A Lexicon-based Sentiment Analysis for the Investigation of the COVID-19 Pandemic's Impact on Culinary and Shopping Tourism of the City of Bandung

Review of Integrative Business & Economics Research

Mario B. Prasetya
Informatics Department, Parahyangan Catholic University

Cecilia E. Nugraheni*
Informatics Department, Parahyangan Catholic University

ABSTRACT

The COVID-19 pandemic has significantly impacted the tourism sector, decreasing visits and revenues worldwide. This article aims to examine the indirect impact of the COVID-19 pandemic on culinary and shopping tourism in the city of Bandung, namely by looking at tourist reviews from TripAdvisor and Google Reviews. A series of text mining stages were carried out to analyze the sentiment of tourist reviews, including data understanding, data cleaning, data transformation, lexicon-based sentiment analysis, data visualization, and interpretation and analysis. The sentiment analysis results were then used to test the hypothesis using the t-test and direct evidence. This study concludes that COVID-19 impacts tourist opinions towards shopping and culinary tourism in the city of Bandung in general and changes in opinion towards health aspects in particular.

Keywords: tourism, COVID-19, sentiment analysis, lexicon-based analysis, t-test.

Received 24 February 2024 | Revised 28 June 2024 | Accepted 21 July 2024.

1. INTRODUCTION

Tourism is essential to Indonesia's economic growth and development (Panggabean & Sipahutar, 2019). Therefore, developing and protecting the tourism industry is necessary for Indonesia to strengthen its position as a major tourist destination in Southeast Asia and the world. Apart from natural and cultural riches, technological developments and innovation also play an essential role in supporting the growth of Indonesia's tourism industry. Various digital applications and platforms have made accessing information and bookings easier for tourists, increasing the efficiency and comfort of their travels (Purwanto *et al*., 2023). Additionally, technologies such as artificial intelligence (AI) have enhanced the travel experience, allowing tourists to explore destinations virtually before visiting them.

Bandung is one of the famous cities in Indonesia It is the capital of West Java province. Bandung, known as the Paris van Java, boasts undeniable culinary and shopping tourism allure. Bandung is one of Indonesia's creative cities and is highly tourist-attracted, especially in the food industry (Chan *et al*., 2017). There is great variation in the culinary Bandung, which is the main attraction for domestic and foreign tourists (Chan *et al*., 2017). Besides, Bandung City, is also known as the centre of trendsetters for today's lifestyle (Raharja *et al*., 2021). In Bandung, plazas and factory outlets are implementing

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

the concept of a comfortable shopping centre equipped with culinary facilities and contemporary cafes, which are becoming tourist attractions (Raharja *et al*., 2021). Figure 1 shows the number of visitors to Bandung from year to year based on information from the Indonesian Central Bureau of Statistics.

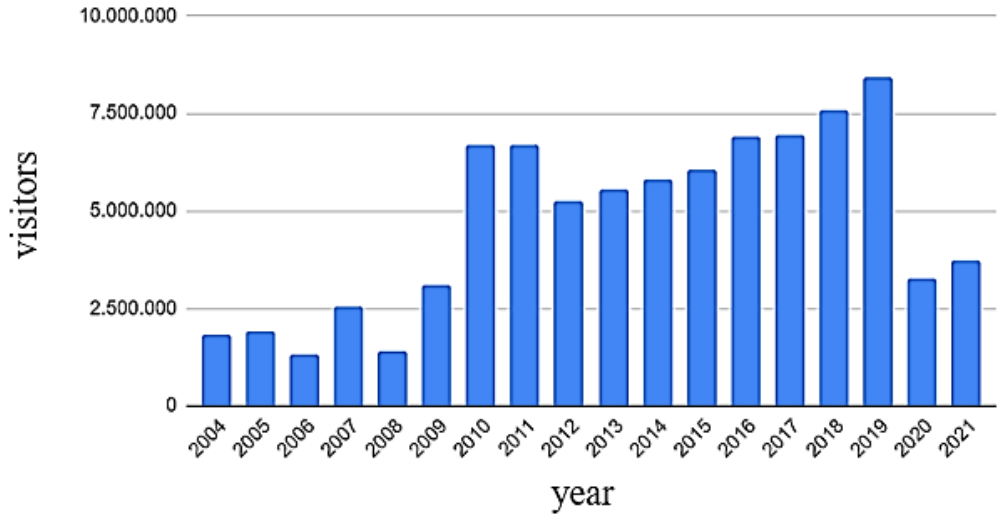

Figure 1. The number of visitors to the city of Bandung from 2004 to 2021.[1]

In culinary and shopping tourism, online platforms like TripAdvisor and various social media channels significantly shape consumer behavior and influence travel decisions. These platforms host a wealth of user-generated content, including reviews, ratings, and testimonials, which offer valuable insights into travelers' experiences and preferences. Data extracted from platforms like TripAdvisor provides a rich source of information for understanding the popularity of dining establishments and shopping venues and the sentiments expressed by visitors regarding their experiences.

