

Summarizing twitter posts regarding COVID-19 based on n-grams

Noralhuda N. Alabid, Zahraa Naseer

Department of Computer Science, Faculty of Education, University of Kufa, Najaf, Iraq

Article Info

Article history:

Received Nov 8, 2022

Revised Mar 27, 2023

Accepted Apr 2, 2023

Keywords:

COVID-19

Extractive summarization

K-mean clustering

Latent dirichlet allocation

N-grams

ABSTRACT

The COVID-19 pandemic announced by the World Health Organization has disrupted human lives at different scales, including the economy, public health, and people's emotions. Social media databases record huge accumulated information concern this pandemic. Twitter platform is considered one of the most active social media that enable users to tweet in different conversations they are concerned about. The problem arises when tweeters want to search about a specific topic. They can only sort tweets by its recency to understand conversation and not by relevancy. This makes tweeters read through the most tweets to understand what was firstly discussed about the related topic. Some strategies were developed for summarizing tweets but summarizing topics of COVID-19 are still at the beginning. The current research aims to introduce a technique to present a short summary related COVID-19 topics with consuming little time and effort. Thus, summarization task started by clustering topics based on latent dirichlet allocation (LDA) method and K-means clustering and then selected the important sentences to format summarization. The study also compares bigram-based and unigram-based summarization. Different metrics were used to evaluate results and experiments at each stage, and the output of the proposal system was evaluated using ROUGE metrics.

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



Corresponding Author:

Noralhuda N. Alabid

Department of Computer Science, Faculty of Education, University of Kufa

Najaf, Iraq

Email: noralhudan.hadi@uokufa.edu.iq

1. INTRODUCTION

The spreading of using the web platforms and applications has a role in changing the activity of users reaction from inefficient reader to active content inventor. Content is created on different platforms of social media such as Twitter, Facebook and Instagram posts and so on. Also customer reviews on e-commerce websites provide sensitive contents. One of the most popular sources for gathering contents is comments posted on Twitter. This system allows accumulating a lot of comments. Taking a deep view on these dialogues with revealing the real meaning is an impossible task. In most times users tend to follow some of the strategies to read these comments such as reading the newest, oldest, or most popular ones. Thus, context and sense of the diversity of the discussion could be lost. Users may feel that these conversations are full of petty arguments, repeated information, and useless data, but actually these posts provide multiple perspectives and a lot of information and experience that gathered from different sources. Some discussions can be important conversations related to sensitive topics, in which users reveal their opinions, and exchange of experiences. Therefore finding a system that provides an overview, impressions, and correctly analyzes this crowded information can help many organizations. For example, it can also help journalists and politicians to enrich their perceptions and experiences with people's opinions and feelings. Recently, along with the news of COVID-19, it is spreading very fast.

in the social networks. Some of them can be used to determine reliable readings and probabilistics for many important topics related to this epidemic. Subsequently, World Health Organization can use these readings to understand the overview impression of people's opinion and their behaviours on how to prevent infection, deal with infection, and understand impressions about vaccinations and their vital role in reducing infection. With the advent of clusters and topic modeling, it has become possible to focus and examine user opinions and behavior in a comprehensive manner. Roy *et al.* [1] examined twitter data by SeqLDA model to determine who is accused of spreading the Ebola epidemic in West Africa From the point of view of social media users. Several recent studies [2], [3] implemented a technique to analysis twitter posts to study users' emotion regards COVID-19's vaccinations.

