Journal of Emerging Information Systems and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

TOPIC MODELING OF UNESA LAKE REVIEW ON GOOGLE MAPS USING LATENT DIRICHLET ALLOCATION (LDA) METHOD

Kurrotul Uyun[1], I Kadek Dwi Nuryana[2]

[1, 2] Information Systems Study Program, Faculty of Engineering, Surabaya State University

[1] kurrotul.20093@mhs.unesa.ac.id, [2] dwinuryana@unesa.ac.id

ABSTRACT

Abstract – User reviews on digital platforms hold valuable information that can be used to improve service quality. This study aims to explore the topics that appear in visitor reviews of Lake UNESA based on rating categories with a topic modeling approach using the Latent Dirichlet Allocation (LDA) method. The analysis process follows the stages in the Knowledge Discovery in Databases (KDD) framework, starting from the selection of Google Maps review data, text preprocessing (cleaning, letter normalization, tokenization, word normalization, and stopword removal), and data transformation into bag-of-words representation through bigram-trigram formation and dictionary-corpus creation. Topic modeling is performed using LDA, and the results are evaluated and interpreted through pyLDAvis and wordcloud visualization. Model validation is carried out through Word Intrusion Task and Topic Intrusion Task testing, with accuracy levels of 0.91 and 0.88, respectively. The results show that LDA is able to identify topics optimally. Each rating category produces different topics that represent visitor perceptions of aspects that are not yet available, still k-aspects such as atmosphere, cleanliness, culinary, and facilities. These findings are expected to provide data-based insights to support the development and management of Lake UNESA more effectively.

Keywords: Topic Modeling, Latent Dirichlet Allocation, User Reviews, Google Maps, UNESA Lake, KDD Framework

Article Info:
Article History:
Accepted July 09, 2025
Revised July 10, 2025
Approved July 15, 2025

Corresponding Author
Kurrotul Uyun
Information Systems Study Program, Faculty of Engineering, Surabaya State University
kurrotul.20093@mhs.unesa.ac.id

1. INTRODUCTION

The advancement of information technology has opened up great opportunities for the application of data analysis in various fields, including the education and tourism sectors. One of the rapidly developing approaches is data mining, which is the process of exploring and extracting important information from large-scale data sets . One branch of data mining is text mining, which focuses on text-based data. This technique allows the discovery of hidden patterns in text through the analysis process, and has been widely used to support more effective and targeted decision making [1].

One interesting object to analyze using this approach is UNESA Lake, a public space located in the Universitas Negeri Surabaya environment. Although it has not been officially confirmed as a tourist attraction, this lake has attracted public attention because of its beautiful environment and its multifunctional function as a place to relax, exercise, and do campus activities. This is reflected in the number of visitor reviews on Google Maps which has exceeded 6,000 reviews. The large number of reviews shows that UNESA Lake has great potential to be developed, but has not been accompanied by a systematic and data-based management strategy.

Digital review platforms such as Google Maps store various visitor opinions that can be used as a valuable source of data. Each review, whether in text or rating form, reflects the user's experience and perception of a place. Google Maps itself is a digital mapping service from Google that provides map views, satellite photos, 360° street views (Street View), and open user reviews [2], [3]. To extract information from this large amount of data, text analysis methods such as topic modeling are needed. One of the widely used algorithms is Latent Dirichlet Allocation (LDA), which allows automatic grouping of topics in text based on word distribution [4], [5].

Previous studies have shown that LDA is effective in understanding public opinion through text. For example, Puspita et al. [6] successfully identified popular topics in online news about beauty brands, while Dewi et al. [7] evaluated user comments on a novel web platform. Junaedi et al. [8] showed that LDA can be used to extract patterns in Tokopedia consumer reviews, which are useful in service development strategies. Another study by Tondang et al. [9] also showed that LDA is able to effectively identify and group reviews of BNI, BCA, and BRI applications, strengthening the evidence that this method is very useful for evaluating digital reviews on various platforms. Considering the success of LDA in these various contexts, this approach is appropriate for analyzing reviews about UNESA Lake.

This study aims to group UNESA Lake visitor reviews based on rating categories (positive, neutral, negative) and identify the main topics that are often discussed in each category using the LDA method. The contribution of this study is to provide a data-based approach to support more structured public space management. The research findings are expected to help managers in formulating development strategies, improving services, or improving facilities that are right on target, based on visitor perceptions and needs recorded in digital review data.

2. METHODS

Research methodology acts as the main guideline in conducting research, which helps ensure that the research process is carried out in a structured and systematic manner with the Knowledge Discovery in Databases (KDD) framework. The research process has several stages as shown in Figure 1.

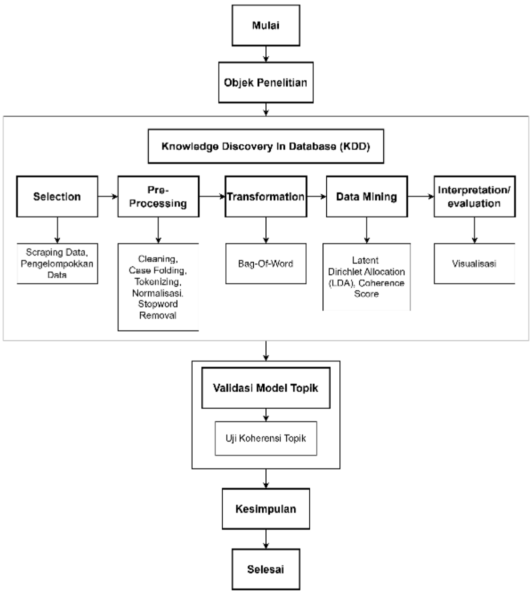

Mulai
Objek Penelitian
Knowledge Discovery In Database (KDD)
Selection
Pre-Processing
Transformation
Data Mining
Interpretation/ evaluation
Scraping Data, Pengelompokkan Data
Cleaning, Case Folding, Tokenizing, Normalisasi, Stopword Removal
Bag-Of-Word
Latent Dirichlet Allocation (LDA), Coherence Score
Visualisasi
Validasi Model Topik
Uji Koherensi Topik
Kesimpulan
Selesai

