing the method of lexical chain analysis that can overcome multiple problems with text mining, such as summarization, classification, and sentiment analysis. Kiyoumars [9] compares different automatic text summarization methods that use fuzzy with the summarization that made by human and conclude that there is no big difference between them and the fuzzy method developed are more economical.

Sah *et al.* [10] also use video summarization with two tasks: first, divide the video into parts as joint photographic experts group (JPEG) files and then arrange it. Then they record the parts of the result according to word frequency analysis of speech transcripts. Several elements may be clearly examined for YouTube videos: visual pictures, video metadata, (such time and creator), music, and transcripts. One approach to defining video content is by transcribing the videos relying on video transcripts opens it up to computational textual analysis techniques to look for word interrelationships, often used words or sentences, and themes [11].

This research seeks to obtain an overview of the contents of the original text. Abstract creation is one of the last goals of automated summary work. Applying the summary to the YouTube transcription by first fetching, saving the scripts in a file, and then using the suggested technique is different from the previous investigations (the term frequency-inverse document frequency (TF-IDF) method) i.e., the research takes the content of the descriptive video and summarizes it using the keywords obtained by TF-IDF to give the shortest description text. Summaries generated automatically may lack the coherence and intelligence of summaries created by humans. The majority of the time, however, the readers were able to comprehend the summaries presented. The remaining paper is structured accordingly. The techniques of the summary are presented in section 2. Section 3 describes the proposed study's "research technique," while section 4 discusses the outcomes and debates, as well as the experimental work. Section 6 concludes the study by emphasizing the important points and conclusions.

2. METHODS OF SUMMARIZATION

Today, text summarization is based on extracting the important sentence model. Several features show the relationship between the phrases and the document or multi-documents being used as the basis of the text translation. Sentence selection methods might be based on statistical data or on some informed summarizing theory that considers linguistic and semantic information [12]. There are many ways to deal with the sentence [4]:

2.1. Multi-document summarization

A multi-document summary is a collection of interrelated papers that sum up the contents in a concise manner. There are different ways to link documents, for example, documents that have the same structure can be related, that talk about the same topic, or because documents are about the same event. Detection and elimination of duplication and determination of contradictions are issues that arise when summarizing inputs [12]. Taskiran *et al.* [13], the multi-report outline issue is concentrated concerning multi-source data extraction in explicit areas. Here layouts started up from different reports are combined utilizing explicit administrators targeting identifying indistinguishable or conflicting data.

2.2. Multilingual summarization

Most researches of summarization executed maybe for the English languages because data and the evaluation resources are available; the summarization of other languages is not infrequent. Summarization practices in Japanese were included in the challenges of text summarization [14], and evaluations' series of

text summary systems with tasks such as summarization of one document and summarization of several documents on the same subject. In 2005, multi-language summary assessment was absorbed by describing combined inputs in Arabic and English when automatic translations were required [15].

2.3. Extraction method

The extractive summary method selects certain words, phrases, or sentences from the source text to create the summary. It's also focused on what the summary should be. It is necessary for the recovery of sentences from the text. It is used as a title process as a statistical methodology, for keyword collection and critical text phrases, use the term frequency-inverse document frequency (TF-IDF) technique, location method, and word process [16], [17].

2.4. Abstraction method

The abstractive summarization is used the way to generate the sentence by using semantic representation and use the technique of natural language to produce an overview similar to the human summary. The idea is to provide a summary including the words which are not used in the initial text. It uses a linguistic method such as the lexical chain, WordNet, graph theory, and clustering to explain and summaries the original text [18].

3. RESEARCH METHOD

This paper proposes a text summary of the video content system based on TF-IDF. Proposed model employs python 3.5 programming language to implemented a system functions and steps. The main aim of the proposed system is generating a summary text to the content of the YouTube video containing all important information. The general block diagram of the proposed system stages illustrated in Figure 1. The convolutional neural network (CNN)-dailymail-master dataset is used in this paper which is a dataset for text summarization, it contains online news stories written by journalists at CNN and the daily mail; it has paired with multi-sentence summaries and the models are evaluated using ROUGE-1, ROUGE-2, ROUGE-L [19], [20]. The details of the steps are demonstrated in the following subsections.