Many studies have been conducted to solve tourism-related problems using exploratory data analysis, sentiment analysis, data mining, data science, and other techniques. Dhiratara *et al*. (2016) analyzed social media data for tourism to explore the relationship between social media dynamics and visitors' visiting patterns to touristic locations in real-world cases. Liapakis analyzed sentiment toward the food and beverage industry based on customer reviews found on the internet in Greek using lexicon-based techniques (Liapakis, 2020). Zhang analyzed the necessity of innovation in tourism marketing management and the existing problems of Intelligent Tourism Management in China (Zhang, 2021). Vu *et al*. presented a method for studying tourist activities based on a new type of data, venue check-ins, by analyzing a large-scale data set from 19 tourist cities in France (Vu *et al*., 2020). Arianto and Budi conducted an aspect-based sentiment analysis using Google Maps user reviews of Indonesia’s tourism destinations, which are Borobudur and Prambanan Temple, using five aspects, namely attractions, amenities, accessibility, image, price, and human (Arianto&Budi, 2020). Chan et al. have mapped tourist interest in Bandung by using social media data, especially from TripAdvisor, to find important information about tourist attractions (Chan *et al*. 2022).

The COVID-19 pandemic has significantly impacted the tourism sector, sharply declining visits and revenues worldwide. The city of Bandung also experiences this situation. As shown in Figure 1, there is a decline in tourists' visits in 2020. Travel restrictions and health concerns have changed how society behaves and responds to the

[1] https://www.nyenang.com/2022/08/pengunjung-dari-tahun-ke-tahun-di-kota.html

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

industry. This research aims to determine changes in tourist behavior in Bandung after the COVID-19 pandemic based on opinions on social media. Furthermore, this research aims to determine changes in attention to health aspects after the COVID-19 pandemic by tourists in Bandung.

This study was based on two hypotheses:

- Hypothesis 1. Opinions on shopping and culinary tourism in Bandung have changed after the COVID-19 pandemic.
- Hypothesis 2. Opinions on the health aspects of shopping and culinary tourism in Bandung have changed after the COVID-19 pandemic.

Figure 2 illustrates the research model used. Sentiment analysis was conducted in parallel to the review of opinions before and after COVID-19. We used March 2020 as the start of the COVID-19 pandemic. Four approaches are commonly used to analyze sentiment: machine learning, lexicon-based, rule-based, and statistics (Collomb *et al*., 2014). The approach used in this research is lexicon-based analysis. In addition, data visualization (Choy, J., 2013) is used to present the analysis results. The results of the sentiment analysis are used to test the first hypothesis. The second hypothesis can only be proven if the first hypothesis is valid.

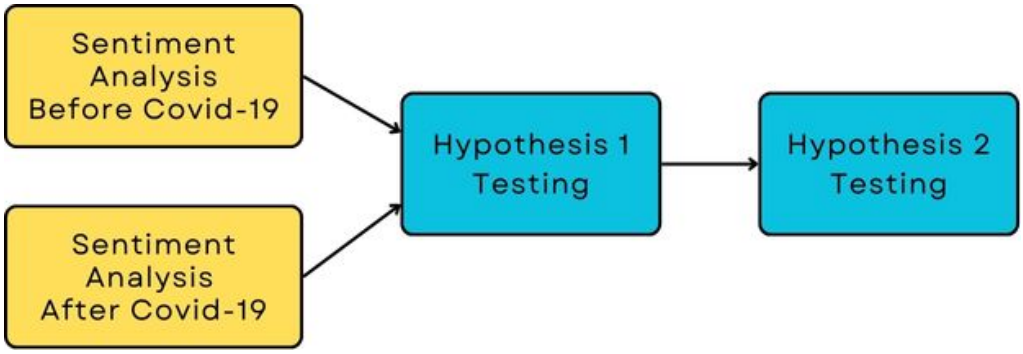

Figure 2. Research Model.

The rest of the paper is organized as follows: Section 2 presents the methodology used in this research. Section 3 describes the results, and Section 4 concludes the paper.

2. METHODOLOGY

2.1 Sentiment Analysis

Exploration and visualization of data are fundamental processes in data analysis aimed at gaining insights, identifying patterns, and communicating findings effectively. Data exploration involves examining and understanding the dataset's structure, content, and quality through various statistical and graphical techniques. This phase often includes data cleaning, summarization, and initial analysis to uncover potential relationships or anomalies within the data.

