ISSN 2085-4552

Analysis of User-generated Content in Visitor Reviews of Tourist Attractions Using Semantic Similarity

Ni Made Satvika Iswari[1], I Gede Juliana Eka Putra[2]
[1,2]Informatics Study Program, Universitas Primakara, Denpasar, Indonesia
[1]iswari@primakara.ac.id, [2]gedejep@primakara.ac.id

Accepted 24 March 2023
Approved 28 June 2023

***Abstract*— The tourism industry plays an important role in the world economic sector because it makes a significant contribution to the global economy, creates jobs, and strengthens economic growth. Therefore, visitor satisfaction is very important in this industry. Currently tourists use information in online media to find tourist attractions that suit their needs and expectations. Reviews of tourist attractions are growing on the internet. Customers can post reviews, recommendations, or ratings of a tourist spot. Online reviews in the form of User-generated Content (UGC) can provide benefits for business managers to obtain feedback from customers and improve certain product attributes or service characteristics to increase business value and support marketing activities. In this research, a user-generated content analysis method has been produced in Visitor Reviews of Tourism Objects Using Semantic Similarity. The results of the sentiment analysis can be seen that positive, negative, and neutral sentiments have similar percentages for each category. This is because the designed algorithm cannot cut sentences for only one category that is assessed. The algorithm designed in this study has limitations, that is only able to analyze up to the semantic similarity stage but has not been able to cut which sentences are relevant to the category to be analyzed for sentiment.**

***Index Terms*— semantic similarity; sentiment analysis; tourist attraction; User-generated Content (UGC); visitor review**.

I. INTRODUCTION

The tourism industry plays an important role in the world economic sector. In 2018, global tourism increased by 6% based on the United Nations World Tourism Organization (UNWTO) and is expected to increase by 3% to 4% in 2019. According to data from the World Tourism Organization (UNWTO), around 1.4 billion people travel international travel in 2018. The tourism industry makes a significant contribution to the global economy, creates jobs, and strengthens economic growth. Therefore, visitor satisfaction is very important in this industry[1].

However, the worldwide spread of Covid-19 in 2020 caused the tourism industry to face a catastrophic crisis as migration and travel became restricted. The ongoing pandemic has had an impact on tourist behavior. Micro, Small, and Medium Enterprise (MSME)s are most national tourism service providers. However, MSMEs lack the infrastructure and resources to provide products and services to more tourists. Estimated tourist arrivals can help MSMEs to significantly reduce the risk of not achieving sales targets, and even allow MSMEs to start building personalized tour packages based on the interests and needs of tourists [1].

Currently, tourists use information in online media to find tourist attractions that suit their needs and expectations. Reviews of tourist attractions are growing on the internet. Customers can post reviews, recommendations, or ratings of a tourist spot. Online review in the form of user-generated content (UGC) give business management advantages by enabling them to get client feedback and enhance features of products or services to boost brand value and help marketing initiatives [2].

Several related studies have provided significant findings about feedback in tourism. Most of these studies use data collection methodologies such as customer surveys, interviews, or Focus Group Discussion (FGD)s to analyze the factors influencing tourist behavior. Although this research provides many perspectives on tourist behavior, the Big Data Analytics approach can provide significant insights [3]. In this study, data collection was carried out from one of the UGC tourism data sources, namely visitor reviews on Google Maps. Data analysis was performed using the semantic similarity processing method. Analysis of the data will produce comprehensive information about tourist's feedback.

II. LITERATURE REVIEW

A. Related Study

Data is a vital asset to support business progress, including in the tourism sector. Information is necessary, but analytical outcomes are as important, particularly with big data technology. Big Data is used by businesses and organizations to assist in decision-

making. As a result, the analysis must be correct. The wrong choice will be made because of incorrect analysis, which will eventually be disastrous. Due to feature learning, Deep Learning algorithms—which are utilized in massive data analysis—can now make judgments that are comparable to those of humans. Big data analysis can improve intelligent computing by helping it to analyze unstructured real-world data[4].