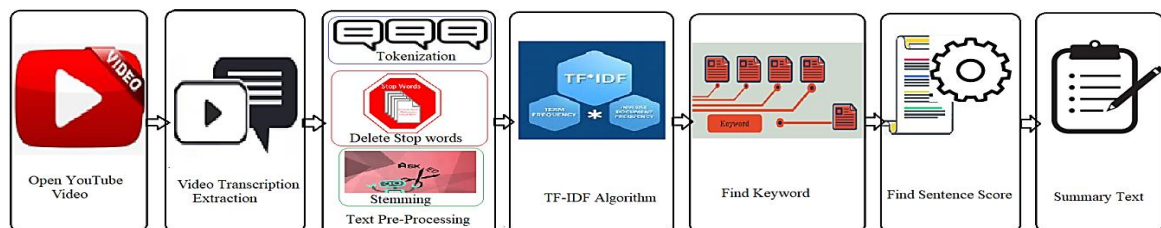


Figure 1. The proposed system block diagram

3.1. Open YouTube video and transcription extraction

YouTube has several beneficial functions, such as title and description translation. Likewise, you can transcribe YouTube videos in a variety of ways. Because voice recognition software has advanced significantly in recent years, you may now obtain a dependable automatic transcription that can be readily modified to perfection with little to no work. It is fairly simple to transcribe YouTube videos because YouTube automatically transcribes the majority of them as soon as they are uploaded. The first step in the proposed approach; which is done handly; is to access the YouTube website and select a video. Video transcription is done to translating a video's audio into text automatically by retrieve the transcription of the video id and save it in a pdf file to be used in the subsequent steps. The transcription text (The text extracted from YouTube) which is called original text. The retrieved text varies in length and may differ from video to the next.

Original text: *The article on cricket and submitted it on the Friday itself while Bo's plan to submit it on Monday. Manu was punctual and regular in all his work. Bo's was bit lazy guy and used to keep things postponing both, you saw the noticeboard there is a competition to be held by next week you know the topic? I haven't seen the noticeboard tell me what the topic is it's about cricket? and you know what the prize is for the winner?. A cricket gear what a cricket gear oh I will take part in this competition me too, we have to write a two-page essay on the game cricket and submit it by Monday the next week. Monday, we have still 5*

days for that a lot of time but we need to collect the articles, let's go to the library tomorrow. The next day both come, let's go to the library no I'm going to play Saturday's holiday I will go to library on that day you carry on man you wrote.

3.2. Text preprocessing

The initial processing step is the first significant stage in natural language processing which consists of three stages. The first one is tokenization which splits each phrase into a series of words or terms. The second is to eliminate English stop words, which is a way to efface letters and words with no denotement in the sentence and reiterate more than once in the text so that the text will be pristine from stop words. Table 1 shows a sample of the stopwords. The last stage is word-stemming; the central concept is to handle the word that cessations or beginning by minimizing the phrases or words to their word roots, kened as a lemma. Stemming is typically performed before the word's final assignment to the index by deleting all affixed suffixes and prefixes (affixes) from index words.

Table 1. English stop words

Stop words					
a	about	above	after	again	all
am	and	any	are	as	at
been	because	below	both	being	by
but	couldn't	could	cannot	do	didn't
for	from	had	has	have	my
no	nor	not	on	off	once
only	my	myself	most	our	we

3.3. Generate unique word frequency

In this step, the unique words will be computed using the unique word frequency (UWF). The iteration of each unique word will be counted to see which of the words is essential in the text in the summarization text. The iterations numbers had sorted in descending order according to their repetition. Table 2 shows a sample of the word frequency in our example.

Table 2. Word frequency

Word	Count	Word	Count	Word	Count
Cricket	5	Gear	2	keep	1
next	3	Regular	1	Seen	1
noticeboard	2	Work	1	tell	1
competition	2	held	1	Manu	1
week	2	Punctual	1	prize	1
know	2	Lazy	1	winner	1
topic	2	Used	1	things	1

3.4. Find keywords

Finding keywords is the consequential step in the system, to filter the words in the text, TF-IDF was utilized this approach measures the words consequential in a sentence and the number of times a word is included in a text. The word is very paramount if it is reiterated in a sentence, but less reiterated in a document [21], [22].