Figure 1. Research flow

2.1 Research Object

The object of this research is in the form of visitor review data for UNESA Lake taken from the Google Maps platform . Reviews include ratings (1–5) and comments related to visitor experiences, including aspects of facilities, cleanliness, and accessibility. A total of 4,761 reviews were successfully collected using web scraping techniques. This data is used as a basis in the topic exploration process in each rating category (positive, neutral, and negative).

2.2 KDD

Data analysis in this study was carried out through stages in the KDD framework which include [10] :

1. Selection

   Data was collected automatically using web scraping techniques [11] , then stored in Excel format that includes information on user name , review content, review time, and rating. After that, reviews were grouped by rating category to facilitate further exploration according to user perception.

2. Pre-Processing

   This stage aims to clean the data from irrelevant elements and simplify the text structure. The process includes cleaning to make the data cleaner [12] , case folding, tokenizing, normalizing non- standard words (for example "klo" becomes "kalang"), and stopword removal. This stage is important to improve accuracy in the next analysis process.

3. Transformation

   At this stage, text data is transformed into a numeric representation using the bag-of-words (BoW) approach. In addition, bigram and trigram techniques are applied to capture the meaning of combined words that often appear together, such as "banyak sampah". The output of this stage is a dictionary and corpus which are input for the topic modeling process.

4. Data Mining

   Topic modeling was performed using the Latent Dirichlet Allocation (LDA) method implemented with the Gensim library. The optimal number of topics was determined based on the highest coherence score [13] . This model groups words into topics based on their co-occurrence in documents, which then represent the dominant themes in each review category.

5. Interpretation/evaluation

   Visualization of the model results was done using wordcloud and pyLDAvis to assist in interpreting the meaning of each topic. Interpretation was done manually by the researcher by observing the dominant words in each topic to understand the context and their relationships. This stage ensures that the modeling results are in accordance with the research objectives.

2.3 Topic Model Validation

At this stage to ensure that the results of the LDA model are easy to understand by humans, a validation process is carried out using the word intrusion task and topic intrusion task. Both of these methods test a group of words from each topic to see if the words are related to each other and form a clear meaning. If respondents can easily recognize inappropriate words or irrelevant topics, then the model is considered to produce good and easy-to-understand topics [14] .

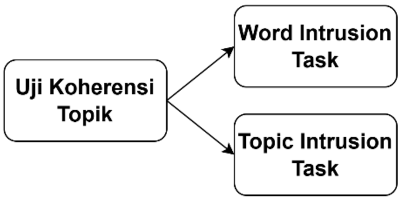


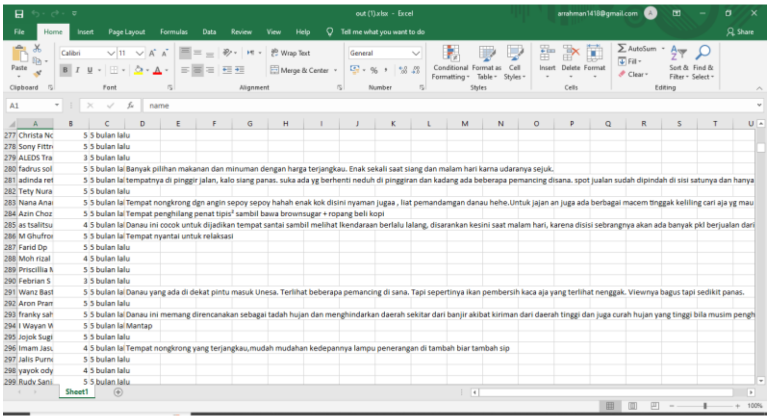
Figure 2. topic model validation

3. RESULTS AND DISCUSSION

3.1 Selection

1. Data scraping

Review data about UNESA Lake was obtained through web scraping techniques from Google Maps using the Python programming language. The information data that has been obtained includes 'name', 'rate', 'created_at', 'review'. All data that has been obtained is then stored in excel format. Data and source code were obtained from kaggle owned by @dewanakretarta . The data that was successfully collected was 4,761 reviews.