Exploration and visualization conducted in this research were carried out in the following stages: data understanding, data cleaning, data transformation, data mining, data visualization, and interpretation and analysis, as shown in Figure 3.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

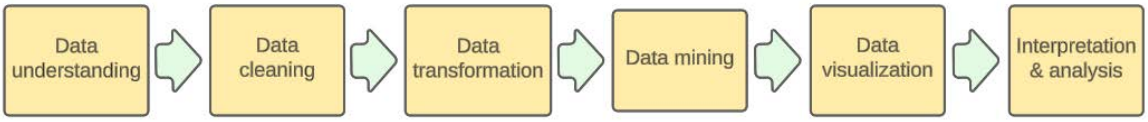

Figure 3. Exploration and visualization stages.

Each stage is explained briefly as follows:

a. Data understanding
   In this stage, understanding the text data involves recognizing the source of the text data to be analyzed. The process includes understanding the format of the text data, the type of information contained within the text (such as reviews, comments, or social media posts), and metadata associated with the text (such as dates, locations, or users who created the text).
b. Data cleaning
   The data cleaning process for text involves specific steps to remove unwanted characteristics, such as punctuation marks, emoticons, or URLs. The text data must also be cleaned from irrelevant stop-words and checked for spelling or orthographic errors.
c. Data transformation
   The data transformation stage for text involves converting the text into a format suitable for further analysis. The process may include tokenizing the text (splitting the text into tokens or individual words), removing stopwords (common words that do not carry meaning), and normalizing the text (converting the text into a uniform format, such as converting all letters to lowercase).

   One of the activities at this stage is stemming, which is changing the form of words into their most basic form. The stemming algorithm used in this research uses the Pysastrawi library, which implements the algorithm created by Nazief and Adriani (Pramudita, 2014).
d. Data mining
   In this stage, text data mining, in particular text mining (Tan, A.-H, 1999; Liu, B., 2012; Jo, T., 2018; Anandarajan, 2019), is performed to extract relevant information or patterns from the text. This stage may involve techniques such as analyzing word frequency (to find frequently occurring keywords), sentiment analysis (to determine positive, negative, or neutral sentiments from the text), or topic analysis (to identify topics or subjects discussed in the text).

   One of the most efficient text-based sentiment analysis methods is lexicon-based sentiment analysis. This approach involves using a predefined lexicon or dictionary of sentiment words and scoring the text based on the presence of positive, negative, or neutral words. The sentiment score of the text is then calculated based on the overall sentiment polarity of the words. This process, known as data labeling, aims to prepare training data to create a sentiment classification model without manual labeling. This lexicon dictionary of positive and negative words results from research by Wahid Devid (Wahid&Azhari, 2016; Liu *et al*., 2005). Examples of positive and negative words in this dictionary are shown in Table 1.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Table 1. Examples of positive and negative words.

Positive words	Negative words
cakap (competent)	*lalai* (negligent)
enak (delicious)	*jijik* (disguisting)
bersih (clean)	*kotor* (dirty)

A word-by-word examination determines whether the review contains positive or negative terms. A group of words containing positive words gets a score of +1, while a group containing negative words gets a score of -1. The scoring system also applies to words consisting of a negation prefix followed by a sentiment word (e.g., "don't like"). In this case, if the preceding word is a negation and the following word is from the positive word dictionary, both words are considered negative, and the score is reduced. Likewise, scores increase when a word that begins with a negation is followed by a word that has a negative sentiment, such as "didn't disappoint".

e. Data visualization
Once the sentiment analysis is complete, the next step is visually presenting the findings. This presentation can be done using various types of visualizations, such as sentiment histograms (to show the distribution of sentiments within the text), word clouds (to display the most commonly occurring sentiment words), or sentiment heatmaps (to visualize the sentiment patterns over time or across different categories).

f. Interpretation and analysis
In the final stage, the sentiment analysis results are interpreted to gain insights into the overall sentiment of the text. This process involves understanding the sentiment distribution, identifying patterns or trends in sentiment, and evaluating the implications of these findings for the analysis goals or research questions posed.

2.2 Hypothesis Testing

In general, there are four hypothesis testing steps: null hypothesis formulation, a significance level definition, the test statistic calculation, and then a conclusion analysis based on the null hypothesis.

One approach to hypothesis testing is the *t*-test. If we want to prove a hypothesis (H0) involving two data groups x_1 and x_2, then the significance level is symbolized by α, and the statistical calculation is with the following formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (1)$$

where

- $\bar{x}_1$ and $\bar{x}_2$ are sample means generated from group 1 and group 2.
- n_1 and n_2 are sample size of group 1 and group 2.
- $s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
  where s1 and s2 is the standard deviation of group 1 and group 2, respectively.