Topic modeling (TM) processes allowed finding categorical relationships between documents. Accordingly, several algorithms for text summarization have been developed. Rautray and Chandra [4] provided an efficient algorithm for multi-document summarization. A number of recently researchers [5]-[8] described a work for summarizing based on applying a neural network to create exemplification for sentences. Then, they applied a binary classifier for detecting the importance of these sentences. Other researchers have examined another kind of topic clustering which was K-mean clustering such as [9], [10]. They proposed K-mean clustering for multi-document summarization process and produced a promising accuracy result. While many researches focused on using latent dirichlet allocation (LDA) method to cluster documents. LDA was firstly discovered in 2003 by Blei *et al.* [11]. The main assumption of this method for detecting topics is based on randomly mixers words of documents over latent topics. LDA assumes that documents contain a mixture of topics and the topics themselves consist from a mixture of words. Thus, LDA builds a topic per document and word per topic. In normal LDA, latent topic is characterized by words that frequently co-occur. So, the words that are frequently mentioned over multiple documents are classified within the same topic. Yao and Wang [12], authors conducted a study to summarize twitter data based on LDA. Wang *et al.* [13] summarized online reviews regards two of competing products by using LDA. Dieng and Blei [14] presented a system for two topic models based on word embedding. At the first period of pandemic utilized topic model to analysis a small dataset of tweets [15]. They distinguished different aspects of people's awareness as well as their negative and positive emotion on spreading of the virus. Since the new emergency of COVID-19, a limited literature was available on topic-specific modeling of COVID-19. Xue *et al.* [16] implemented the basic LDA to analyze 1.9 million Tweets and examined the psychological impact of this disease on human. Recently study done by Bogdanowicz and Guan [17] for tracking the dynamic topic modeling of COVID-19 based on using Seq LDA.

The domain of summarization is diverse in the scenario of different applications. One of the earliest approach focuses on summarizing a single document [18] then extended to summarize multiple documents [19], email thread [20], recognize the specific online arguments and dialogues [21], [22], and timeline summarization [23]. Comments summarization in social media concentrates on determining which posts are most relevant to a particular topic. It allows understanding the hidden events and feelings about various incidents. With this domain many work was implemented to summarize blog [24], consumer products reviews [25]. El-fishawy *et al.* [26], machine learning based summarization is developed for Twitter posts. Other authors considered the unigram based mode is the baseline for summarization [27]. While, some other researchers used each of n-grams comprehensive with syntax information for detecting a specific topic [28]-[30].

The basic idea behind text summarization was by filtering, grouping similar information and summarizing contents [31]. When these groups are detected they are summarized by one of summarization approach which are extractive or abstractive summarization method. The extractive approach selects an important sentence from the data in the source group without alteration and introduces it as the summary of that group. For detecting the important terms in each group, some of methods such as term frequency inverse document frequency (TF-IDF), mutual information (MI) and pageRank can be used to determine the weight of these terms [32]. Carmel *et al.* [33] suggested a standardized ranked framework for ranking multi-objective comment. Typically, they supposed that comments can be ranked by different criteria depending on modernity, sorting, or user's profile. The other approach, abstractive methods bases on modifying data in the group to an analogous form which identifies summarization. This conversion needs to an extensive, complex natural language processing for paraphrasing sentences.

There are some critical issues related using LDA. It is capable of summarizing multiple documents only as a whole and cannot group them by topic. LDA generates topics based on the distribution of words in the documents, without considering the coherence or logical flow of ideas. As a result, it may produce a disorganized summary output. Therefore, we are motivated to aggregate LDA and K-mean. So that, K-mean cluster is implemented on each topic of LDA to classify subjects and select summary. As well as summarizing tweets and summarizing news documents are two distinct tasks that require different approaches and techniques. This is because of tweets and documents differ significantly in terms of their content, purpose, and length. Tweets are short and concise messages that are limited to 280 characters, making it necessary to convey the main idea in a limited space. Therefore, we relied on n-grams to identify the main ideas and structure the summary in a concise and readable manner. This paper presents a system for extractive summarization of

Twitter posts that have been considered multi-topic. It focuses on the problem of how to produce a brief summary for tweets related to COVID-19 in less time and effort. We investigated bigrams and unigrams analysis based the topic modelling. Afterward, readers should be capable of assessing the primary sub-topics that were addressed and the viewpoints held by individuals regarding these subjects. Each LDA as well as K-mean clustering approaches has been used for grouping topics. Once grouped, all comments in each cluster is ranking by TF-IDF to given a score that determines the efficiency of this phrase to create summaration. The stages of proposed strategy was evaluated by using both of topic coherence, Silhouette measure, and the totally result was compared with the well known human-written summaries. This is through implements ROUGE matrix. This article is organized as follows: section 1 introduces introduction and overview on relevant work. Section 2 presents the methodology for Tweet acquisition, preprocessing, clustering topics, rank them, and evaluation senarion. Then, results and discussions are stated in section 3. Finally, section 4 outlines the conclusion.