Compared to conventional tourist survey data, a larger volume of tourist data is available online. The data allows decision makers to understand tourist behavior in a more granular way. Social media and social networking websites are an online platform that allows users to share experiences and opinions on a product or service [5]. Currently, tourists use information in online media to find tourist attractions that suit their needs and expectations. Given the large amount of unstructured content from social media, big data analytics approach like sentiment analysis, geographic visualization, and frequency analysis have been widely applied for User Generated Content (UGC) analysis in several research domains. UGC content offers the opinions and sentiment analysis of numerous individuals in diverse areas[6]. UGC is currently growing because it provides a valuable source of data for many parties to extract information that can be used as a competitive advantage[7].

Sentiment analysis-based neural networks are utilized to assist the information architecture of the tourism industry [1]. Big data analysis and content analysis are utilized in the restaurant industry to examine user-generated content (UGC) relevant to the meal experience for diners who have food allergies [8]. Supervised machine learning approach used on big data analytics cloud-based to support customer insight-based design innovation for the SME domain [9]. In this study, the UGC analysis was carried out using the natural language processing approach. The data analysis will produce comprehensive information about customer satisfaction.

B. Semantic Similarity

In computer science, semantics is the mathematical justification for a valid string that is specified by a programming language. The rules regulating program, or syntax, are the exact reverse of this. It's simpler to conceive of semantics as meaning and syntax as structure. Computers can now read, compare, and extract meaning from human languages thanks to the field of computer science known as natural language processing. To create models from text and speech, natural language relies on the disciplines of linguistics and computer science [10].

The question of "how two words/phrases/documents are similar to each other?" is a fundamental one for research and applications in Natural Language Processing (NLP). To determine how closely two words, phrases, or documents resemble one another, use text similarity. This similarity could be lexical or in meaning.

Lexical similarity refers to the similarity of the word set, whereas semantic similarity refers to the similarity of the meaning. The cosine distance between the two embedded vectors in a sentence serves as an indicator of its semantic similarity. Although many people consider this computation to be difficult, building word or phrase embeddings is significantly more difficult. Text must first be transformed into a vector of features before the algorithm can choose an appropriate features representation, such as TF-IDF. Similarity research on text representation vectors is the last step.

There are numerous methods for determining text similarity, whether they take semantic relations into account or not. In addition to these methods:

1. Jaccard Similarity
2. Cosine Similarity
3. K-Means
4. Latent Semantic Indexing (LSI)
5. Etc.

The degree to which two texts' meanings are similar is gauged by their semantic similarity. Typically, this measurement is scored between 0 and 1. 0 denotes that they are utterly unrelated, and 1 indicates that they nearly share the same meaning.

The fields of linguistics and natural language processing are actively conducting research in text semantic similarity. Additionally, it participates in a variety of applications for informatics and natural language processing. In many natural language processing (NLP) applications, such as sentiment analysis, natural language comprehension, machine translation, question answering, chatbots, search engines, and information retrieval, we make use of semantic similarity.

Applications for informatics sciences exist in the biological and geo-informatics fields. Semantic similarity approaches are mostly used in biomedical informatics to develop biological ontologies, such as Genes Ontology. Ontologies for geographic features used in geo-informatics rely on topological and statistical measures of semantic similarity. The OSM Semantic Network, which calculates the semantic similarity of tags in OpenStreetMap, is one of the most well-known tools for this kind of application [11].

III. METHODOLOGY

This research was conducted in 3 stages of activity, namely data collection, data analysis, visualization, and data interpretation, as shown in Figure 1. The first stage carried out was the collection of User-generated Content (UGC) data which would be processed in the

next stage. The data will be taken from the Google Maps website, which consists of visitor review data.

Because UGC analysis provides valuable information like opinions, evaluations, suggestions, experiences, or customer demands, it can be very helpful to organizations. The internet, social media, blog posts, tweets, product evaluations, and survey responses with accompanying images all provide open access to this information. UGC typically only exists in unstructured or semi-structured formats, making it impossible to process using conventional data mining methods.

Figure 2 shows visitor reviews on Google Maps which are available in several categories, such as scenery, massage, malls, traffic hours, tattoos, dusk, surf boards, etc. While Figures 3 and 4 show several user reviews based on word occurrence per category. The appearance of the word is the same word as the category or result of the translation. For example, for the category of massage, the results of a review that contains the word massage appear. Meanwhile for the scenery category, the results of a review that contain the word scenery or its translation in other languages appear .