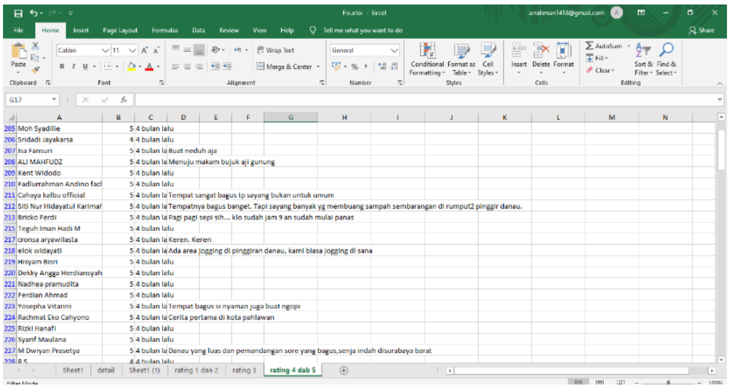


	A	B	C	D
277	Christa Nc	5	5 bulan lalu	
278	Sony Fittr	5	5 bulan lalu	
279	ALEDS Tra	3	5 bulan lalu	
280	fadrus sol	5	5 bulan lal	Banyak pilihan makanan dan minuman dengan harga terjangkau. Enak sekali saat siang dan malam hari karna udaranya sejuk.
281	adinda ret	5	5 bulan lal	tempatnya di pinggir jalan, kalo siang panas. suka ada yg berhenti neduh di pinggiran dan kadang ada beberapa pemancing disana. spot jualan sudah dipindah di sisi satunya dan hanya
282	Tety Nura	5	5 bulan lalu	
283	Nana Anai	5	5 bulan lal	Tempat nongkrong dgn angin sepoy sepoy hahah enak kok disini nyaman jugaa , liat pemandamgan danau hehe.Untuk jajan an juga ada berbagai macem tinggak keliling cari aja yg mau
284	Azin Choz	5	5 bulan lal	Tempat penghilang penat tipis² sambil bawa brownsugar + ropang beli kopi
285	as tsalitsu	4	5 bulan lal	Danau ini cocok untuk dijadikan tempat santai sambil melihat lkendaraan berlalu lalang, disarankan kesini saat malam hari, karena disisi sebrangnya akan ada banyak pkl berjualan dari
286	M Ghufroi	5	5 bulan lal	Tempat nyantai untuk relaksasi
287	Farid Dp	5	5 bulan lalu	
288	Moh rizal	4	5 bulan lalu	
289	Priscillia M	5	5 bulan lalu	
290	Febrian S	3	5 bulan lalu	
291	Wanz Bast	5	5 bulan lal	Danau yang ada di dekat pintu masuk Unesa. Terlihat beberapa pemancing di sana. Tapi sepertinya ikan pembersih kaca aja yang terlihat nenggak. Viewnya bagus tapi sedikit panas.
292	Aron Pram	5	5 bulan lalu	
293	franky sah	5	5 bulan lal	Danau ini memang direncanakan sebagai tadah hujan dan menghindarkan daerah sekitar dari banjir akibat kiriman dari daerah tinggi dan juga curah hujan yang tinggi bila musim pengh
294	I Wayan W	5	5 bulan lal	Mantap
295	Jojok Sugi	5	5 bulan lalu	
296	Imam Jasu	4	5 bulan lal	Tempat nongkrong yang terjangkau,mudah mudahan kedepannya lampu penerangan di tambah biar tambah sip
297	Jalis Purn	5	5 bulan lalu	
298	yayok ody	4	5 bulan lalu	
299	Rudy Sani	5	5 bulan lalu	

Figure 3. Data scraping results

2. Rating grouping

After the data is collected, the reviews are grouped into three categories based on the ratings given by visitors. Reviews with ratings 4 and 5 are included in positive reviews, ratings 3 as neutral reviews, and ratings 1 and 2 into negative reviews. This grouping aims to facilitate the identification and analysis of topics in each review category, so that each important aspect in the group can be analyzed more deeply.



	A	B	C	D
205	Moh Syadilie	5	4 bulan lalu	
206	Sridadi Jayakarsa	4	4 bulan lalu	
207	Isa Fansuri	5	4 bulan la	Buat neduh aja
208	ALI MAHFUDZ	5	4 bulan la	Menuju makam bujuk aji gunung
209	Kent Widodo	5	4 bulan lalu	
210	Fadlurrahman Andino fad	5	4 bulan lalu	
211	Cahaya kalbu official	5	4 bulan la	Tempat sangat bagus tp sayang bukan untuk umum
212	Siti Nur Hidayatul Karimah	5	4 bulan la	Tempatnya bagus banget. Tapi sayang banyak yg membuang sampah sembarangan di rumput2 pinggir danau.
213	Bricko Ferdi	5	4 bulan la	Pagi pagi sepi sih... klo sudah jam 9 an sudah mulai panas
215	Teguh Iman Hadi M	5	4 bulan lalu	
217	cronsa aryawilasta	5	4 bulan la	Keren. Keren
218	elok widayati	5	4 bulan la	Ada area jogging di pinggiran danau, kami biasa jogging di sana
219	Hisyam Bisri	5	4 bulan lalu	
220	Dekky Angga Herdiansyah	5	4 bulan lalu	
221	Nadhea pramudita	5	4 bulan lalu	
222	Ferdian Ahmad	5	4 bulan lalu	
223	Yosepha Vitarini	5	4 bulan la	Tempat bagus si nyaman juga buat ngopi
224	Rachmat Eko Cahyono	5	4 bulan la	Cerita pertama di kota pahlawan
225	Rizki Hanafi	5	4 bulan lalu	
226	Syarif Maulana	5	4 bulan lalu	
227	M Dwiyan Prasetya	5	4 bulan la	Danau yang luas dan pemandangan sore yang bagus,senja indah disurabaya barat
228	R S	4	4 bulan lalu	

Figure 4. Rating grouping results

3.2 Pre-Processing

The stages carried out in the *preprocessing process* can be seen in the flow diagram shown in Figure 5.