2. METHOD

2.1. Data collection

Twitter platform offers an open-source for gathering public opinions and thoughts that contain keywords of interest for data analytics. Several twitter application programming interfaces (API) services are available such as 30-days Sandbox, and full archive data (tweets from twitter archive), and standard streaming API. We choosed to steam tweets through Twitter's API endpoint. This option allows to collect tweets based on pre-define conditions and parameters and store them in an internal database. Since we are only interested in tweets related to Covid-19 we used set of keywords to filter tweets related COVID-19 which were ["coronavirus", "corona", "Covid19", "Covid", "Covid-19", "face mask", "pandemic", "outbreak", "infect", "vaccine", "omicron"]. Between March 1 and July 31, 2022 we archived around 100,000 tweets. Also, because we collected tweets at unconnected times, the Twitter interface API may re-collected the same tweets with each connection. So, we created a dictionary to test whether tweet_id is previously streamed in our internal database preprocessing.

2.2. Data structure

By default, all twitter API outputs are encoded tweets using JavaScript Object Notation (JSON). JSON is a collection of name-value pairs, which is named attributes and related values. These attributes and their values are used to describe a tweet object. The main component of tweets object is tweets and twitters users. Each tweet has different attributes associated with it such as: author, a message, a unique ID, and time of posted. Each of twitter users has Twitter name, unique ID, location, number of follows.

2.3. Cleaing text and removing non-English words

It was found that the most effective way to perform an in-depth analysis of this volume of tweets and raise the accuracy value is by reducing the dataset and extracting qualitative samples. Therefore it is important to eliminate non-semantic content such as mentions and links. This can be implemented by removing popular, identical, and duplicate content. Thus, we utilized natural language toolkit (NLTK) on python language pipeline and Gensim's library to apply set of data filtration techniques.

Since hashtags mostly contain efficient phrases as well as popular words, we used a split function on hashtags such that the format of "#Covid19Vaccine" will be transformed to be " Covid 19 Vaccine". Also we used the standard NLP libraries to filter out all of non-English words. For cleaning, we applied Gensim's library on python programming software to remove stopwords, tokenize our texts, split text based on whitespace, and lemmatizing word. Also and by using a simple code, we removed words less than 2 characters.

2.4. Generating n-grams

N-grams are sequences of items that frequently appeared next to each other in most documents. These pairs have different names based on size of n-grams. The n-grams of size two is named bigram, size three is trigram, and size four is quadgram. It may be useful to recognize these pairs, as they can make more logical meaning than their individual formations. For example, taking the sentence "coronavirus can spread through coughs", we can extract the word bigram in this form", "coronavirus can", "can spread", " spread through", " through coughs". After we produced the bigram tokenization for each tweets, we combined all occurrences of each word pairing in the bigram list into a single token, in same approach that was presented in [3]. This means for example, if phrase "Covid spread" was extracted as a bigram. It will be replaced to be as a single token "Covid_spread". The most 10 top common unigrams inclusive the frequencies of appearances are shown in Figure 1.

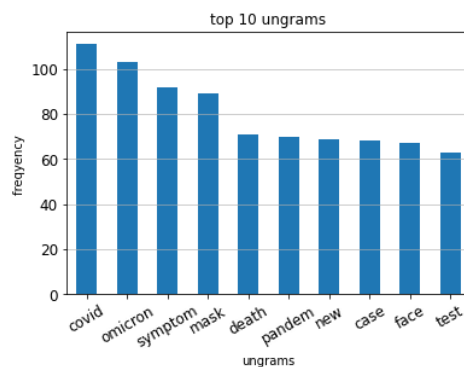


Figure 1. Most common unigrams

2.5. Features selection

To deal with topic models techniques, we needed to convert our tokenized tweets into vector representations with m rows (tweet samples) and n columns (size of vocabularies). We trained TfidfVectorizer model from sklearn.feature_extraction library to code each of the tokenizer tweets and produce TF-IDF n-grams features. TfidfVectorizer converted each tweet into integer or floating-point value. Additionally, TfidfVectorizer algorithm is not prone by the presence of stop words so missing them while preprocessing will not dominant on results.