TF-IDF is equal to TF*IDF, both TF and IDF were computed (1) and (2) respectively:

$$TF(i, j) = f(i, j) / \sum_{i' \in j} f(i', j) \tag{1}$$

where $f(i, j)$ is the number of repetitions of the word i in document j . It's worth noting that the numerator is just the entire number of phrases in document j (counting each occurrence of the same term separately).

$$IDF(i) = \log(|D| / d_i) \tag{2}$$

where $|D|$ is the total number of sentences in the input text, and d_i is the number of sentences where the word i appears. In our example, the number of sentences is 12 after splitting it according to the end with one of the special characters (“.”, “,”, “;”, “?”, “!”) and the keywords are: let, cricket, bit lazy, saturday, play, week, monday, tell, page, essay, library, noticeboard, tomorrow, things postponing, come, articles and article.

3.5. Sentences score

After the calculation is culminated, the words must sort in descending order according to their value. The sorting of all words is very consequential to test the TF-IDF rank. Afterward, the sentence's consequentiality value should be calculated utilizing the sum of each verb and entity in it, the values should be sorted in decrementing order. For example, the value of the sentence *Manu was punctual and regular in all his work=4* because *Manu* (1), *punctual* (1), *regular* (1) and *work* (1) while the sentence *A cricket gear what a cricket gear oh I will take part in this competition me too* after calculating the sentence value, the result was *A cricket* (5), *gear* (2), *take* (1), *part* (1) and *competition* (2)=11.

3.6. Summary text

The final step in the proposed system is summary generation according to the paramount sentences calculated in the precedent step. All pristine sentences that remained eligible during the precedent process had been designated as "summary". If a sentence is found twice or two sentences with similar keyword collections, the approach removes such a sentence and provides the user with a final summary.

Summary text: *Manu punctual regular work. Bo 's bit lazy guy used keep things postponing saw noticeboard competition held next week know topic n't seen tell cricket prize winner gear oh take part write two - page essay game submit monday still days lot time need collect articles let go library tomorrow day come 'm going play Saturday holiday carry man wrote article submitted Friday Bo plan.*

4. RESULTS AND DISCUSSIONS

The program has been implemented by python and the program's interface is engendered utilizing the python graphical utilizer interface package called the Tkinter. There is a supplemental package including a natural language toolkit for text processing, Math, and PyPDF2. During the execution, the user was asked to pick a video from YouTube to be summarized, fetch the subscription, preserve it as a PDF file. Firstly, the preprocessing step such as tokenizer, stop words abstraction, and stemming were executed. Secondly, utilizing a statistical approach to frequency-inverse document frequency, the algorithm defines the features of a document and culls words and a verbalization containing the features. Finally, the summary is output in the section of the interface to show the important sentences in the documents. For the evaluation of our research, we have used recall-oriented understudy of the gisting evaluation (Rouge) to equate a summary method with a collection of comparison summaries (human summaries).

There are various measures of the rouge for different lengths, some of which are [23], [24]:

- ROUGE-N: N-grams, including ROUGE-1 (unigrams that used for measuring text similarity [25]), ROUGE-2 (bigrams), and so on to n-grams, overlap between the proposed system and comparison summaries.
- ROUGE-L: the advantage of LCS (longest common) is that because it includes immediately the longest common n-grams, it does not need a predefined n-gram length (using the LCS to measure the longer matching word chain).
- ROUGE-SU: bigrams with a maximum skip distance of 4 between bigrams.

Our work is based on ROUGE 2.0 (Java implementation of ROUGE). It uses a synonym dictionary to allow capturing of semantic overlap and evaluating of specific topics or a subset of content based on CNN-dailymail-master data set. The result is explained in Table 3, the quality and readability evaluation outcomes are high in contrast with the human summaries.