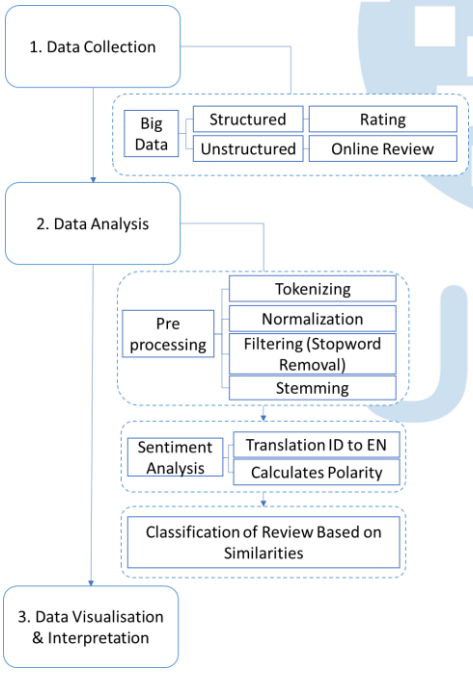

Fig. 1. Research Methodology

In this study, the Semantic Similarity approach was used to find review results that not only contained the exact same keywords as the categories but were also able to find review results that did not contain the exact same keywords but had the same semantics (meaning) as the categories. Semantics in computer science is the mathematical reasoning behind legal string specified by the programming language. It is the opposite of syntax, the rules that govern computer programs. It's easiest to think of syntax as structure and semantics as meaning.

Semantic analysis is a process of extracting meaning from text. Grammatical analysis and recognition of the relationships between certain words in certain contexts allow computers to understand and interpret phrases, paragraphs or even manuscripts. Using semantic analysis can help businesses in many ways, such as interpreting customer reviews more comprehensively. In terms of analyzing visitor reviews, semantic analysis can produce comprehensive information about tourist feedback. This analysis is needed for visitor review data that reviews a category for a tourist spot that does not explicitly state the name of the category[12].

The second stage of this study is data analysis, which includes pre-processing stages, sentiment analysis, and review classification based on similarity of context (similarities). Pre-processing is the first step taken in data mining techniques to convert raw data into cleaner data and ready to be used for further analysis. Some of the pre-processing stages carried out include [13]:

1. *Tokenizing* is the stage of splitting the string into tokens. Previously, the process of number removal and punctuation removal was first carried out.
2. *Normalization* is a process of cleaning words which are words in everyday language/*slank words* be the correct Indonesian word.
3. *Filtering* is a stop word removal process. At this stage, the Indonesian stop word was used which was obtained from the NLTK library (https://www.nltk.org/).
4. *Stemming* is the process of removing word affixes and changing them into basic words

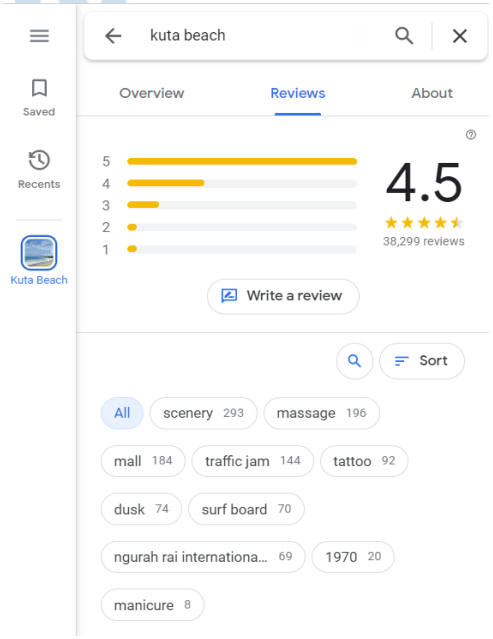

Fig. 2. Visitor Reviews on Google Maps Available in Several Categories

After going through the pre-processing stages, the data is ready to be processed in the semantic similarity analysis process. At this stage, 13 categories have been prepared, which are keywords that are reviewed quite a lot by visitors to tourist attractions. The categories consist of scenery, massage, mall, traffic jam, tattoo, dusk, surfboard, Ngurah Rai Int Airport, Manicure, Facilities, Location, Meals, and Sanitation. This semantic similarity analysis process will produce reviews related to the 13 categories mentioned above.