2.6. Clustering

In order to obtain best results for clustering comments into topics we built a cooperative model combined the efficiency works of the topic modeling represented by LDA and clustering technique demonstrate by K-mean model. LDA topic modeling is a generative probabilistic mode produced to identify latent topics included in a text document. The Gensim library 3.4.0 on python platform was utilized in this research for implement LDA clustering. The supposition is there are two inputs have to feed the LDA classifier. One of them is the vector of token tweets which contains related words. In our corpus, set of words relating to COVID-19 pandemic were used. To manipulate the LDA assumption that described data of analysis as mixture of topic and these topics are used to pick up words according their probability. To address this assumption, different hashtags, related to various issues of COVID-19 pandemic, were used for mining tweets. The other input to LDA model was the number of topics. In our research we extracted 20 topics, but with multiple experiments to determine the optimum number of topics, it turned out to be 7. The last output from LDA is a list of all the determined topics as well as a distribution of each topic across each particular document.

One of the critical issues related the field of documents summarization is keeping coherent and removing redundant content. This problem arises more often with summarizing multiple documents than the process of summarizing single documents. This is because of the overlapping information between topics. For this case we applied a clustering method, which was K-means clustering, on each topic to select the important phrases and consider them as the summary of related topic. K-means clustering is one of the simplest clustering algorithms. It requires two major inputs: vector of comments and random points taken from vector space. Each point is allocated with a specific cluster based on measuring its distance to the centroid vector. This clustering was implemented by using sklearn.cluster.Kmeans library 0.23.2. However, utilizing K-means cluster requires defining the number of clusters. Typically, tweets in mostly state tend to be short text with approximately three or four sentences. For this reason, as well as with multiple experiments, the appropriate number of clusters is decided to be two.

2.7. Ranking clusters

After clustering phase, we needed to rank the more informative words in each cluster to define the summarization. Users want to know the main comments that explain the main ideas of all other comments. We needed a way to identify one or more comments in each cluster which provided the best explaining for topics. For that we used TF-IDF for ranking. TF-IDF is a widely used measure to identify the importance terms in documents. This scale can be used to calculate the amount of information each term participates to the cluster. TF is the frequency of words in the comments of cluster. IDF measure normalizes its ratio by comparing the number of the related words with total number of words in that cluster. It measures the amount of information that term provides to a cluster. This matrix specifically used to measure the importance of rare words by calculating the logarithm of total number of clusters divided by the number of cluster that the related word

appears in. An average score of TF-IDF is computed for each comment in each cluster. Typically, this approach prefers short comment with a limited number of important terms.

2.8. Evaluation

For evaluating the process of topic modeling, we measured the quality of LDA model by using topic coherence matrixes. It is a harmonic mean that receives two inputs: topics and the reference dataset to outputs a single value defines the overall topic coherence. The concept of topic coherence includes a set of measures to assess the semantic coherence of the topics that the model infers. Thus, we implemented each of c_v and u_{mass} coherence score. The c_v measure builds content vectors of items using their iterations, then it is normalized scores by using each of normalized pointwise mutual information (NPMI) and cosine similarity. u_{mass} scores is applied to calculate how often the pairs of words are frequently appear together in a document. We also utilized form c_v matrix to find the optimal number of topics. Figure 2 shows the optimal number of topics for the corpus, with Figure 2(a) demonstrating the results for bigram models, and Figure 2(b) providing an explanation of the ideal number of topics for unigram models. For assessing how well the K-mean clustering was implemented with topics of LDA, we executed Silhouette matrix. For each cluster, the Silhouette result is calculated by measuring the distances of data points in one cluster to these points located in other cluster. Human-produced summary procedure was utilized to create a gold standard to evaluate the automatically-produced summary. Three participants were asked to summarize tweets, and other three to evaluate these manual summaries. Consequently, we used one of the most fundamental measures for summary evaluation which is ROUGE measure. It includes number of measures to automatically determine the quality of summarization, such as Rouge-N that bases on counting the number of the overlapping n-gram, Rouge-L that measures the overlapping between longest common subsequences to formed summary. In addition, Line [27] concluded that each of ROUGE-1, ROUGE-2, ROUGE-S4, and ROUGE-SU4 scores perform efficient evaluations especially when removing stop words. These scores were calculated for each comment in our corpus after stemming items and removing stop words.