To evaluate the efficacy of our method, we compare it to the findings in [26], which included many traditional methodologies and used the CNN/dailymail data set. A high assessment score is assigned to a good summary. Including various traditional methods and using CNN/dailymail data set. High score evaluation measure is assign to a good summary. NN-SE and SummaRuNNer are examples of extractive models, SummaRuNNer-abs, on the other hand, is an extractive model that is similar to SummaRuNNer then is trained straight on abstractive summaries [27], [28]. Furthermore, several baselines were compared, including the baselines described in [27], Although only 500 samples of the test set have been tested. Listener registration (LREG) is a linear regression technique based on functionality. The word extraction model NN-WE limits the production of words from the source document [27]. Lead-3 is a powerful extractive baseline that summarizes the first three phrases. In the sequence-to-sequence architecture neural employs a graphic-based attention mechanism for solving the difficult job of abstract document resuming [28]. The neural network abstractive baseline (NN-ABS) is a basic hierarchical variation of the neural abstractive baseline (NAB) [29]. The findings in Table 4 reveal that our strategy is significantly better and has a higher score than others.

Table 3. Evaluation result using Rouge-1, Rouge-2, Rouge-L and Rouge-SU

DATA SET		Aston Villa manager Tim Sherwood	Blackburn Rovers	Blind Woman	Gabriel set	Glenn Maxwell	Hook line and stinker	Liverpool without Steven Gerrard	Scotland full-back Stuart Hogg
ROUGE-1	recall	0.5913	0.4219	0.8936	0.8442	0.7182	0.4790	0.4575	0.4188
	Precision	0.9465	0.8439	0.9843	0.9824	0.8293	0.8322	0.305	0.7371
	f-score	0.7278	0.5626	0.9368	0.9081	0.7698	0.6080	0.366	0.5341
ROUGE-2	recall	0.5555	0.3558	0.8785	0.8131	0.6458	0.3959	0.2748	0.3289
	Precision	0.8984	0.7246	0.7246	0.9698	0.7707	0.6955	0.1835	0.5872
	f-score	0.6865	0.4773	0.9248	0.8846	0.7027	0.5046	0.2201	0.4217
ROUGE-L	recall	0.1946	0.1102	0.3076	0.1228	0.2582	0.1190	0.1261	0.1214
	Precision	0.2340	0.1415	0.3	0.1196	0.2041	0.1369	0.0585	0.1486
	f-score	0.2125	0.1239	0.3038	0.1212	0.2280	0.1273	0.08	0.1353
ROUGE-SU	recall	0.5466	0.3535	0.8676	0.7908	0.6167	0.3992	0.2676	0.3196
	Precision	0.8936	0.7333	0.9725	0.9677	0.7589	0.7090	0.1789	0.5787
	f-score	0.6783	0.4771	0.9171	0.8703	0.6804	0.5108	0.2145	0.4117

Table 4. Comparison results using rouge recall at variant length

Method	Rouge 1	Rouge 2	Rouge L
LREG	18.5	6.9	10.2
NN-ABS	7.8	1.7	7.1
NN-WE	15.7	6.4	9.8
Lead-3	21.9	7.2	11.6
NN-SE	22.7	8.5	12.5
SummaRuNNer-abs	23.8	9.6	13.3
SummaRuNNer	26.2	10.8	14.4
Graph_based A. Neural	27.4	11.3	15.1
Our method	28.3	22.0	13.9

5. CONCLUSION

The summaries that create automatically may not be as cohesive and smart as the summarization created by humans. The readers, on the other hand, were able to comprehend the produced summaries for the most part. Furthermore, the large number of text documents and the long video on the web introduces to the user summary of every document with more facility to find the desired documents or the suitable video. The automatic summary feature allows the user to quickly obtain a sense of all the material in the document and to identify the papers that are most relevant to the user's requirements.

In this research, we show the automatic summary for the transcription of YouTube long videos. Firstly, we fetch the text from the video, execute the pre-processing steps in the English text, and expound the text summarization process based on extractive type. Then, the algorithm of TF-IDF was used to summarize the transcript of the YouTube video depending on the important sentence in it. Through this experiment, it can be seen that it is a suitable method is used for extractive summary. By using this method, TF-IDF had proven that it is a strong method to produce the value that decides which word inside the text is important. That value assists the program for choose which sentence can be used in the summary. The dataset that used in this study was CNN-dailymail-master dataset.