The hypothesis H0 is accepted whenever *t* obtained is greater than α.

3. RESULTS

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

3.1 Data Understanding

Data collection was done by collecting review data from tourists or visitors who visited restaurants related to culinary and shopping tourism in Bandung and its surroundings. Data was taken from two review provider websites, namely Tripadvisor and Google Reviews. Tripadvisor and Google Reviews provide suggestions for tourist attractions and facilities (restaurants, accommodations, and many others) in various regions, and users can provide reviews of places they have visited. From visitor reviews, potential visitors can consider whether a place they want to visit suits their wishes.

The restaurants chosen are those that are unique to Bandung in terms of food or places that are quite busy with tourism, such as Lembang, Dago, and other places that attract tourists to visit. This data was scraped using a web scrapping tool called WebHarvy. The data generated from WebHarvy is in the form of Excel data.

The following are the names of the restaurants chosen for review, namely Dusun Bambu, The Stone, Kampung Daun, Batagor Kingsley, Cuangki Serayu, Sunda Pavilion, Sudirman Street, Sate Bu Ngantuk, Alas Daun, Mie Baso Akung. As explained previously, this restaurant or cafe was chosen for several reasons. The first reason is that the restaurant is close to tourist centers in the Bandung area, such as Lembang and Dago, which offer natural beauty. Apart from that, some of the restaurants above were chosen because they are busy with tourists, judging from the number of reviews on the Tripadvisor and Google Review websites. The third reason is that several restaurants have characteristics specific to the Bandung area. This characteristic can be seen from the menu: typical Bandung foods such as Batagor, Cuanki, and other typical foods.

The dataset used in this research was collected from the Tripadvisor and Google Review websites. The data collected amounted to 1673 rows of data. The dataset was collected from several restaurants that have characteristics of Bandung in terms of food or restaurants that are quite visited because of tourism and culinary delights, such as Lembang, Dago, and places that attract tourists. For shopping tourism, some places reviewed were recommended by *TripAdvisor*, such as Factory Outlet, PVJ, Ciwalk, and several others. Data collection was taken from September 2022 to April 2023.

3.2 Data Cleaning

3.2.1 Case folding and numbers-punctuation removal

At this stage, all capital letters in the review column text will be changed to lowercase. Numbers are also necessary to delete because they are less important for analysis, and punctuation to reduce tokens/words that have the same meaning but are different because of the addition of punctuation marks.

3.2.2 Slang replacing and stemming

At this stage, slang words are replaced with standard words. Moreover, a stemming process is carried out to remove affixes from a word.

3.2.3 Stop-word removal

The purpose of stop-word removal is to remove words or tokens that have no value or meaning. Examples of stop words are conjunctions and question words. The list of stopwords used in this research was obtained from the PySastrawi library by importing the StopWordRemoverFactory module.

3.3 Data Labeling

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Several words in the positive and negative lexicon dictionary were deleted because they were not appropriate, such as the words 'hamlet' and 'goreng '. 'Hamlet' is the name of a place, and 'goreng' (fried) means a way of cooking food. The reason 'goreng' was included in the negative word category is because it was used in a negative context in the collected data. However, upon further review, it was deemed that this word should not be called negative in this research dataset.

Upon completing the value per word calculation, the scores of each review were methodically categorized into three distinct groups: positive, negative, and neutral.

3.4 Data exploration and visualization

After the data preparation process, which includes preprocessing and the crucial data labeling stages, the next step is data exploration and visualization. This process aims to dig up information or insight that can be used to assist in decision-making. In this stage, several visualization graphs are used to visualize text into graphs, such as barplots and pie charts.

The data collected amounts to 1673 rows, 1238 rows before the COVID-19 pandemic and 435 rows after it. Table 2 gives the detailed results of the labeling process. Furthermore, Figure 4 and Figure 5 show the percentage of each label in the form of pie charts.

Table 2. Data labelling results.

Label	Before COVID-19	After COVID-19
Positive	759	145
Negative	319	213
Neutral	160	77
Total	1238	435

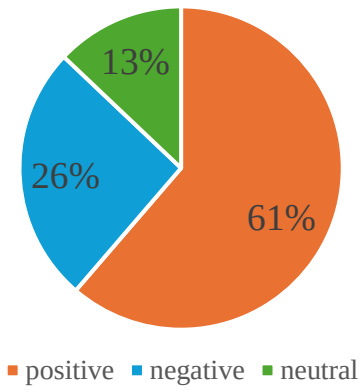

Figure 4. A pie chart for the percentage of sentiment labels before the COVID-19 pandemic.