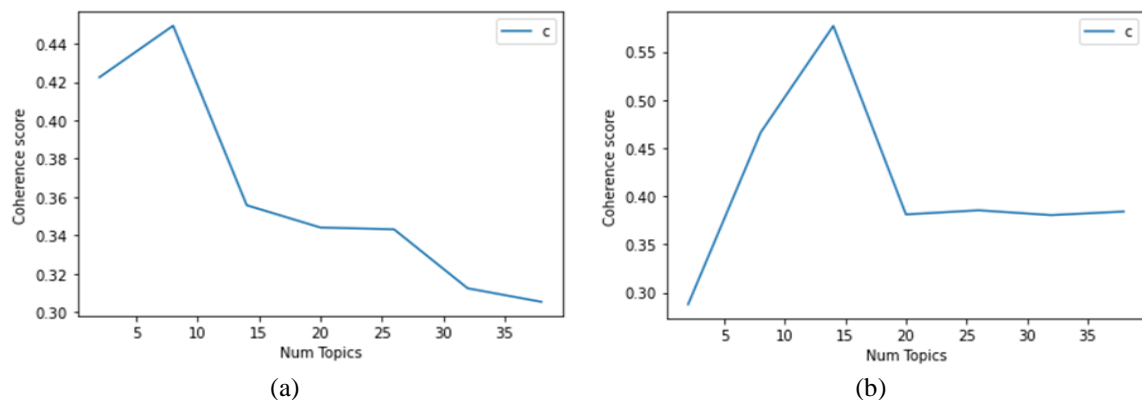


Figure 2. LDA coherence scores (c_v score) for varied numbers of topics (a) bigram model and (b) unigram

3. RESULTS AND DISCUSSION

In this section we presented the results of our proposed system of bigrams model and unigrams model. Table 1 represents the results of coherence value based on different number of topics for LDA models that was evaluated using each of c_v and u_{mass} metrics. The LDA model was tested with different topic values starting from 1 to 20, for each of bigram and unigram sequences. With the c_v score, the maximum value denotes to the case of the optimal coherence of the topic and in the state of u_{mass} score, the value close to zero denotes the highest topic coherence. With bigram based model, the highest coherence value of c_v was estimated to be 0.6 when the number of topics was 7 with preprocessing corpus. While the topic coherence values that was generated by LDA model using u_{mass} metric was 0.2 when the topic number was 15. This result was close to u_{mass} value when the number of topics was 7. In case of unigram model, the maximum value of c_v was estimated with topic number 14 to be 0.54 while the optimal u_{mass} value was detected as 1.37. It can be seen from Table 1 that the LDA modeling with 7 topics presented the best appropriate cluster distribution than the modeling LDA with 15 topics for bigrams sequences. Also, the best results of unigrams combination were detected with 14 topics. However, despite there were not enough differences in results, a definitive conclusion were conducted by looking at the optimal results in both topics number 7 and 14.

Table 1. The performance of LDA with different setting

System	Number of Topic	c_v	u_mass
LDA trained bi-gram with preprocessing	7	0.46	0.18
LDA trained bigram without preprocessing	7	0.40	1.9
LDA trained bigram with preprocessing	15	0.35	0.2
LDA trained bigram without preprocessing	15	0.32	3.8
LDA trained unigram with preprocessing	20	0.39	0.39
LDA trained unigram without preprocessing	20	0.32	4.23
LDA trained unigram with preprocessing	14	0.57	0.87
LDA trained unigram without preprocessing	14	0.48	3.39

The next set of testing involved various experiments to determine the best scenario of combination between number of topics of LDA and number of cluster that used in K-mean algorithm. Different results were conducted by changing the number of topics and fixing number of clusters, see Table 2. It can be extracted that the best results was when clustering LDA with 7 topics in bigram based model and 14 topics in unigram based model.