After going through the semantic similarity process, the next stage is the sentiment analysis process. The sentiment analysis process is carried out using the TextBlob Library (https://pypi.org/project/textblob/). TextBlob is a python library used to analyze data in the form of text. After going through this process, visitor review data is obtained based on the category and sentiment analysis results (positive, negative, or neutral).

The final stage is the interpretation and visualization of the data. Data that has been processed will be easier to understand if displayed in a visual form. Business managers may ultimately find this information useful in gathering client feedback and refining features of their products or services in order to boost their company's value and assist marketing initiatives [2].

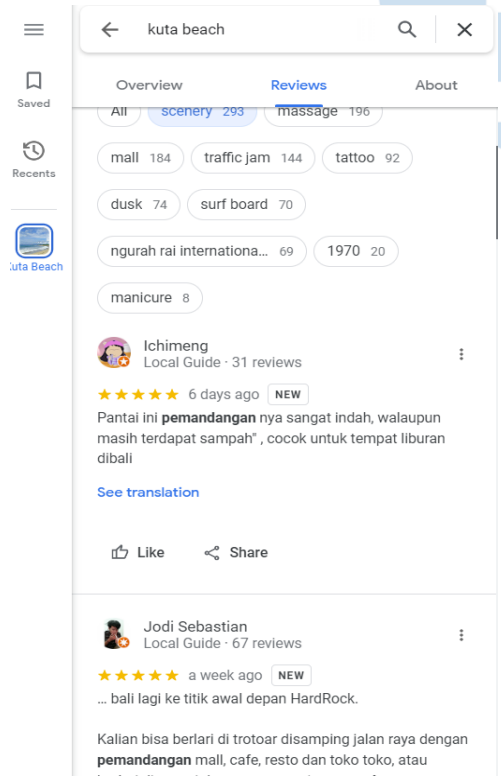

Fig. 3. Classification of Reviews Based on the Appearance of the Word “Massage” Category

IV. Result

A. Tools Used

To carry out the semantic similarity process, the SBERT and Cosine Similarity algorithms are used. Sentence-BERT (SBERT) leverages the advanced performance of BERT, with a different architecture. This allows things like cosine similarity to be found more quickly. For example, searching for sentence similarities for 65 hours on BERT would take 5 seconds with SBERT [10].

To implement the SBERT algorithm, we use SentenceTransformers (https://www.sbert.net/) in Python. SentenceTransformers is a Python framework for sophisticated embedding of sentences, text, and images. This framework can be used to calculate sentence/text embedding for more than 100 languages. This embedding can then be compared, e.g., with cosine-similarity to find sentences with similar meanings. This can be useful for semantic textual similarity, semantic search, or paraphrase mining.

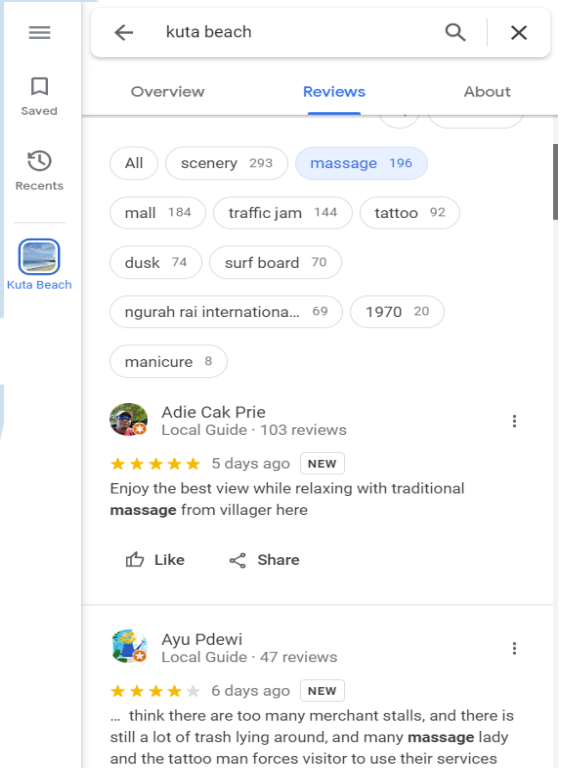

Fig. 4. Classification of Reviews Based on the Appearance of the Word "Massage" Category

Installing Kali Linux requires basic knowledge of The sentiment analysis process is carried out using the TextBlob Library (https://pypi.org/project/textblob/). TextBlob is a python library used to analyze data in the form of text. After going through this process, visitor review data is obtained based on the category and sentiment analysis results (positive, negative, or neutral).

