






TikTok Video Cluster Analysis Based on Trending Topic

Juhaida Abu Bakar^{1,2} , Nur Azmielia Muhammad Sharimi²,
Mohd Azrul Edzwan Shahril², Nur Syafiqah Azmi², Nor Hazlyna Harun^{2,3} ,
Hapini Awang³ , and Nur Syafiqah Abu Bakar⁴

¹ Data Management and Software Solution Research Lab, School of Computing,
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

juhaida.ab@uum.edu.my

² Data Science Research Lab, School of Computing, Universiti Utara Malaysia, 06010 Sintok,
Kedah, Malaysia

³ Institute for Advanced and Smart Digital Opportunities, School of Computing,
Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

⁴ Faculty of Medicine and Health Sciences, Universiti Sains Islam Malaysia, 71800 Nilai,
Negeri Sembilan, Malaysia

Abstract. TikTok is a popular social networking application that offers trend research and is a valuable source for users. However, this is often misconstrued for content, which may not be suitable for children due to inappropriate content. This study aims to improve user perception of TikTok by using topic modelling and clustering techniques to identify trending topics in TikTok videos. The research uses Latent Dirichlet Allocation (LDA) and *K*-means clustering techniques to enhance the recognition of local and global topics across text documents. The methodology includes data collection, data pre-processing, clustering, topic modelling, and results. Ten subjects associated with trending TikTok videos are displayed using the LDA algorithm, and the generated topics are used to produce an Inter-topic Distance Map. The method's effectiveness is evaluated using log-likelihood score and perplexity measurements. It has a log-likelihood score of 5579 and a perplexity score of 287. A good model is one with a higher log-likelihood and lower perplexity. The study extracts popular TikTok topics using both the LDA topic modelling technique and the *K*-means clustering algorithm.

Keywords: social networking application · topic modelling · clustering analysis · latent dirichlet allocation · tiktok

1 Introduction

Trending analysis is the practice of gathering data and attempting to identify patterns or trends in data. It can also determine whether an organization's objectives have been met. Nowadays, Tik Tok is a new source and one of the most interesting and useful sources of information in social networking applications involved in trend analysis. TikTok was the most common application that grew faster and attracted 1.5 billion active users [1] of the

hundreds of millions of users from children and young adults. This application lets users create short video content, sharing 15-s and up to 60-s videos on any topic. Influencer Marketing Hub said Tik Tok was formerly known as Musical.ly until ByteDance, a Chinese company, took over the app, and users were transferred to Tik Tok.

TikTok is a popular app among children and adults, often viewed as having disadvantages rather than advantages [2]. However, it offers interesting content such as business tips, cooking tips, and many other types of information. The study aims to classify video content into categories to improve user perception of TikTok. Users can better understand its features and avoid inappropriate content by analyzing the content.

The primary contribution of this study is the use of clustering and topic modelling methods to uncover hot topics from Tik Tok videos. The goal of the effort was to improve the recognition of local topics within a single document and a group of global topics across a series of text documents using Latent Dirichlet Allocation (LDA) and K-Means clustering algorithms. The remainder of the document is structured as follows: Sect. 2 includes a literature review; Sect. 3 describes data preparation and methodology; Sect. 4 illustrates experiments and results; and Sect. 5 concludes the paper.

2 Related Work

A topic modelling experiment based on user comments on social media is shown in the study [3]. The author conducted an experiment using two datasets from Yahoo and Tokyo Electric Power Company (TEPCO), which covered the most popular news stories and video streaming comments, respectively. LDA was used to implement topic clustering based on topic modelling. This experiment received 15,000 comments throughout the same period. The results of the modelling are displayed in Fig. 1.