Table 2. Diversity of values for different number of LDA topics

System	Number of topic	c_v
Clustering the 7 topics of LDA and 2 clusters	Bigram	0.158
Clustering the 15 topics of LDA and 2 clusters		0.116
Clustering the 14 topics of LDA and 2 clusters	Unigram	0.033
Clustering the 20 topics of LDA and 2 clusters		0.030

The following step was how to select the appropriate phrases to form summation. The TF-IDF ranking algorithm was used to choose the top ranked phrases to identify summary. As a baseline, 3 top phrases were extracted from each cluster to conclude summary.

Table 3 shows the performances of the proposed system with different setting. To evaluate the obtained summary, various score of ROUGE matrix were applied which are (ROUGE-1, ROUGE-2, ROUGE-S4, ROUGE-SU4) as explained in Table 3. The outperforming of unigram on bigrams model could be because of the effects of compositional bigrams are much mixed. Consider the following phrases for example COVID-19 symptom and corona symptoms which are both refer to the same topic. By representing these phrases to be a single phrase in the form COVID-19 symptom and corona-symptoms, the ability of the model to directly model the association between these phrases is retracted. However, ROUGE metrics also have a limitation. This is because it depends on tracking the overlapping phrases between both machine-extractive summaries and human- summaries. In fact the language of the automated summaries is tended to be different from the human-written summaries. Thus, the results of the ROUGE matrices cannot be highly relied upon to evaluate abstracts. For this case, three experts were asked to assess and process differences between human-written summaries and the original documents to reduce the error rate. Another issue with bigram model is that this model may have a harder time to fit dataset due to generation of a large number of vocabularies.

Table 3. Average value of of ROUGE matrices for different number of LDA topics

System	R-1			R-2			R-S4		R-SU4			
	F	R	P	F	F	R	F	R	F	F	R	
Ranking bigram phrase from each cluster (3 top unit)	0.08	0.12	0.09	0.23	0.19	0.17	0.21	0.02	0.03	0.42	0.05	0.09
Ranking bigram phrase from each cluster (5 top unit)	0.06	0.1	0.05	0.21	0.18	0.15	0.21	0.18	0.15	0.29	0.05	0.09
Ranking unigram phrase from each cluster (3top unit)	0.25	0.20	0.19	0.07	0.01	0.04	0.07	0.01	0.04	0.48	0.04	0.07
Ranking unigram phrase from each cluster (5 top unit)	0.26	0.21	0.18	0.04	0.03	0.03	0.04	0.03	0.03	0.4	0.07	0.12

4. CONCLUSION

In this paper we examined the problem of summarizing comments over a collection of Twitter comments specifically regarding the COVID-19 pandemic. We demonstrated a robust technical solution that expanded the entire data processing line, from data acquisition to topic summarizing, covering the entire operations of preprocessing, vectorization, clustering and ranking. Our approach is differentiated in scope, utilizing LDA to interpret the emerging topics about COVID-19 at a scale that few research have been done. An important challenge is how to resolve the overlap of topics between multiple comments. Therefore, to solve

this obstacle, clustering topics were used by allowing K-mean clustering to group similar phrases. Next step was to implement TF-IDF ranking algorithm based on context similarity to define contents which will be introduced as the summarization of comment. By using c_v Topic coherence score, we could determine the optimal number of topics for training LDA. The Silhouette metric was applied to measure the range of convergence of data point spread around the center points in each cluster. The efficiency of the system in forming summarization was identified by using ROUGE metrics. However, this measure is not the efficient matrix to evaluate extractive summary. This is because ROUGE bases on calculate the overlapping between words and phrases in machine-written summary against human-written summary. Thus, this is still a challenge task for NLP researchers. It could be possible to consult some of linguistic expert to evaluate or even forming the human-written summaries. We are optimistic in the future to expand the size of dataset, use semantic clustering to group topics, and create abstractive summarization instead of extractive summarization.