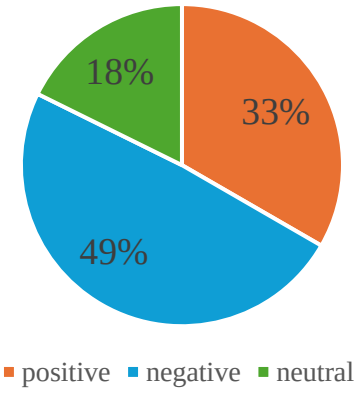

Figure 5. A pie chart for the percentage of sentiment labels after the COVID-19 pandemic.

The next step is to explore in more detail the words that occur most frequently for each category (positive, negative, and neutral). Figure 6, Figure 7, and Figure 8 display the 15 positive, negative, and neutral sentiment words that occur most often before the

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

COVID-19 pandemic, respectively. Figure 9, Figure 10, and Figure 11 display the 15 positive, negative, and neutral sentiment words that occur most often after the COVID-19 pandemic, respectively. These words are taken from all the reviews with corresponding sentiments (positive, negative, or neutral) and are included in Wahid Devid's word list.

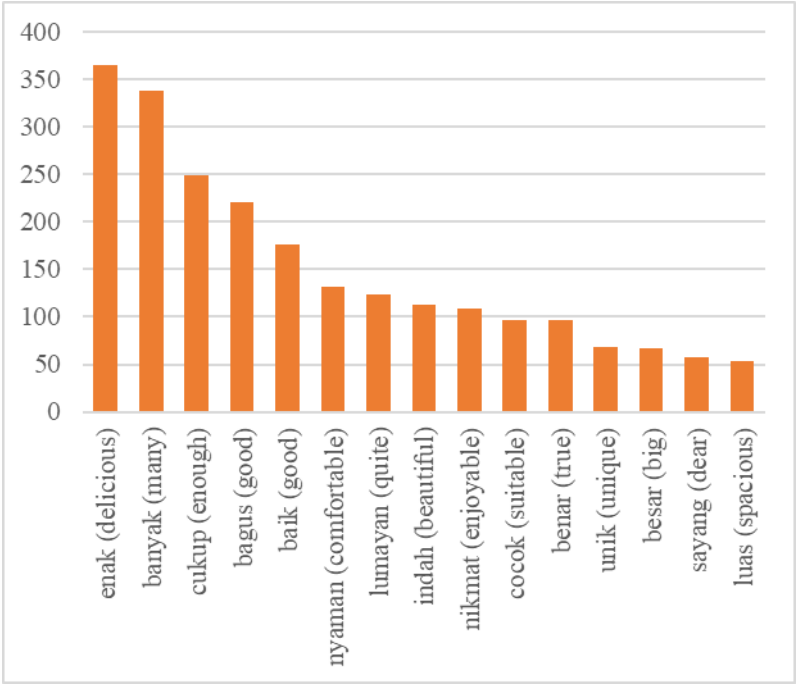

Figure 6. The 15 frequently occurring words with positive label before the COVID-19 pandemic.

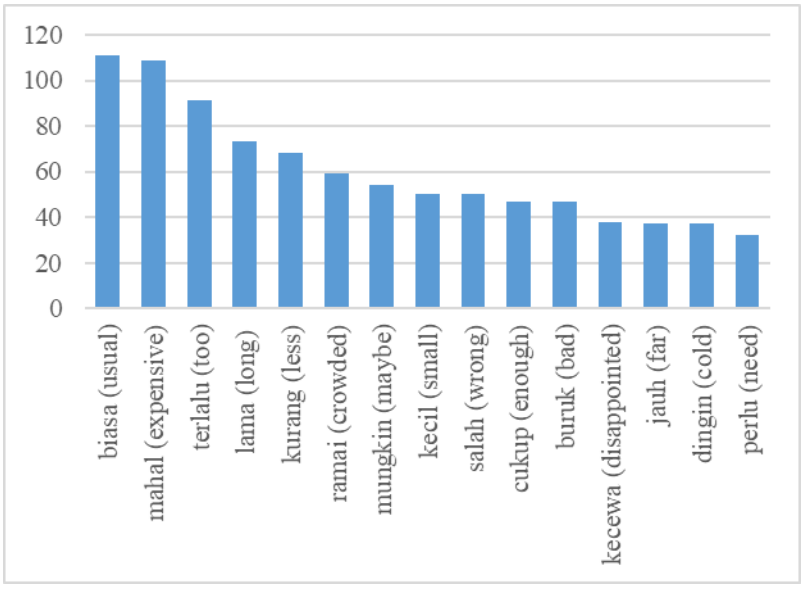

Figure 7. The 15 frequently occurring words with negative label before the COVID-19 pandemic.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