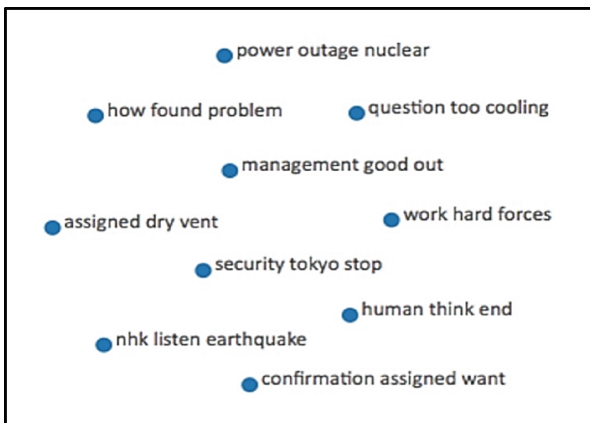


Fig. 1. Modelling's outcome

In the [4] study, Twitter social media was searched for information about the Covid-19 epidemic from March 3rd to March 31st. This author copied ten thousand tweets. The outcomes of clustering based on latent semantic analysis produced more clusters than clustering based on LDA. Because the largest cluster would show the day of trend, it was used as a comparison. The total number of confirmed cases in each nation and worldwide are shown in Fig. 2 [4].

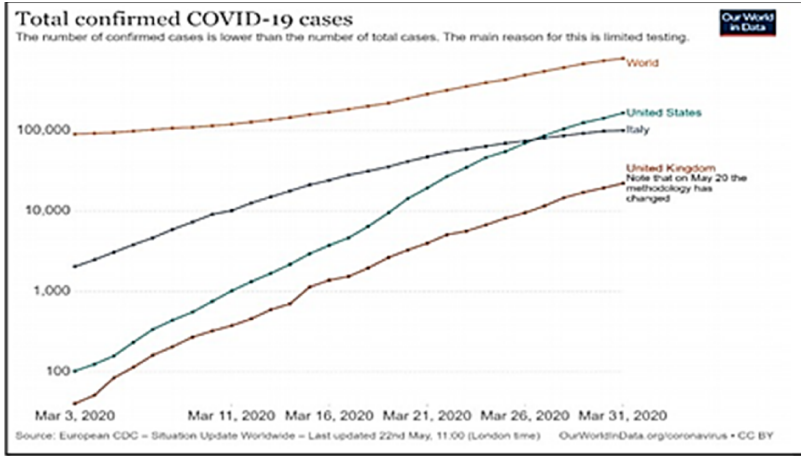


Fig. 2. Number of confirmed cases

This study [5] focuses on the two-step problem of extracting semantically relevant themes and trend analysis of these subjects from a large temporal text corpus utilising an end-to-end unsupervised technique. The author first created word clouds based on the frequency of terms in each cluster of abstract text. As a result, terms that were less prevalent and significant to the cluster were deleted from word clouds. The author generated word clouds based on the TF-IDF scores of phrases belonging to a cluster. The TF-IDF based on a word cloud of four distinct clusters is displayed in Fig. 3 [5].

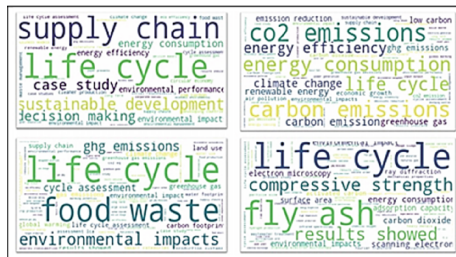


Fig. 3. TF-IDF based on a word cloud of 4 different clusters

The study from [6] has addressed the issue by focusing on Facebook fan pages. The type of special account that has more significance than standard Facebook accounts.

Using a combination of LDA-based topic distribution and post interaction indices, the authors presented a vector for Facebook fan pages. Figure 4 shows the process of LDA to obtain the topic distribution vector for a specific document in a fan pages text corpus.