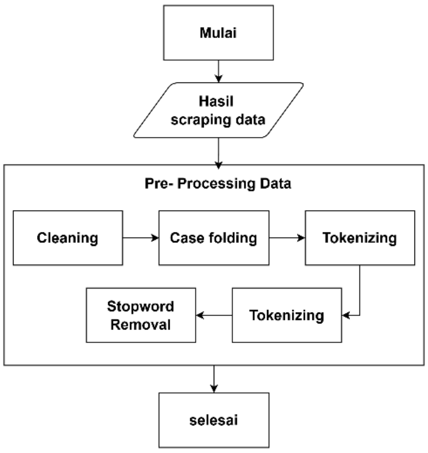


Figure 5. Preprocessing flow

1. Cleaning

Cleaning Removing irrelevant special characters for the cleaning process is done by deleting characters or words that are not meaningful such as punctuation, emoticons, tags, urls , symbols, numbers, mentions. In table 1 below there are comparison results before and after the cleaning stage:

Table 1. Cleaning Results

Before	After
The lake is cool , even though it's hot the wind is gentle ,	The lake is cool even though the wind is hot
The place is nice , especially in the afternoon for those who like hunting 😊	The place is good for those who like hunting in the afternoon.

2. Folding case

case folding changes uppercase letters to lowercase. In table 2 below there are comparison results before and after the case folding stage:

Table 2. Case Folding Results

Before	After
Danau sejuk, walaupun panas anginnya sepoi2,	danau sejuk walaupun panas anginnya sepoi
Tempatnya bagus kalau sore yang suka hunting	tempatnya bagus kalau sore yang suka hunting

3. Tokenization

Tokenizing to break down sentences in data into separate words [15] . Table 3 below shows the comparison results before and after the tokenizing stage:

Table 3. Tokenizing Results

Before	After
danau sejuk walaupun panas anginnya sepoi	[danau, sejuk, walaupun, panas, anginnya, sepoi]
tempatnya bagus kalau sore yang suka hunting	[tempat, bagus, kalau, sore, yang, suka, hunting]

4. Normalization

Normalization changes slang words into formal language, using a reference dictionary stored in the file "kamus slang.csv". This file contains a list of slang vocabulary along with its formal vocabulary in Indonesian. This process is done by reading the contents of the dictionary and saving it in the form of a dictionary named "kata_normalisasi_dict" which functions as a reference for replacing words. Furthermore, a loop is carried out to check each word in the text. If the word found is in the slang dictionary, then the word will be replaced with its formal version according to what is listed in the dictionary. Table 4 below shows the results of the comparison before and after the normalization stage:

Table 4. Normalization Results

Before	After
Tempat nongkrong	Tempat kumpul
Klo malam pengamennya kaya preman	Kalau malam pengamennya kaya preman

5. Stopword removal

Stopword Removal to remove unimportant words such as prepositions or conjunctions "although", "which", "and" or "if", words that do not help provide meaningful information in text analysis [15] . Table 5 below shows the comparison results before and after the stopword removal stage:

Table 5. Stopword Removal Results

Before	After
danau sejuk walaupun panas anginnya sepoi	danau sejuk panas anginnya sepoi
tempatnya bagus kalau sore yang suka hunting	tempatnya bagus sore suka hunting

3.3 Transformation

The stages carried out in the *transformation process* can be seen in the flow diagram shown in Figure 6 .

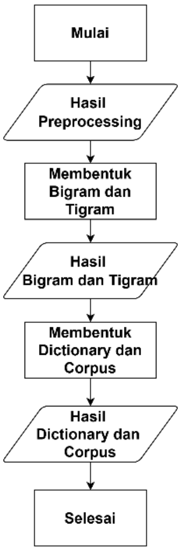

Mulai
Hasil Preprocessing
Membentuk Bigram dan Tigram
Hasil Bigram dan Tigram
Membentuk Dictionary dan Corpus
Hasil Dictionary dan Corpus
Selesai

Figure 6. Transformation flow

1. Bigram and Trigram

   The process of creating bigrams and trigrams is done by combining two or three words that often appear together in the text. The *Phrases function* from the Gensim library is used to detect meaningful word combinations when combined. In its application, *the min_count = 2 (trigram) and 3 (bigram) parameters are used* , which means that word combinations must appear at least twice and three times in the dataset to be considered a bigram or trigram. This approach helps maintain relevant phrases and reduces the occurrence of meaningless word combinations. The results of the bigram and trigram process can be seen in table 6 .

Table 6. Bigram and trigram results

Bigram	Trigram
banyak_sampah	kotor_banyak_sampah
banyak_pedagang	nyaman_banyak_pedagang
danau_unesa	pemandangan_danau_unesa

2. Dictionary and Corpus

   After the bigrams and trigrams are formed, the next step is to create a dictionary and corpus. A dictionary stores a collection of unique words that have been processed, with each word given a special index. This process is done using the *corpora module* from Gensim. Furthermore, the data is converted into a corpus in the form of *bag-of-words* (BoW), which represents the frequency of words in a document without regard to word order. The result of this process is a corpus data structure that is ready to be used as input for LDA modeling for topic extraction . The results of the dictionary and corpus formation process can be seen in table 7 .

Table 7. Dictionary and corpus results

Dictionary	Corpus
'banyak_sampah', 1	(33, 1)
'buang_sampah', 1	(34, 1)
'jualan', 2	(14, 2)
'enak', 1	(36, 1)

3.4 Data Mining

The stages carried out in the *data mining process* can be seen in the flow diagram shown in Figure 8.