REFERENCES

- [1] M. Roy, N. Moreau, C. Rousseau, A. Mercier, A. Wilson, and L. Atlani-Duault, "Ebola and localized blame on social media : analysis of Twitter and Facebook conversations during," *Culture, Medicine, and Psychiatry*, vol. 44, no. 1, pp. 56–79, 2020, doi: 10.1007/s11013-019-09635-8.
- [2] N. Alabid and Z. Dalaf, "Sentiment analysis of Twitter posts related to the COVID-19 vaccines," *Indonesian Journal of Electrical Engineering and Computer Science (IJECCS)*, vol. 24, no. 3, pp. 1727–1734, 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.
- [3] J. H. Lau, T. Baldwin, and D. Newman, "On collocations and topic models," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 10, no. 3, pp. 1-14, 2013, doi: 10.1145/2483969.2483972 ACM.
- [4] R. Rautray and R. Chandra, "An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA," *Applied Computer Informatics*, vol. 14, no. 2, pp. 134–144, 2018, doi: 10.1016/j.aci.2017.05.003.
- [5] X. Zhang, M. Lapata, F. Wei, and M. Zhou, "Neural latent extractive document summarization," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 779–784, doi: 10.48550/arXiv.1808.07187.
- [6] E. Crawford and J. C. K. Cheung, "Extractive : summarization as a contextual bandit," in *Proceedings of the EMNLP Conference*, 2019, pp. 3739–3748, doi: 10.18653/v1/D18-1409.
- [7] J. Cheng and M. Lapata, "Neural summarization by extracting sentences and words," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 484–494, doi: 10.18653/v1/P16-1046.
- [8] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural abstractive text summarization with sequence-to-sequence models," *ACM/IMS Transactions on Data Science*, vol. 2, no. 1, pp. 1–37, 2020, doi: 10.1145/3419106.
- [9] D. Abuaiadah, "Using bisect K-means clustering technique in the analysis of arabic documents," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 15, no. 3, pp. 1–13, 2016, doi:10.1145/2812809.
- [10] U. Behera and S. Gupta, "An approach of categorization and summarization of news using topic modeling," in *12th International Conference on Cloud Computing, Data Science and Engineering*, 2022, pp. 470–476, doi: 10.1109/Confluence52989.2022.9734216.
- [11] D. M. Blei, A. Y. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning*, vol. 3, pp. 993–1022, 2003.
- [12] F. Yao and Y. Wang, "Tracking urban geo-topics based on dynamic topic model," *Computers, Environment and Urban Systems*, no. September, p. 101419, 2019, doi: 10.1016/j.compenvurbsys.2019.101419.
- [13] W. Wang, Y. Feng, and W. Dai, "Topic analysis of online reviews for two competitive products using latent dirichlet allocation," *Electronic Commerce Research and Applications*, no. 18, 2018, doi: 10.1016/j.elerap.2018.04.003.
- [14] A. B. Dieng and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020, doi: 10.1162/tacl_a_00325.
- [15] S. Boon-itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on Twitter : sentiment analysis and topic modeling study," *JMIR Public Health and Surveillance*, vol. 6, no. 4, pp. 1–17, 2020, doi: 10.2196/21978.
- [16] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic : using latent dirichlet allocation for topic modeling on Twitter," *PLOS ONE*, vol. 15, no. 9, pp. 1–12, 2020, doi: 10.1371/journal.pone.0239441.
- [17] A. Bogdanowicz and C. Guan, "Dynamic topic modeling of twitter data during the COVID-19 pandemic," *PLOS ONE*, vol. 17, no. 5, pp. 1–22, 2022, doi: 10.1371/journal.pone.0268669.
- [18] Y. Liu, I. Titov, and M. Lapata, "Single document summarization as tree induction," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, 2019, pp. 1745–1755, doi: 10.18653/v1/N19-1173.
- [19] M. T. Nayeem and T. A. Fuad, "Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1191–1204.
- [20] A. Khan *et al.*, "Sentence embedding based semantic clustering approach for discussion thread summarization," *Hindawi*, vol. 2020, p. 11, 2020, doi: 10.1155/2020/4750871.
- [21] A. Misra, P. Anand, J. F. Tree, and M. Walker, "Using summarization to discover argument facets in online ideological dialog," in *Human Language Technologies: The 2015 Annual Conference of the North American*, 2015, pp. 430–440, doi: 10.3115/v1/N15-1046.
- [22] R. Rameshkumar and P. Bailey, "Storytelling with dialogue : a critical role dungeons and dragons dataset," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5121–5134. doi: 10.18653/v1/2020.acl-main.459.
- [23] S. Zhang, A. Celikyilmaz, J. Gao, and M. Bansal, "EMAILSUM: abstractive email thread summarization," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 6895–6909, doi: 10.18653/v1/2021.acl-long.537.
- [24] H. Xu, H. Yuan, B. Ma, "Where to go and what to play : Towards summarizing popular information from massive tourism blogs," *Big Social Data Special Issue*, 2015, doi: 10.1177/0165551515603323.
- [25] H. Hu and Y. Chen, "A novel approach to rate and summarize online reviews according to user-specified aspects," *Journal of Electronic Commerce Research*, vol. 17, pp. 132–152, 2016.