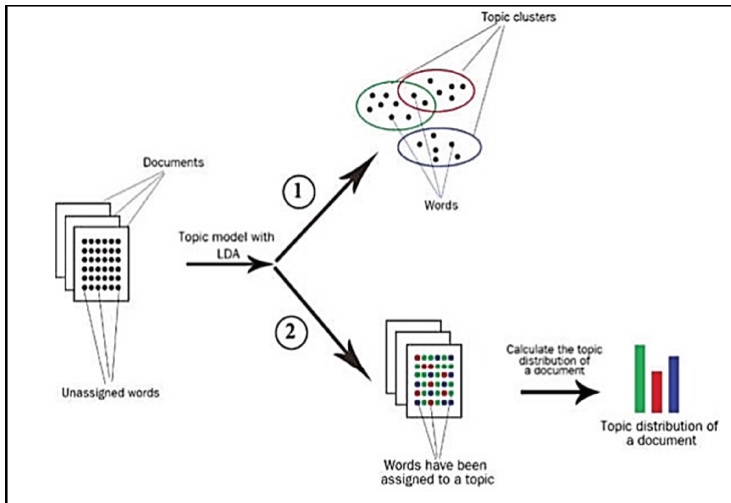


Fig. 4. Process of LDA

In [7], a new spammer classification method is based on the LDA topic model. This technique captures the spamming essence by retrieving global and local data regarding topic distribution patterns. Clustering based on online spam detection is proposed to discover spammers that appear to be posting legal tweets but are difficult to identify using existing spammer categorization methods. The examination of the K-means technique produces an accurate result, and it also identifies spammers on social media.

A past study from [8] explained the process of document clustering by using K-means and K-medoids algorithm (see Table 1.). This study focuses on hundreds of documents collected from Entertainment, Literature, Sports, Political, and Zoology. The authors implement the K-means algorithm on the WEKA tool and K-medoids on the Java platform. Both results of each algorithm are compared to get the best cluster.

Based on Table 1, each cluster defines documents of a particular domain topic. According to the result, the authors conclude that the K-means algorithm is more efficient than the clusters obtained from the K-medoids algorithm [8].

The following word clustering experiment, carried out by [9], focuses on grouping Chinese words using LDA and K-means in accordance with five categories: politics, economics, culture, people's livelihood, and science and technology. The Latent Dirichlet Allocation (LDA) algorithm and the k-means clustering algorithm are combined in a novel approach that is put forth in this work. The highest probability of each topic is picked as the centroids of k-means after some topics are retrieved using LDA. In the final stage, the K-means algorithm is employed to group every word in the text [9]. The authors calculate the K-means centroids using the LDA algorithm findings and utilise the Chinese word similarity calculation method to calculate the distance between the

Table 1. Comparison of K-means and K-medoids algorithm.

<i>COMPARISON</i>			
Algorithm	Cluster number	Number of documents	Efficiency
<i>K-means</i>	0	18	27.78%
	1	3	66.67%
	2	39	25.64%
	3	25	24%
	4	15	26.67%
<i>K-medoids</i>	0	43	18.60%
	1	11	54.15%
	2	7	42.85%
	3	35	14.2%
	4	4	50%

words. The authors calculate the K-means centroids using the LDA algorithm findings and utilise the Chinese word similarity calculation method to calculate the distance between the words. As a result, the suggested algorithm's average similarity is higher than that of the K-means algorithm.

To group Arabic documents, they were analysed using K-means and topic modelling [10]. LDA and K-means clustering methods were coupled in this study's dataset of Arabic text texts. Datasets that represent news stories are obtained from the internet. The experiments choose and apply TF-IDF weighting. This work's authors conclude that using topic modelling techniques during the clustering process enhances the quality of clusters for Arabic text texts [10].

Using the K-means technique, the study in [11] carried out text document clustering. The comparison of K-means clustering and K-Means clustering using Dimension Reduction approaches is covered in this work. The BBCSports dataset, which comprises five categories, including athletics, cricket, football, rugby, and tennis, is used for the comparison. The authors' evaluation metrics include accuracy, precision, recall, and f-measure. The efficiency of K-means clustering with and without DR methods is seen in the following figure [11]. K-means with information gain DR is superior to K-means clustering without dimension reduction approaches, as shown in Fig. 5.

A study on document feature extraction using LDA was conducted in [12]. The information was collected from websites of Indonesian news media by choosing news categories and saving them as text files. The TF-IDF K-Means methodology and the LDA method were used to compare the document clustering results.