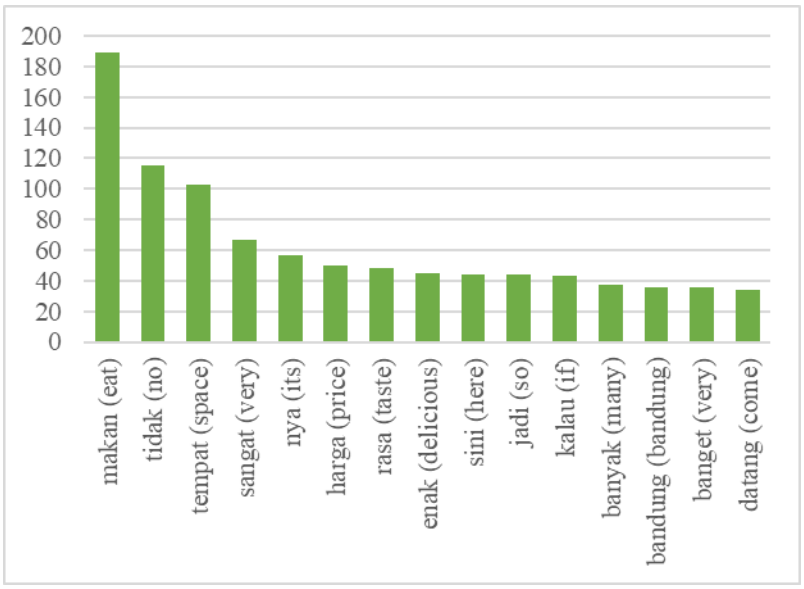

Figure 8. The 15 frequently occurring words with negative label before the COVID-19 pandemic.

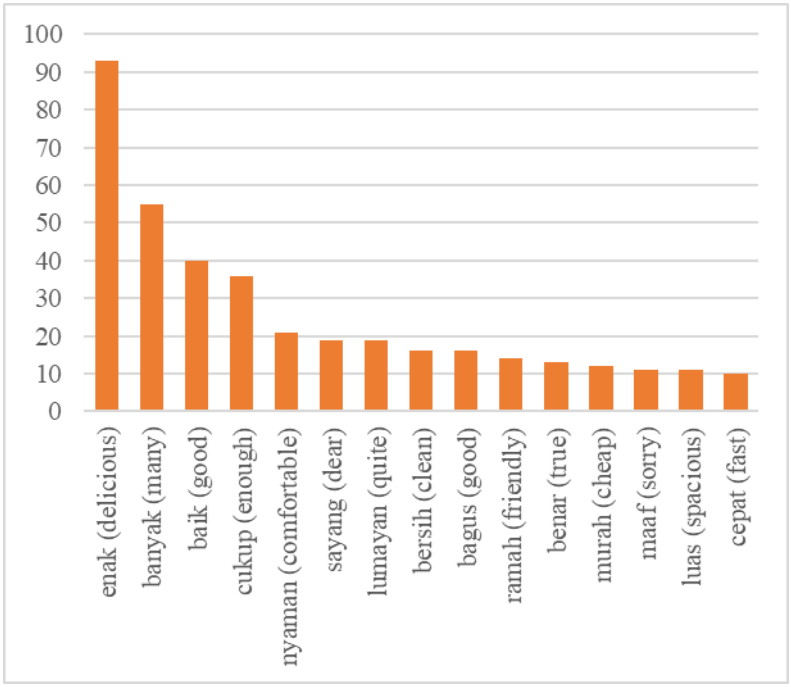

Figure 9. The 15 frequently occurring words with positive label after the COVID-19 pandemic.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

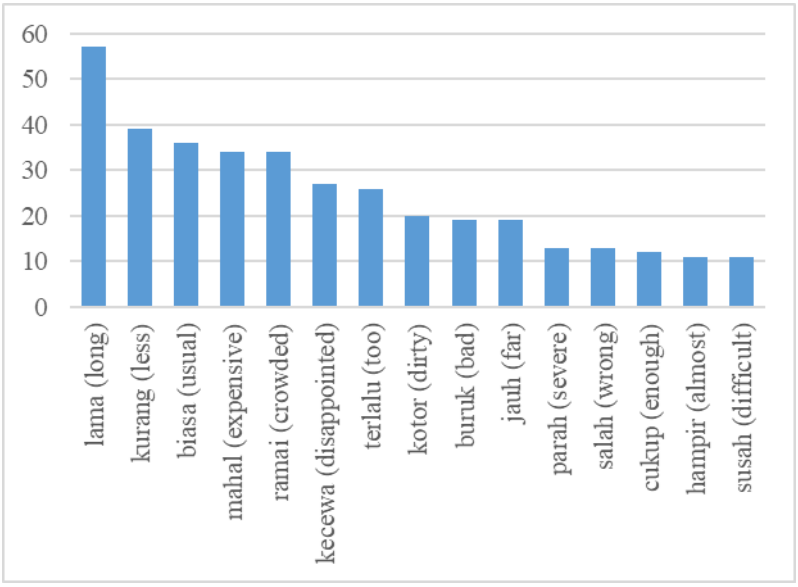

Figure 10. The 15 frequently occurring words with negative label after the COVID-19 pandemic.