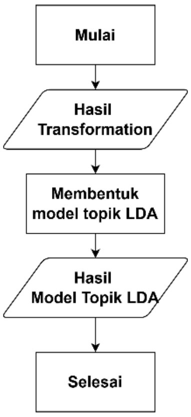
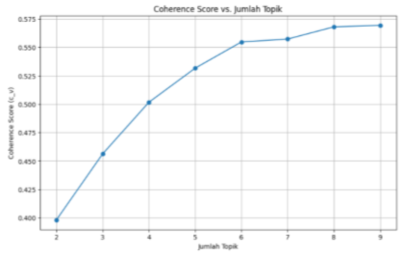
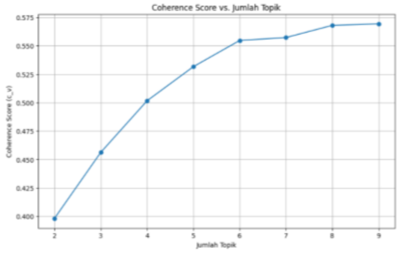
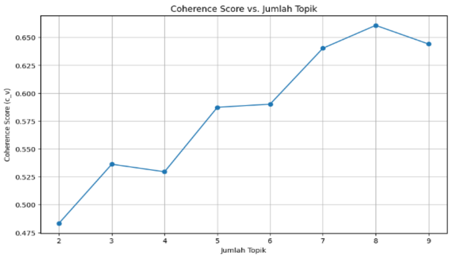


Figure 8. Data mining flow

Latent Dirichlet Allocation (LDA) is used to group frequently occurring words together to form topics from the review data. This process is done with the help of the Gensim library, which allows exploration of various input parameters. The selection of the number of topics is determined based on *the coherence score* with the highest value to ensure the resulting model is accurate and representative of the data content. The results of the coherence score can be seen in table 8.

Table 8. Results Coherence Score

Rating	chart	Mark	Number of topics
4 and 5		0.5694	9
3		0.5642	2
1 and 2		0.6608	8

The LDA model generates several topics, divided by rating category: 9 topics for reviews rated 4 and 5 (positive), 2 topics for ratings 3 (neutral), and 8 topics for ratings 1 and 2 (negative). Each topic displays 10 words with the highest probability weight. The results of the LDA topic model can be seen in table 9.

Table 9. LDA Topic Model Results

Rating	Topics	Words and Probability Weights
4 and 5	1	0.063*"asyik" + 0.059*"tempat_kumpul" + 0.038*"surabaya" + 0.025*"pagi" + 0.021*"jam" + 0.020*"area" + 0.020*"jagung_bakar" + 0.020*"sambil_makan" + 0.020*"kopi" + 0.019*"ceker_pedas"
	2	0.055*"danau_unesa" + 0.049*"makan" + 0.045*"cocok_untuk" + 0.037*"santai" + 0.036*"cocok_buat" + 0.025*"gratis" + 0.022*"sip" + 0.015*"bareng_teman" + 0.014*"air" + 0.013*"mantul"
	3	0.055*"murah" + 0.053*"unesa" + 0.045*"pinggir_danau" + 0.024*"aneka" + 0.024*"sekitar_danau" + 0.022*"asyik_buat_kumpul" + 0.020*"menikmati" + 0.017*"ngobrol" + 0.016*"tempat_wisata" + 0.015*"kalau_siang"

	4	0.135*"bagus" + 0.078*"nyaman" + 0.054*"danau" + 0.044*"makanan" + 0.042*"buat_kumpul" + 0.034*"mantap" + 0.019*"enak_buat_kumpul" + 0.018*"anak" + 0.017*"sih" + 0.016*"pas"
	5	0.071*"mancing" + 0.047*"lumayan" + 0.038*"kalau_malam" + 0.029*"mantab" + 0.026*"kota" + 0.025*"danaunya" + 0.022*"sayangnya" + 0.019*"minum" + 0.019*"tinggal_pilih" + 0.019*"pilihan"
	6	0.059*"tempatnya" + 0.055*"enak" + 0.047*"tempat_yang" + 0.029*"malam" + 0.029*"nya" + 0.027*"ngopi" + 0.026*"jajanan" + 0.025*"cocok" + 0.023*"sore_hari" + 0.021*"adem"
	7	0.079*"sejuk" + 0.044*"keren" + 0.039*"ceker" + 0.039*"malam_hari" + 0.038*"tempat_santai" + 0.033*"enak_buat" + 0.031*"bersantai" + 0.026*"lokasi" + 0.025*"banyak_orang_jualan" + 0.022*"melepas_pusing"
	8	0.081*"indah" + 0.051*"pinggir_jalan" + 0.049*"sore" + 0.037*"teman" + 0.025*"pemandangan" + 0.020*"tempat_ini" + 0.019*"ramai" + 0.019*"suka" + 0.019*"pkl" + 0.019*"tepi_danau"
	9	0.077*"kumpul" + 0.042*"kuliner" + 0.028*"suasana" + 0.026*"nyaman_untuk" + 0.024*"istirahat" + 0.024*"dan_makan" + 0.021*"banyak_orang" + 0.017*"seru" + 0.016*"sangat_nyaman" + 0.015*"bagus_untuk"
3	1	0.048*"bagus" + 0.041*"tempatnya" + 0.040*"nya" + 0.038*"nyaman" + 0.038*"lumayan" + 0.030*"sayang" + 0.026*"enak" + 0.023*"sih" + 0.021*"buat_kumpul" + 0.018*"indah"
	2	0.070*"danau" + 0.029*"makanan" + 0.026*"sore" + 0.025*"sampah" + 0.025*"kumpul" + 0.024*"banyak_sampah" + 0.023*"makan" + 0.020*"kuliner" + 0.019*"sisi" + 0.018*"ngopi"
1 and 2	1	0.187*"orang" + 0.114*"danau_unesa" + 0.078*"cocok" + 0.077*"jualan" + 0.077*"kurang_nyaman" + 0.077*"menikmati" + 0.060*"banyak_sampah" + 0.041*"sampah" + 0.041*"nyantai" + 0.041*"pinggir"
	2	0.121*"nyaman" + 0.082*"panas" + 0.082*"unesa" + 0.082*"ngopi" + 0.082*"jual" + 0.043*"pinggir" + 0.043*"jalan" + 0.043*"orang" + 0.043*"sma" + 0.043*"beli"
	3	0.164*"cocok" + 0.164*"mata" + 0.018*"kotor" + 0.018*"bagus" + 0.018*"jual" + 0.018*"banyak_sampah" + 0.018*"ngopi" + 0.018*"makanan" + 0.018*"tempatnya" + 0.018*"nyaman"
	4	0.186*"makanan" + 0.095*"danau" + 0.095*"teman" + 0.065*"parkir" + 0.065*"jalanan" + 0.065*"malam" +