Thus, based on previous studies, trend topics in social networking sites such as Tik Tok are still new in the current study. Hence, the use of topic modeling and clustering techniques to identify trending topics in Tik Tok videos is proposed in this study. In

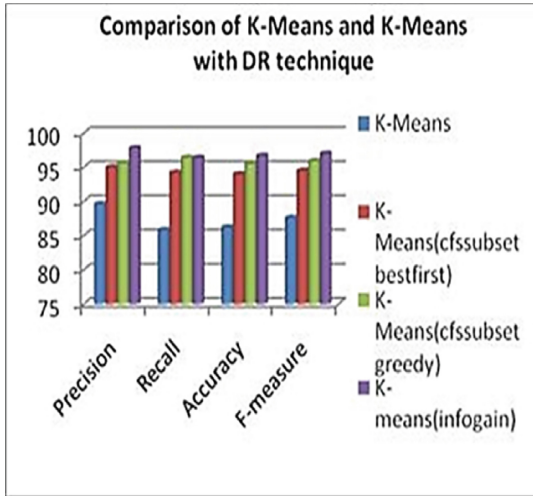


Fig. 5. Comparison of K-Means and K-Means with DR techniques

addition, this research also aims to improve the recognition of local topics in one document and a group of global topics across a collection of text documents using Latent Dirichlet Allocation (LDA) and K-Means clustering techniques.

3 Material and Methods

This study followed the four steps of trending topic analysis of TikTok videos, as shown in Fig. 6: data collection, preprocessing, clustering, and topic modelling algorithms and results. The following subsections clarify each step in detail.

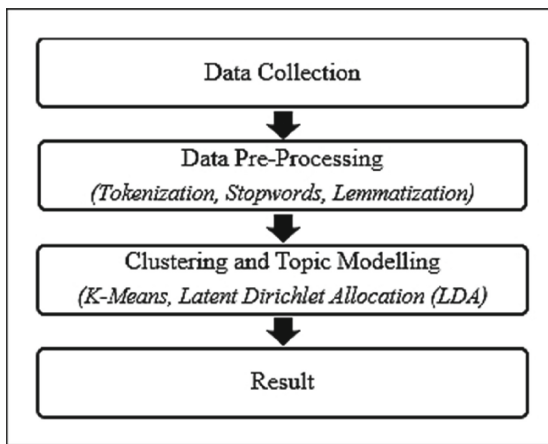


Fig. 6. Steps of Trending Topics Analysis of Tik Tok Video

3.1 Data Collection

This work mined Tik Tok data between April 25th and May 11th, 2021, for our experiment. The data was gathered by randomly extracting Tik Tok metadata such as number of views, number of likes, number of plays, video descriptions, etc. The Tik Tok API was utilized to collect 7552 data elements with 8 attributes at random, and this work used Python language software for data analysis and topic modelling tasks. The data was filtered based on the description's most frequently used terms and hashtags. Figure 7 shows the overview of the dataset.

```
In [9]: dataset.shape
dataset.head()
```

```
Out[9]:
```

	user_name	video_desc	video_length	video_link	n_likes	n_shares	n_comments	n_plays
0	ejaanuar_	Pandai sorok kamera dah sekarang had to do L...	21	https://www.tiktok.com/@ejaanuar_video/69523...	261400	3536	1695	400000
1	khaby.lame	#nPerfavore non fare questo con la Pizza 🍕 #u...	20	https://www.tiktok.com/@khaby.lame/video/69554...	5600000	38100	46900	4130000
2	runningchan3	Bende paling kila takut terjadi lagi. kwinggo.	45	https://www.tiktok.com/@runningchan3/video/695...	48200	2846	1386	324500
3	tra_ahmad	#shouldbema #fyp #esekali #suekhaizhar #rube769	6	https://www.tiktok.com/@tra_ahmad/video/694945...	12500	120	488	271500
6	areppo_	#spm #projekanchorspm #esekali #fyp	15	https://www.tiktok.com/@areppo_video/69555294...	33800	808	504	486800

Fig. 7. Tik Tok Datasets

3.2 Data Preprocessing

The first data preprocessing steps were carried out to eliminate noise from the dataset. It started by removing the following terms as noise from Tik Tok data: text and symbols that are not standard, mentions, hashtags, emoji, null values, and unwanted characters from the Tik Tok descriptions. In addition, to reduce dimensionality and promote the topic coherence, all words were transformed to lowercase. Following that, the preprocessing steps listed below were carried out:

Tokenization: This process changes a string sequence into words, keywords, phrases, symbols, and other items known as tokens. Each sentence is divided into a list of words, with all punctuation and extra characters removed.