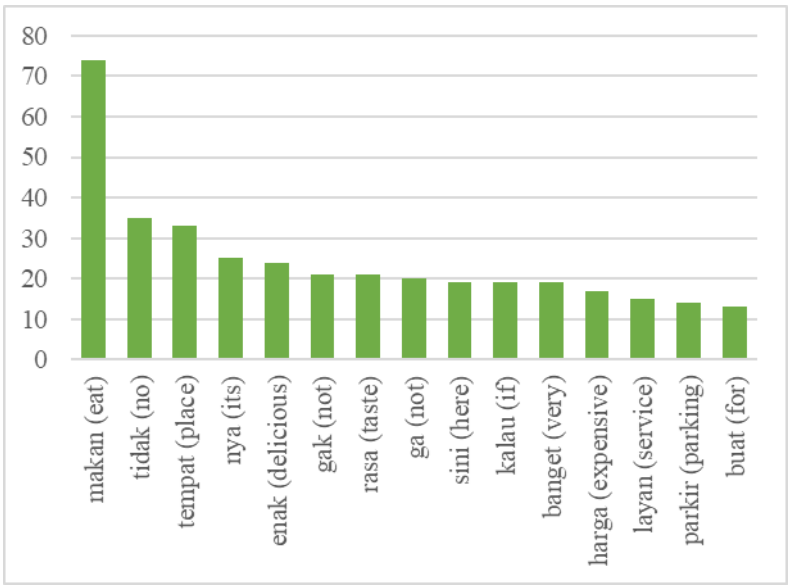

Figure 11. The 15 frequently occurring words with neutral label after the COVID-19 pandemic.

3.5 First Hypothesis Testing

To test the hypothesis regarding the influence of COVID-19 on opinions about culinary and shopping tourism in Bandung City, the *t*-test was used to compare the means of opinion values generated from the sentiment analysis process before and after the COVID-19 pandemic.

The null and research hypotheses are defined as follows:

H0: Opinions on shopping and culinary tourism in Bandung have changed after the COVID-19 pandemic.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

H1: Opinions on shopping and culinary tourism in Bandung have not changed after the COVID-19 pandemic.

A significance level, α, is set to 0.05 (α = 0.05). Using equation (1), we define the means of each group, which is calculated as follows:

$$\bar{x} = \frac{(1)*n_{pos}+(0)*n_{neu}+(-1)n_{neg}}{n_{pos}+n_{neu}+n_{neg}} \quad (2)$$

where $\bar{x}$ is the means of x, n_{pos}, n_{neu} and n_{neg} represents the number of positive, neutral, and negative reviews, respectively. The result t is 10.54. Since $t > \alpha$, we fail to reject H0. Thus, we conclude that there are differences between the opinions towards culinary and shopping tourism in Bandung City before and after the COVID-19 pandemic, implying that H0 is proven to be true.

3.6 Second Hypothesis Testing

The second hypothesis was proven directly by looking at the words that appeared with high frequency related to health. From the sentiment analysis results, the words related to the health aspect were *bersih* (clean) and *kotor* (dirty). These two words did not appear or were not included in the 15 most frequently appearing words before the COVID-19 pandemic. After the pandemic, the word clean was included in the 15 most frequently appearing positive labeled words, while the word dirty was included in the 15 most frequently appearing negative labeled words. Therefore, the second hypothesis can be accepted.

4. CONCLUSION

This study investigates user review data pertaining to culinary and shopping tourism in the city of Bandung on TripAdvisor and GoogleReviewer. Specifically, this paper discusses the evidence of the impact of the COVID-19 pandemic on shopping and culinary tourism in Bandung. The impact of the decline in tourists is seen from changes in the opinions of Bandung city tourists posted on social media, GoogleReview and TripAdvisor.

The findings derived from this investigation are outlined below:

- Before the COVID-19 pandemic, the majority of the data collected on culinary and shopping tourism in Bandung, comprising 61%, was classified as positive. However, the landscape has drastically changed post-COVID, with negative reviews (49%) now outnumbering positive reviews (33%), while neutral reviews are 18%. This significant shift in sentiment underscores the urgent need to address the impact of the pandemic on tourism in Bandung.
- The difference in opinion towards tourists before and after the COVID-19 pandemic, which can be seen from the distribution of review labels, is strengthened by proving the hypothesis using a *t-test*. The results of this proof state that there is a very significant change in opinion.
- From the exploration of sentiment analysis results, it can be concluded that there is a change in opinion related to health issues, namely the emergence of two new words, clean and dirty, whose frequency of occurrence has increased or only emerged after the COVID-19 pandemic.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

The lexicon-based sentiment analysis conducted in this study uses a particular lexicon for Indonesian. If this approach is to be used for different objects using different languages, then an appropriate lexicon must be used.