We can conclude that our method gives a good result compared with the previous method studies. The result of Rouge 1 and Rouge 2 is higher than the rest. Additionally, Rouge L is not the highest one but it is good when compared with the other methods. Finally, if we use a reliable method for summarizing the original document, automatic summaries can be just as useful as human summaries.




REFERENCES

- [1] A. E. B. Ajmal and R. P. Haroon, "Maximal marginal relevance based malayalam text summarization with successive thresholds," *International Journal on Cybernetics and Informatics*, vol. 5, no. 2, pp. 349–356, Apr. 2016, doi: 10.5121/ijci.2016.5237.
- [2] M. Allahyari *et al.*, "Text summarization techniques: A brief survey," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017, doi: 10.14569/ijacsa.2017.081052.
- [3] K. Prudhvi, A. B. Chowdary, P. S. R. Reddy, and P. L. Prasanna, "Text summarization using natural language processing," in *Advances in Intelligent Systems and Computing*, vol. 1171, pp. 535–547, 2021, doi: 10.1007/978-981-15-5400-1_54.
- [4] R. Boorugu and G. Ramesh, "A survey on NLP based text summarization for summarizing product reviews," in *Proceedings of the 2nd International Conference on Inventive Research in Computing Applications, ICIRCA 2020*, Jul. 2020, pp. 352–356, doi: 10.1109/ICIRCA48905.2020.9183355.
- [5] M. H. Maaaloul, I. Keskes, L. H. Belguith, and P. Blache, "Automatic summarization of Arabic texts based on RST technique," in *ICEIS 2010 - Proceedings of the 12th International Conference on Enterprise Information Systems*, 2010, vol. 2 AIDSS, pp. 434–437, doi: 10.5220/0002976104340437.