Stop Word Removal: Words that are often used but have no effect on the data's meaning were also eliminated from the corpus. The LDA algorithm could only use the most important words as input by deleting such sentences from the text, yielding more accurate results. The work combines English and Malay stop words to remove the mixed term from the corpus. The work also extends the stop word by adding related words, as shown in Fig. 8.

Lemmatization: Lemmatization is the process of converting words to their base word. For instance, 'playing becomes 'play,' 'running becomes 'run,' and 'meeting becomes 'meet'. The benefit is that it can minimize the dictionary's overall number of unique terms. As a result, the number of columns in the document-word matrix will be reduced, resulting in a denser matrix with fewer columns.

```
# NLTK Stop words
from nltk.corpus import stopwords
stop_words = stopwords.words('english') + stopwords.words('malay')
stop_words.extend(['foryoupage', 'fyp', 'for', 'you', 'page',
                  'esekeli', 'tiktok', 'foryou', 'fyp', 'foryourpage', 'stitch', 'viral',
                  'forpage', 'part', 'be', 'go', 'point', 'reply', 'fypage', 'untuk', 'fypviral', 'take', 'help',
                  'video', 'get', 'rise', 'always', 'want', 'comment', 'look', 'link', 'try', 'let', 'see', 'time',
                  'do', 'babi', 'come', 'bro', 'lot', 'choice', 'follow', 'new', 'still', 'eskelil', 'trend', 'kira',
                  'tiktokmalaysia', 'masing'])
```

Fig. 8. Stop word by adding some related words.

3.3 Clustering and Topic Modelling

The K-means clustering and LDA topic modelling methods are employed to the sets used in this work to identify patterns for trending topic discovery. The detail of following algorithm listed below:

K-means Clustering: K-means is a conventional clustering technique that seeks for each observation of each attribute value and compares it to the nearest mean, and it is used to define clusters that are similar to one another [13]. The *k*-means algorithm takes the input parameter *k* and partitions a set of *n*-objects into *k*-clusters. Cluster similarity is calculated using the mean value of the objects in the cluster, which serves as the cluster’s centre of gravity. Steps that this work take for text clustering using *k*-means algorithm is (i) determine the *k* value, (ii) identify the segregation of topic clusters, and (iii) clustering evaluation model using Euclidean distance. Figure 9 shows the *k*-means algorithm.

Algorithm 1: *k*-means algorithm

- 1: Specify the number *k* of clusters to assign
 - 2: Randomly initialize *k* centroids.
 - 3: **repeat**
 - 4: **expectation:** Assign each point to its closest centroid.
 - 5: **maximization:** Compute each cluster's new centroid (mean).
 - 6: **until** The centroid positions do not change.
-

Fig. 9. *k*-means algorithm

Latent Dirichlet Allocation (LDA): Topic modelling is unsupervised natural language processing used to represent text with the help of several topics. A generative model called LDA explains sets of observations made by unseen groups and explains why certain parts of the data are similar [14]. LDA is also a generative probabilistic model that is commonly utilized in the field of information retrieval [15]. Depending on the input settings, the generated topics can be highly generic or very specific [16].

3.4 Experimental Setup

The experiment started by preprocessing the TikTok Metadata by subsetting video descriptions with Python code. A word cloud was used to ensure that no unwanted

content was included in the text corpus created from the video descriptions. Additionally, the word cloud can be utilised to verify the results of the preprocessing technique. The corpus was then represented using the bag of words characteristics representation technique. The bag of word words in the text corpus were therefore subjected to the LDA topic modelling technique.