B. Dataset

The dataset used in this study was taken from visitor review data at Kuta Beach, Bali, Indonesia which can be publicly accessed via the URL https://maps.google.com/. Visitor review data totals the top 120 reviews sorted by the most recent review. The original review data consists of various languages, such as English, Indonesian, etc. which are then translated into English before being processed. The review data contains impressions from visitors who have visited Kuta Beach, Bali.

C. Experiments

In the Semantic Similarity Analysis experiment that was carried out, the Cosine Similarity calculation was carried out between words in the specified category and sentences that were already available in the corpus. The cosine similarity results range from minus numbers to positive numbers. The higher the result of cosine similarity, the more similar in meaning the two sentences being compared. In this experiment, the cosine similarity limit for sentences that are considered similar is above 0. Only those sentences will be processed in the sentiment analysis process.

The sentiment analysis process carried out with the TextBlob Library is carried out by calculating polarity. After that, the polarity calculation will return the levels of polarity. Polarity lies between (-1, +1) with -1 meaning negative sentiments and +1 meaning positive sentiments. Polarity is reversed by negation words[14].

TABLE I. PERCENTAGE OF SENTIMENT ANALYSIS PER CATEGORY

No	Category	Sentiment Positive	Sentiment Negative	Sentiment Neutral
1	Scenery	0,908333333	0,075	0,016666667
2	Massage	0,903508772	0,078947368	0,01754386
3	Mall	0,907563025	0,075630252	0,016806723
4	Traffic Jam	0,907563025	0,075630252	0,016806723
5	Tattoo	0,901785714	0,080357143	0,017857143
6	Dusk	0,906779661	0,076271186	0,016949153
7	Surfboard	0,908333333	0,075	0,016666667
8	Ngurah Rai Int Airport	0,929824561	0,052631579	0,01754386
9	Manicure	0,91011236	0,06741573	0,02247191
10	Facilities	0,908333333	0,075	0,016666667
11	Location	0,908333333	0,075	0,016666667
12	Meals	0,910714286	0,071428571	0,017857143
13	Sanitation	0,9	0,081818182	0,018181818

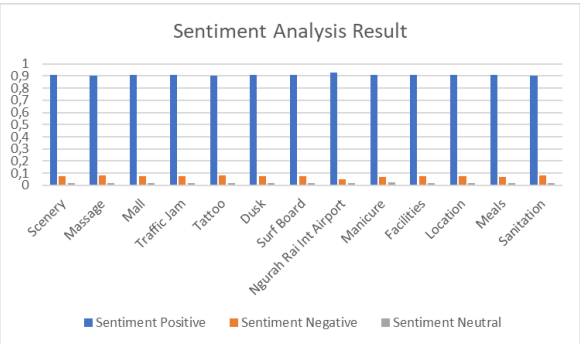

Fig. 5. Sentiment Analysis Result

The results of the sentiment analysis can be seen that positive, negative, and neutral sentiments have similar percentages for each category. This is because the designed algorithm cannot cut sentences for only one category that is assessed. One review being analyzed may contain more than one category, as shown in Figure 5. In this example, the review given by a visitor is a collection of paragraphs. Each paragraph consists of several sentences, where each sentence is the result of a review for one or more types of categories, including scenery, expense, massage, etc.

The algorithm designed in this study is only able to analyze up to the semantic similarity stage but has not been able to cut which sentences are relevant to the category to be analyzed for sentiment. In other words, sentiment analysis will be assessed for one review. So, if there is a review that contains several categories at once, then the overall sentiment analysis will be taken. This is a limitation of this research and will be carried out as future works.

To be able to perform a more comprehensive analysis, it is necessary to improve the designed algorithm. The algorithm should be able to cut sentence by sentence on visitor reviews first. The sentence can then be classified into one or more categories. Sentences that have been classified per new category can then be determined by the sentiment analysis results.