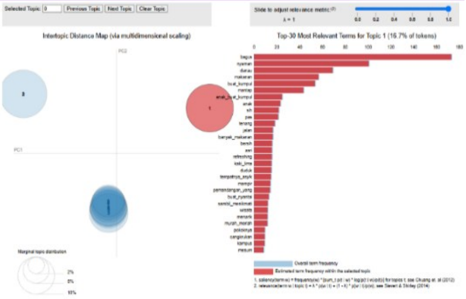
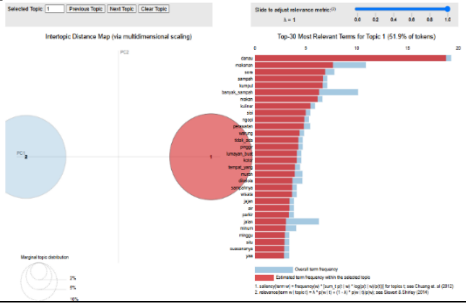
		0.064*"kumpul" + 0.034*"banyak_sampah" + 0.034*"elok" + 0.034*"tempatnya"
	5	0.161*"kaki_lima" + 0.161*"banyak_pedagang" + 0.110*"asri" + 0.058*"menikmati" + 0.058*"kurang_nyaman" + 0.058*"kurang_bersih" + 0.058*"jalan" + 0.058*"pinggir" + 0.058*"kotor" + 0.031*"banyak_sampah"
	6	0.270*"bau" + 0.143*"danau" + 0.016*"kotor" + 0.016*"bagus" + 0.016*"tempatnya" + 0.016*"nyaman" + 0.016*"ngopi" + 0.016*"banyak_sampah" + 0.016*"parkir" + 0.016*"malam"
	7	0.268*"kotor" + 0.093*"enk" + 0.049*"bagus" + 0.049*"nyantai" + 0.049*"beli" + 0.049*"enak" + 0.049*"kurang_bersih" + 0.049*"sma" + 0.049*"mata" + 0.049*"penjual"
	8	0.320*"bagus" + 0.087*"kurang_bersih" + 0.087*"sampah" + 0.087*"tempatnya" + 0.087*"nyaman" + 0.010*"asri" + 0.010*"kotor" + 0.010*"banyak_sampah" + 0.010*"bau" + 0.010*"kumpul"

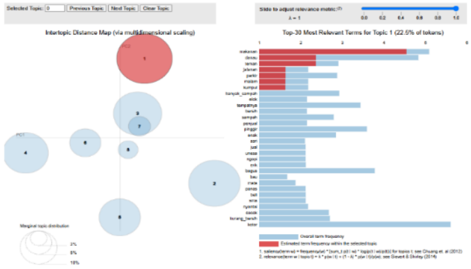
3.5 Interpretation/evaluation

The results of topic modeling will be visualized in two visualizations, namely pyLDAvis visualization and wordcloud, as follows:

The topic model visualization with pyLDAvis consists of two main panels. The left panel displays *an intertopic distance map* of numbered circles representing topics; the distance between circles reflects the degree of difference between topics. The right panel displays *a bar chart* of the 30 most relevant words, with blue indicating word frequency across documents, and red indicating word frequency within a particular topic. The results of one of the pyLDAvis visualizations can be seen in table 10.

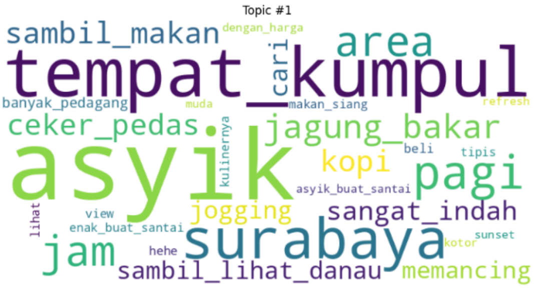
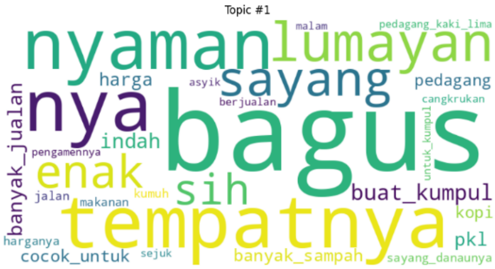
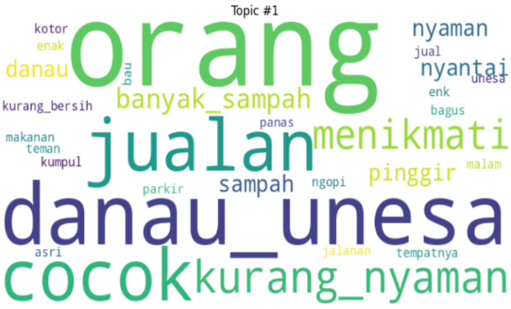
Table 10. Results of pyLDAvis visualization