By creating a topic per document and a word per topic model, LDA algorithms classify documents into themes. A keyword distribution is used to represent each processed video description. The LDA also assumes that the data used for the analysis are a mixture of subjects. The subject will next generate phrases depending on their likelihood matrix. Five topics were initially taken out of the data. However, to determine the ideal number of topics in the corpus, later found to be 10, the best model estimators were used.

4 Experiments and Results

The experiment yielded a word cloud of the Tik Tok description of frequent words, as seen in Fig. 10.

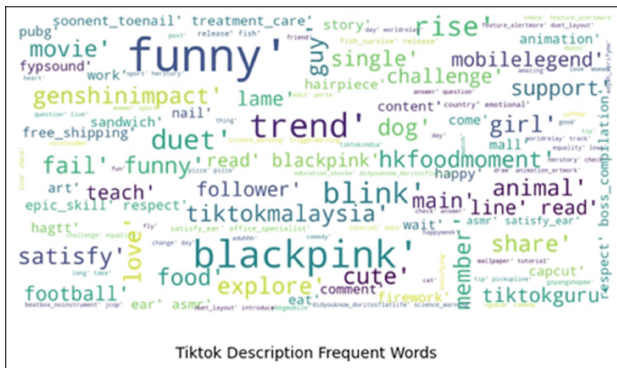


Fig. 10. Word Cloud of The Tik Tok Description

According to the word cloud, the majority of the words throughout the period are related to funny videos, food content, and Korean music videos. Hence, some of their words were included in ‘Food,’ ‘Movie,’ ‘Korean’ and ‘Comedy’ topics. Furthermore, some of the other words are related to the ‘Pets,’ ‘Tutorial’ and ‘Sports’ topics. In addition, this work created an Intertopic Distance Map using the generated subjects. It gives a broad overview of the topics and how they differ from one another. It also allows for a detailed analysis of top phrases associated with each topic. The right panel, as illustrated in Fig. 11, presents a horizontal bar chart representing the top keywords that can be used to understand the topics.

Table 2 shows the results of applying the LDA algorithm to extract ten topics related to trending videos on Tik Tok. Topic 0 is about sharing material regarding drawing and painting. Other common words in this category include ‘artwork,’ ‘animation,’ and so on. Furthermore, as shown in Table 2, topic 1 is about creating Autonomous Sensory

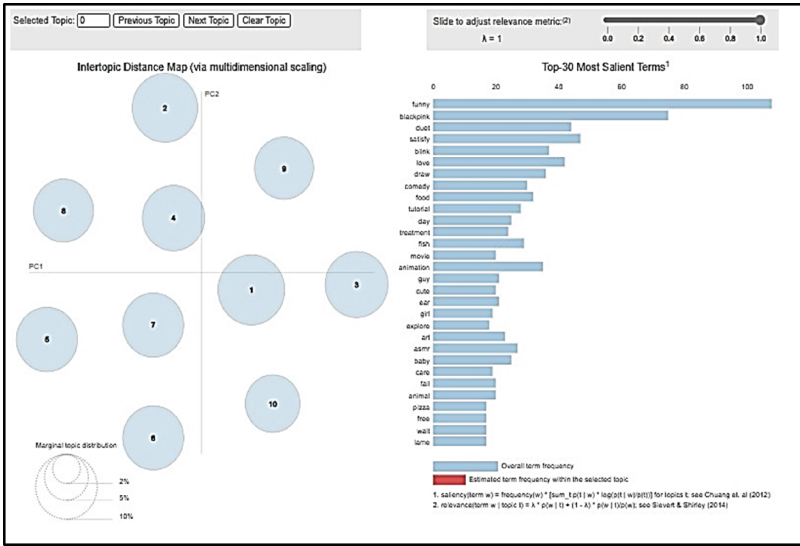


Fig. 11. Inter-topic distance plot of the 10 topics

Meridian Response (ASMR) video material or any enjoyable video. Some popular words associated with this topic include ‘ear,’ ‘foodie,’ ‘food,’ and so on. As a result, this form of ASMR content.