ACKNOWLEDGEMENT

The authors to thank the anonymous reviewer for his/her helpful comments and suggestions.

REFERENCES

[1] Anandarajan, M., Hill, C., and Nolan, T. (2019) *Practical text analytics: Maximizing the value of text data*. Springer.

[2] Arianto, D. and Budi, I. (2020). *Aspect-based Sentiment Analysis on Indonesia's Tourism Destinations Based on Google Maps User Code-Mixed Reviews (Study Case: Borobudur and Prambanan Temples).* https://aclanthology.org/2020.paclic-1.41.pdf.

[3] Chan, A., Tresna, P.W., and Suryadipura D. (2017). *Experiential Value of Bandung Food Tourism.* Review of Integrative Business and Economics Research, Vol. 6(s1), 184-190.

[4] Chan, A., Suryadiputra, D., and Sukmadewi, R. (2022). *Mapping of Tourism Interests Through the Use of Digital Data*. Review of Integrative Business and Economics Research, Vol. 6(s1), 184-190.

[5] Collomb, A., Costea, C., Joyeux, D., Hasan, O., and Brunie, L. (2014). *A study and comparison of sentiment analysis methods for reputation evaluation*. Rapport de recherche RR-LIRIS-2014-002.

[6] Dhiratara, A., Yang, J., Bozzon, A., Houben, G-J. (2016). *Social Media Data Analytics for Tourism A Preliminary Study*. Proceedings of the 2nd International Workshop on Knowledge Discovery on the WEB Cagliari, Italy, September 8 - 10, 2016.

[7] Jo, T. (2018). *Text mining: Concepts, implementation, and big data challenge*. 1st ed. Springer, Seoul.

[8] Liapakis, A. (2020). *A Sentiment Lexicon-Based Analysis for Food and Beverage Industry Reviews*. The Greek Language Paradigm (May 20, 2020). Available at SSRN: https://ssrn.com/abstract=3606071 or http://dx.doi.org/10.2139/ssrn.3606071.

[9] Liu, B. (2012) *Sentiment analysis and opinion mining*. Synthesis lectures on human language technologies, 5, 1–167.

[10] Liu, B., Hu, M., and Cheng, J. (2005). *Opinion observer: analyzing and comparing opinions on the web*. Proceedings of the 14th international conference on World Wide Web, pp. 342–351.

[11] Panggabean, M. & Sipahutar, T. (2019). *Analysis of the Role of Tourism in the Economy in Indonesia*. International Journal of Advances in Social and Economics. 1. 10.33122/ijase.v1i6.126.

[12] Pramudita, H. R. (2014). *Penerapan algoritma stemming nazief & adriani dan similarity pada penerimaan judul thesis*. *Data Manajemen dan Teknologi Informasi (DASI)*, **15**, 15.

[13] Purwanto, N., Amelia, A., Ronald, R., Pancaningrum, E., and Irawan, N. (2023). *How to Build Adoption Intentions through Customer Engagement for Travelling Application in*

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)

Indonesia. Review of Integrative Business and Economics Research, Vol. 12(s3), 143-149.

[14] Raharja, R.J., Muhyi, H.A., and Adiprihadi D. (2021). *Contribution of the Retail Sector Towards City Economy: Study in Bandung City, Indonesia*. Review of Integrative Business and Economics Research, Vol. 10(s2), 19-32.

[15] Tan, A.-H. (1999). Text mining: *The state of the art and the challenges*. Proc. of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases (KDAD'99), pp. 65–70.

[16] Vu, H. Q., Luo, J. M., Li, G., and Law, R. (2020). *Exploration of Tourist Activities in Urban Destination Using Venue Check-In Data*. Journal of Hospitality & Tourism Research, 44(3), 472-498.

[17] Wahid, D. H. and Azhari, S. (2016). *Peringkasan sentimen esktraktif di twitter menggunakan hybrid tf-idf dan cosine similarity*. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 10, 207–218.

[18] Zhang, H. (2021). *Exploration and Analysis of Tourism Marketing Management Innovation Based on Big Data*. J. Phys.: Conf. Ser. 1744 032086.

ISSN: 2304-1013 (Online); 2304-1269 (CDROM); 2414-6722 (Print)