Rating	pyLDAvis visualization
4 and 5	 Selected Topic: 0 Previous Topic Next Topic Clear Topic Slide to adjust relevance metric:(2) λ = 1 0.0 0.2 0.4 0.6 0.8 1.0 Intertopic Distance Map (via multidimensional scaling) PC2 PC1 1 2 Top-30 Most Relevant Terms for Topic 1 (16.7% of tokens) 0 20 40 60 80 100 120 140 160 180 bagus nyaman danau makanan buat_kumpul mantap enak_buat_kumpul enak sih pas tenang jalan banyak_makanan bersih asri refreshing kaki_lima duduk tempatnya_asyik mampir pemandangan_yang buat_nyantai sambil_menikmati wisata menarik murah_meriah pokoknya cangkrukan kampus mesum Marginal topic distribution 2% 5% 10% Overall term frequency Estimated term frequency within the selected topic 1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012) 2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)
3	 Selected Topic: 1 Previous Topic Next Topic Clear Topic Slide to adjust relevance metric:(2) λ = 1 0.0 0.2 0.4 0.6 0.8 1.0 Intertopic Distance Map (via multidimensional scaling) PC2 PC1 1 2 Top-30 Most Relevant Terms for Topic 1 (51.9% of tokens) 0 5 10 15 20 danau makanan sore sampah kumpul banyak_sampah makan kuliner sisi ngopi perawatan warung tidak_ada pinggir lumayan_buat kotor tempat_yang murah dikelola sampahnya wisata jajan air parkir jalan minum minggu situ suasananya yaa Marginal topic distribution 2% 5% 10% Overall term frequency Estimated term frequency within the selected topic 1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012) 2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

1 and 2	 Selected Topic: 0 Previous Topic Next Topic Clear Topic Slide to adjust relevance metric:(2) λ = 1 0.0 0.2 0.4 0.6 0.8 1.0 Intertopic Distance Map (via multidimensional scaling) PC2 PC1 1 2 3 4 5 6 7 8 Marginal topic distribution 2% 5% 10% Top-30 Most Relevant Terms for Topic 1 (22.5% of tokens) makanan danau teman jalanan parkir malam kumpul banyak_sampah tempatnya bersih sampah penjual pinggir enak jual unesa ngopi enk bagus bau panas beli nyantai cocok kurang_bersih kotor Overall term frequency Estimated term frequency within the selected topic 1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012) 2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Visualization in the form of a word cloud is also displayed for each topic, which shows 20 main words with a larger font size for words that have a high probability weight in the topic. The results of one of the word cloud visualizations can be seen in table 11

Table 11. Wordcloud visualization results

Rating	Workcloud visualization
4 and 5	
3	
1 and 2	

Based on the pyLDAvis and wordcloud visualizations, interpretations were made for each topic by observing the distribution of words that have the highest weight in the model. The results of the topic interpretation can be seen in table 12.

Table 12. Topic Interpretation Results

Rating	Topics	topic interpretation
4 and 5	1	Tempat Asyik untuk Berkumpul dan Jajan di Surabaya
	2	Danau UNESA sebagai Tempat Santai dan Gratis bersama Teman
	3	Tempat Murah untuk Kumpul di Sekitar Danau
	4	Lingkungan yang Bagus dan Nyaman untuk Makan dan Kumpul
	5	Aktivitas Mancing dan Minum dengan Banyak Pilihan

	6	Tempat Indah dan Adem untuk Ngopi di Sore Hari
	7	Tempat Sejuk untuk Bersantai dan Melepas Pusing
	8	Ramai di Sore Hari dengan Pemandangan yang indah
	9	Tempat Nyaman untuk Kulineran dan Istirahat
3	1	Tempat yang Bagus dan Nyaman untuk Kumpul
	2	Aktivitas Kuliner di Sore Hari dengan Isu Sampah
1 and 2	1	Aktivitas Santai di Pinggir Danau dengan Isu Sampah
	2	Tempat Pinggir Jalan untuk Ngopi namun Panas
	3	Tempat Ngopi dengan Pemandangan Kurang Bersih
	4	Kegiatan Malam dan Parkir di Sekitar Danau yang Kurang Bersih
	5	Suasana Asri dengan Banyak Pedagang tapi Kurang Bersih
	6	Tempat Bagus tapi Kotor dan Bau pada Malam Hari
	7	Jualan Makanan yang Enak tapi Kurang Bersih
	8	Tempat Bagus tapi Kotor dengan Banyak Sampah

3.6 Topic model validation

At this stage using the topic coherence test method aims to ensure that the results of topic modeling have a good level of coherence. The results of the topic coherence test using two methods, namely word intrusion task and topic intrusion task, are as follows:

Table 13. Word intrusion task results

question	Rating	Topics	Number of respondents	Correct	Wrong	Average
1	4 & 5	1	20	20	0	1.00
2		2	20	18	2	0.90
3		3	20	20	0	1.00
4		4	20	20	0	1.00
5		5	20	20	0	1.00
6		6	20	20	0	1.00
7		7	20	18	2	0.90
8		8	20	18	2	0.90
9		9	20	18	2	0.90
10	3	1	20	18	2	0.90
11		2	20	18	2	0.90
12	1 & 2	1	20	18	2	0.90
13		2	20	17	3	0.85
14		3	20	16	4	0.80
15		4	20	16	4	0.80
16		5	20	18	2	0.90
17		6	20	17	3	0.85
18		7	20	17	3	0.85
19		8	20	18	2	0.90
Total average						0.91

The level of respondent accuracy in the Word Intrusion Task reached 0.91 , indicating that the LDA model successfully identified topics well and respondents were able to understand the relationship between words in the UNESA Lake review.