Table 2. Results of Applying The LDA Algorithm

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topics
draw	art	share	hkfoodmom	single	satisfaction	vietmyusa	freepalestine	mobilelegend	drawe	Art/Creative Content
satisfy	ear	asmr	rich	poor	country	teach	office	firework	hairpiece	Asmr/Satisfy Content
treatment	animation	care	soonent	respect	skill	toenail	epic	compilation	boss	Animation/Treatment/Sing Content
food	movie	main	satisfying	cartoon	eat	music	nailsmanicure	popularscience	hoove	Food/Movie/Music Content
blackpink	fish	guy	cute	member	release	read	support	amazing	content	Korean/Game Content
duet	comedy	explore	genshinimpac	good	malaysian	check	feature	male	ngakak	Comedy/Cooking Content
free	wait	lame	shipping	live	answer	ice	question	giant	work	Business/Work Content
funny	blink	fail	animal	challenge	tiktokmalaysia	tiktokguru	football	funnygame	pedicure	Funny/Sports Content
tutorial	day	edit	education	science	didyouknow	trick	change	warning	hagtt	Tutorial/Education/Tricks Content
love	girl	baby	pizza	car	beautiful	sport	woman	sikit	nail	Love Content

This work uses log likelihood score and perplexity measures to assess the algorithm’s performance on the dataset. It acquired a perplexity score of 287 and a log likelihood score of 5579. A model with a greater log likelihood and a lower perplexity is regarded as good. The degree of perplexity is a measure of how well a model predicts a sample. Following that, this model was put to the test by predicting the topic using the newly developed LDA Model. Before predicting the topics, the work uses a random text using the same preprocessing technique. The transformation order and the result are shown in Fig. 12. Figure 13 shows the code to get similar documents.

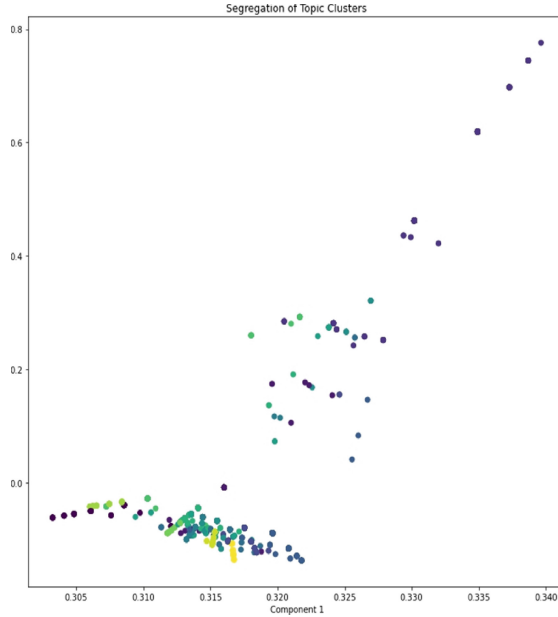


Fig. 12. Segregation of Topic Clust

```
# Get similar documents
mytext = ["i like to eat food and cook"]
doc_ids, docs = similar_documents(text=mytext, doc_topic_probs=lda_output, documents = data, top_n=1, verbose=True)
print('\n')

Topic Keywords: ['movie', 'main', 'satisfying', 'cartoon', 'eat', 'music', 'nailsmanicure', 'popularscience', 'hoove', 'Food/Movie/Music Content']
Topic Prob Scores of text: [[0. 0. 0. 0.7 0. 0. 0. 0. 0.]]
Most Similar Doc's Probs: [[0. 0. 0. 0.7 0. 0. 0. 0. 0.]]
Topics: Food/Movie/Music Content
```

Fig. 13. Similar Documents

In conclusion, this work uses K-Means clustering on the document's topic probability matrix with $k=10$ since the optimal model has 10 clusters to cluster video descriptions that share similar subjects and plots. To depict X and Y columns, singular value decomposition (SVD) is employed. SVD ensures that the two columns collect the most information available for the LDA model. Hence, Euclidean distance and the likelihood score are employed to find similar subjects. The themes that are the most comparable are classified with the shortest distance.