- [6] I. Keskes, M. M. Boudabous, M. H. Maaloul, and L. H. Belguith, "Étude comparative entre trois approches de résumé automatique de documents arabes," in *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, 2012, pp. 225–238.
- [7] H. Froud, A. Lachkar, and S. A. Ouatik, "Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering hanane," *arXiv preprint arXiv*, 2013, doi: 10.48550/arXiv.1302.1612.
- [8] E. Mahdipour, "Automatic persian text summarizer using simulated annealing and genetic algorithm," *International Journal of Intelligent Information Systems*, vol. 3, no. 6, p. 84, 2014, doi: 10.11648/j.ijis.s.2014030601.26.
- [9] F. Kiyomarsi, "Evaluation of automatic text summarizations based on human summaries," *Procedia - Social and Behavioral Sciences*, vol. 192, pp. 83–91, Jun. 2015, doi: 10.1016/j.sbspro.2015.06.013.
- [10] S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'hommeaux, and R. Ptucha, "Semantic text summarization of long videos," in *Proceedings-2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, 2017, pp. 989–997, doi: 10.1109/WACV.2017.115.
- [11] N. L. Garlic, "Computational text analysis of Youtube video transcripts - Loretta C. Duckworth scholars studio," *sites.temple.edu*, 2019. [Online] <https://sites.temple.edu/tudsc/2019/04/03/computational-text-analysis-of-youtube-video-transcripts/> (accessed Jun. 24, 2021).
- [12] H. Saggion and T. Poibeau, "Automatic text summarization: past, present and future," in *Theory and Applications of Natural Language Processing*, Multi-Sour., Berlin, Heidelberg: Springer, 2013, pp. 3–21.
- [13] C. M. Taskiran, Z. Pizlo, A. Amir, D. Ponceleon, and E. J. Delp, "Automated video program summarization using speech transcripts," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 775–790, Aug. 2006, doi: 10.1109/TMM.2006.876282.
- [14] H. Saggion, S. Teufel, D. Radev, and W. Lam, "Meta-evaluation of summaries in a cross-lingual environment using content-based metrics," in *Proceedings of the 19th international conference on Computational linguistics*, 2002, vol. 1, pp. 1–7, doi: 10.3115/1072228.1072301.
- [15] M. Okumura, T. Fukusima, H. Nanba, and T. Hirao, "Text summarization challenge 2 text summarization evaluation at NTCIR workshop 3," *ACM SIGIR Forum*, vol. 38, no. 1, pp. 29–38, Jul. 2004, doi: 10.1145/986278.986284.
- [16] M. J. Witbrock and V. O. Mittal, "Ultra-summarization: A statistical approach to generating highly condensed non-extractive summaries," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, 1999, pp. 315–316, doi: 10.1145/312624.312748.
- [17] M. S. Ansary, "A hybrid approach for automatic extractive summarization," in *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, Feb. 2021, pp. 11–15, doi: 10.1109/ICICT4SD50815.2021.9396855.
- [18] H. Christian, M. P. Agus, and D. Suhartono, "Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF)," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 7, no. 4, p. 285, Dec. 2016, doi: 10.21512/comtech.v7i4.3746.
- [19] K. M. Hermann *et al.*, "Teaching machines to read and comprehend," in *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems*, Jun. 2015, vol. 1, pp. 1693–1701, doi: 10.5555/2969239.2969428.
- [20] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gülçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings*, 2016, pp. 280–290, doi: 10.18653/v1/k16-1028.
- [21] M. G. Ozsoy, I. Cicekli, and F. N. Alpaslan, "Text summarization of turkish texts using latent semantic analysis," in *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*, 2010, vol. 2, pp. 869–876.
- [22] D. Debnath, R. Das, and P. Pakray, "Extractive single document summarization using multi-objective modified cat swarm optimization approach: ESDES-MCSO," *Neural Computing and Applications*, pp. 1–16, 2021, doi: 10.1007/s00521-021-06337-4.
- [23] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004, pp. 74–81.
- [24] K. Ganesan, "ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks," *arXiv preprint*, 2018, doi: 10.48550/arXiv.1803.01937.
- [25] M. Jabardi and A. S. Hadi, "Twitter fake account detection and classification using ontological engineering and semantic web rule language," *Karbala International Journal of Modern Science*, vol. 6, no. 4, pp. 404–413, 2020, doi: 10.33640/2405-609X.2285.
- [26] J. Tan, X. Wan, and J. Xiao, "Abstractive document summarization with a graph-based attentional neural model," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2017, vol. 1, pp. 1171–1181, doi: 10.18653/v1/P17-1108.
- [27] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016, vol. 1, pp. 484–494, doi: 10.18653/v1/p16-1046.
- [28] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017, pp. 3075–3081.
- [29] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for sentence summarization," in *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 379–389, doi: 10.18653/v1/P17-1108.

BIOGRAPHIES OF AUTHORS






Assist. Lec. Rand Abdulwahid Albeer    B. Sc at University of Kerbala, Collage of Science, Computer Department in 2011 and M.Sc at University of Babylon, College of Information Technology in 2017. She is a lecturer at University of Kerbala, Collage of Computer Science and Information Technology, Computer Department. Here research interests include Petri net, Cryptography and Data mining. She can be contacted at email: rand.a@uokerbala.edu.iq.






Assit. Prof. Huda F. AL-Shahad    member of Computer science and Information Technology College, Kerbala University, Karbala, Iraq. She did get her M.Sc. and B.Sc. from Al-Nahrain University. Her research interest is artificial intelligent, natural language, and data mining. She can be contacted at email: huda.alshahad@uokerbala.edu.iq.



Assist. Prof. Hiba J. Aleqabie    Faculty member of Computer science and Information Technology College, Kerbala University, Karbala, Iraq. She did get her Ph.D. From The University of Babylon.M.sc.and B.Sc. from Al-Nahrain University. Her research interest in multimedia processing and data mining as well as text, twitter and socila networks mining. She can be contacted at email: hiba.jabbar@uokerbala.edu.iq.



Assist. Prof. Dr. Noor D. Al-Shakarchy    B.Sc and M.Sc at University of Technology, Computer science and information systems- information systems Department, Bagdad, Iraq in 2000 and 2003 respectively. She is a lecturer at University of Kerbala, Collage of Computer Science and Information Technology, Computer Department. Here research interests include computer vision, pattern recognition, information security, artificial intelligent and deep neural networks. She can be contacted at email: noor.d@uokerbala.edu.iq.