Table 14. Topic intrusion task results

question	Rating	Topics	Number of respondents	Correct	Wrong	Average
1	4 & 5	1	20	19	1	0.95
2		2	20	19	1	0.95
3		3	20	20	0	1.00
4		4	20	19	1	0.95
5		5	20	18	2	0.90
6		6	20	20	0	1.00
7		7	20	20	0	1.00
8		8	20	20	0	1.00
9		9	20	18	2	0.90
10	3	1	20	18	2	0.90
11		2	20	18	2	0.90
12	1 & 2	1	20	15	5	0.75
13		2	20	14	6	0.70
14		3	20	15	5	0.75
15		4	20	15	5	0.75
16		5	20	16	4	0.80
17		6	20	16	4	0.80
18		7	20	16	4	0.80
19		8	20	16	4	0.80
Total average						0.88

of respondent accuracy in the Topic Intrusion Task reached 0.88 , which indicates that the LDA model successfully identified topics optimally and respondents were able to understand the relationships between topics in the UNESA Lake review.

CONCLUSION

This study successfully identified the dominant topics of UNESA Lake visitor reviews based on rating categories using the LDA method. Positive reviews highlighted comfort, cool atmosphere, and cheap culinary, while negative reviews discussed many issues of cleanliness and traders. Model validation showed good results with a Word Intrusion score of 0.91 and Topic Intrusion of 0.88, indicating that the resulting topics were easy to understand. These findings contribute to the use of digital reviews as a basis for decision making for public space management.

ACKNOWLEDGEMENTS

The author would like to express his deepest gratitude to Mr. I Kadek Dwi Nuryana for the guidance and direction that was very helpful during the research process. Thanks are also conveyed to the State University of Surabaya for the facilities and financial assistance that have been provided. Sincere gratitude is also addressed to family and friends who always provide encouragement and support during the preparation of this research.

REFERENCES

[1] O. Y. Findawati, M. M. Muhammad, A. Rosid, S. Kom, and M. Kom, Buku Ajar Text Mining. Sidoarjo: UMSIDA Press, 2020.

[2] R. Walalayo, E. A. W. Manuputty, A. J. R. Ufie, A. Niaga, and P. N. Ambon, “Pemanfaatan Google Maps dalam Mempromosikan Objek Wisata Tebing Makariki Negeri Yaputih Kecamatan Tehoru Kabupaten Maluku Tengah,” 2022.

[3] E. R. Susanto, “Sistem Informasi Geografis (GIS) Tempat Wisata di Kabupaten Tanggamus,” Jurnal Teknologi dan Sistem Informasi (JTSI), vol. 2, no. 3, pp. 125–135, 2021. [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[5] R. Alghamdi and K. Alfalqi, “A Survey of Topic Modeling in Text Mining,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 6, no. 1, pp. 147–153, 2015.

[6] E. Puspita, D. F. Shiddieq, and F. F. Roji, “Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc),” MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 2, pp. 481–489, 2024. doi: 10.57152/malcom.v4i2.1204.

[7] D. Rosmala and R. C. Nugroho, “Analisis Sentimen Web Novel Menggunakan Metode Latent Dirichlet Allocation (LDA),” Merkurius: Jurnal Riset Sistem Informasi dan Teknik Informatika, vol. 2, no. 2, pp. 44–53, 2024. doi: 10.61132/merkurius.v2i2.74.

[8] F. R. Junaedi, D. F. Zahra, and T. S. Ardan, “Analisis Ulasan Pembelian Produk Elektronik di Marketplace Tokopedia dengan menggunakan Topic Modelling,” Julyxxxx, vol. x, no. x, pp. 1–5, 2023.

[9] B. A. Tondang, M. R. Fadhil, M. N. Perdana, A. Fauzi, and U. S. Janitra, “Analisis Pemodelan Topik Ulasan Aplikasi BNI, BCA, dan BRI Menggunakan Latent Dirichlet Allocation,” INFOTECH: Jurnal Informatika dan Teknologi, vol. 4, no. 1, pp. 114–127, 2023. doi: 10.37373/infotech.v4i1.601.

[10] F. Gullo, “From Patterns in Data to Knowledge Discovery: What Data Mining Can Do,” Physics Procedia, vol. 62, pp. 18–22, 2015. doi: 10.1016/j.phpro.2015.02.005.

[11] A. Nayoan, “Apa itu Web Scraping? Pengertian, Teknik, dan Manfaatnya,” Niagahoster Blog, Jan. 13, 2020. [Online]. Available: https://www.niagahoster.co.id/blog/web-scraping/

[12] M. U. Albab, P. Y. Karuniawati, and M. N. Fawaiq, “Optimization of the Stemming Technique on Text Preprocessing President 3 Periods Topic,” Transformatika, vol. 20, no. 2, pp. 1–10, 2023. doi: 10.26623/transformatika.v20i2.5374.

[13] Listari, “Topic Modeling Menggunakan Latent Dirichlet Allocation (Part 2): Topic Modeling with Gensim (Python),” Medium.com, Aug. 1, 2019. [Online]. Available: https://medium.com/@listari.Tari/topic-modeling.

[14] P. S. Nautika and W. Yustanti, “Analisis Pinjaman Online pada Sosial Media Twitter Menggunakan Latent Dirichlet Allocation,” Journal of Informatics and Computer Science, vol. 6, 2024.

[15] V. Gurusamy and S. Kannan, “Preprocessing Techniques for Text Mining,” ResearchGate, 2014. [Online]. Available: https://www.researchgate.net/publication/273127322