5 Conclusion

This study encompasses the LDA topic modelling technique and the K-means clustering algorithm, both used to extract trending topics from Tik Tok. The study aimed to locate and extract the most popular Tik Tok content. The LDA method obtained a perplexity score of 287 and a log likelihood score of 5579 from 7552 video data. The perplexity score evaluates how well the probability models predict. However, this measure is insufficient

unless accompanied by a manual topic evaluation. Some of the study findings are trendy topics, such as comedy and sports content. As a result, because data is a multi-language combination, future work will need to incorporate the extraction of keywords from the data. This work is also interested in evaluating different topic extraction algorithms, such as a combination of natural language processing and machine learning, and comment analysis [17].

Acknowledgment. This research was supported by the Ministry of Higher Education (MoHE) through the Fundamental Research Grant Scheme (Ref: FRGS/1/2021/ICT02/UMK/02/1). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of MoHE, Malaysia. The authors would also like to extend their gratitude to the Data Management & Software Solution Research Lab, School of Computing, Universiti Utara Malaysia for their generous support in sponsoring the publication of this article.

References

1. Weimann, G., Masri, N.: The virus of hate: Far-right terrorism in cyberspace. International Institute for Counter-Terrorism (2020). https://www.ict.org.il/Article/2528/The_Virus_of_Hate
2. Jing, P., et al.: Listen to social media users: Mining Chinese public perception of automated vehicles after crashes. *Transport. Res. F: Traffic Psychol. Behav.* **93**, 248–265 (2023)
3. Ramamonjisoa, D.: Topic modeling on users's comments. In: 2014 Third ICT International Student Project Conference (ICT-ISPC), pp. 177–180. IEEE (2014)
4. Sheikha, H.: Text mining Twitter social media for Covid-19: Comparing latent semantic analysis and Latent Dirichlet Allocation (2020)
5. Mann, J.K.: *Semantic Topic Modeling and Trend Analysis* (2021)
6. Phan, V.H., Ninh, D.K., Ninh, C.K.: An effective vector representation of Facebook fan pages and its applications. In: Hernes, M., Wojtkiewicz, K., Szczerbicki, E. (eds.) ICCCI 2020. CCIS, vol. 1287, pp. 674–685. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63119-2_55
7. Jose, T., Babu, S.S. Detecting spammers on social network through clustering technique. *J. Ambient Intell. Humanized Comput.*, 1–15 (2019)
8. Rakesh Chandra Balabantaray, C.S.: Document clustering using K- means and K-Medoids. *Elixir Inter. J.*, 16773–16777 (2013)
9. Qiu Lin, X.J.A.: Chinese Word Clustering Method using Latent Dirichlet Allocation and K-means. In: 2nd International Conference on Advances in Computer Science and Engineering , pp. 267–270 (2015)
10. Alhawatat, M., Hegazi, M.: Revisiting k-means and topic modeling, a comparison study to cluster arabic documents. *IEEE Access* **6**, 42740–42749 (2018)
11. Ramkumar, A.S., Nethravathy, R.: Text document clustering using k-means algorithm. *Int. Res. J. Eng. Technol* **6**, 1164–1168 (2019)
12. Prihatini, P.M., Suryawan, I.K., Mandia, I.N. Feature extraction for document text using Latent Dirichlet Allocation. *J. Phys. Conf. Ser.* **953**(1), 012047 (2018)
13. Sapul, M.S C., Aung, T.H., Jiamthaphaksin, R.: Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms. In: 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE) (2017)
14. Tong, Z., & Zhang, H. A text mining research based on LDA topic modelling. In: International Conference on Computer Science, Engineering and Information Technology, pp. 201–210 (2016)

15. Chyi-Kwei Yau, A.L.: Clustering scientific documents with topic modeling . *Scientometrics*, 767–786 (2014)
16. Blair, S.J., Bi, Y., Mulvenna, M.D.: Aggregated topic models for increasing social media topic coherence. *Appl. Intell.* **50**(1), 138–156 (2020)
17. Chumwatana, T.: Comment analysis for product and service satisfaction from Thai customers' review in social network. *J. Inform. Commun. Technolo.* **17**(2), 271–289 (2018)