- [26] N. El-fishawy, A. Hamouda, G. M. Attiya, and M. Atef, "Arabic summarization in Twitter social network," *Ain Shams Engineering Journal*, vol. 5, pp. 411–420, 2013, doi: 10.1016/j.asej.2013.11.002.
- [27] C. Llewellyn, C. Grover, and J. Oberlander, "Summarizing newspaper comments," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2015, pp. 599–602, doi: 10.1609/icwsm.v8i1.14575.
- [28] S. Gilda, "Notice of violation of IEEE publication principles: evaluating machine learning algorithms for fake news detection," *2017 IEEE 15th Student Conference on Research and Development*, 2017, pp. 110–115, doi: 10.1109/SCORED.2017.8305411.
- [29] K. Conroy, L. Victoria, L. Rubin, and J. Niall, "Automatic deception detection : methods for finding fake news," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, 2015, doi: 10.1002/ptra2.2015.145052010082.
- [30] M. Junction, "Fake news detection with semantic features and text mining," *International Journal on Natural Language Computing*, vol. 8, no. 3, pp. 17–22, 2019, doi: 10.5121/ijnlc.2019.8302.
- [31] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, and Y. Li, "Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, 2019, doi: 10.1007/s11042-018-6894-4.
- [32] E. Khabiri, J. Caverlee, and C. Hsu, "Summarizing user-contributed comments," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media Summarizing*, 2011, pp. 534–537, doi: 10.1609/icwsm.v5i1.14192.
- [33] D. Carmel, L. Lewin-eytan, A. Lazerson, and L. L. Eytan, "Multi-objective ranking optimization for product search using stochastic label aggregation," in *Proceedings of Web Conference*, 2020, pp. 373–383, doi: 10.1145/3366423.3380122.

BIOGRAPHIES OF AUTHORS



Noralhuda N. Alabid    works as a lecturer and researcher in the Department of Computer Science at the Faculty of Education, University of Kufa (UOK), located in Najaf, Iraq. She has been employed at UOK since 2012 and holds an M.S. in Advanced Computer Science. She obtained her undergraduate degree from the Department of Computer Science at the Faculty of Science, University of Sheffield in the UK. Her academic background is in computer science, with particular expertise in image processing, bioinformatics, natural language processing, text processing, and medical statistics. She can be contacted at email: noralhudan.hadi@uokufa.edu.iq.



Zahraa Naseer    graduated from the Computer Science and Mathematics Department at Kufa University in Iraq with a B.Sc. degree in computer science in 2015, and received her M.Sc. degree in computer science from the same institution in 2018. She is currently employed as an Assistant Teacher at Kufa University, where she focuses on information security, artificial intelligence and natural language processing in her research. She can be contacted at email: zahraan.albakaa@student.uokufa.edu.iq.