V. CONCLUSION

In this research, a user-generated content analysis method has been produced in Visitor Reviews of Tourism Objects Using Semantic Similarity. The algorithm designed in this study has limitations, that is only able to analyze up to the semantic similarity stage but has not been able to cut which sentences are relevant to the category to be analyzed for sentiment. In other words, sentiment analysis will be assessed for one review. So, if there is a review that contains several categories at once, then the overall sentiment analysis will be taken. This is a limitation of this research and will be carried out as future works.

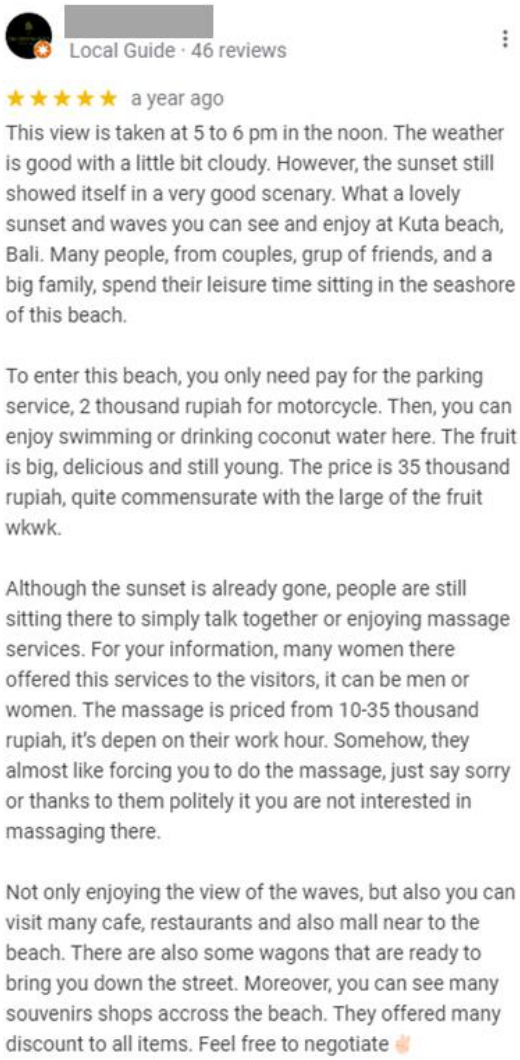

Fig. 6. Example of a review consisting of several categories

As future works, it is necessary to improve the algorithm designed in this study. The algorithm needs to be designed to be able to cut reviews consisting of several sentences, to then be classified based on the review category. Thus, the resulting sentiment analysis can be more detailed, namely on a sentence-by-sentence basis and not generalized for an overall review.

REFERENCES

[1] C. Science and J. M. Moguerza, "Murga J . et al . (2020) A Sentiment Analysis Software Framework for the Support of Business Information Architecture in the Tourist Sector . In : Hameurlain A . et al . (eds) Transactions on Large-Scale Data- and Knowledge-Centered Systems XLV . Lect," vol. 12390, 2020.

[2] F. Kitsios, M. Kamariotou, P. Karanikolas, and E. Grigoroudis, "Digital marketing platforms and customer satisfaction: Identifying ewom using big data and text mining," *Applied Sciences (Switzerland)*, vol. 11, no. 17, 2021, doi: 10.3390/app11178032.

[3] J. Q. Dong and C. H. Yang, "Business value of big data analytics: A systems-theoretic approach and empirical test," *Information and Management*, vol. 57, no. 1, p. 103124, 2020, doi: 10.1016/j.im.2018.11.001.

[4] N. Balakrishnan, D. Pelusi, and S. Ganesan, "Special issue on ' Big Data Analytics and Deep Learning for E - Business Outcomes ,'" *Information Systems and e-Business Management*, vol. 18, no. 3, pp. 281–282, 2020, doi: 10.1007/s10257-020-00486-0.

[5] W. Lu and S. Stepchenkova, "User-Generated Content as a Research Mode in Tourism and Hospitality Applications: Topics, Methods, and Software," *Journal of Hospitality Marketing and Management*, vol. 24, no. 2, pp. 119–154, Feb. 2015, doi: 10.1080/19368623.2014.907758.

[6] N. Afriliana, N. M. S. Iswari, and Suryasari, "Sentiment Analysis of User-Generated Content: A Bibliometric Analysis," *Journal of System and Management Sciences*, vol. 12, no. 6, pp. 583–598, 2022, doi: 10.33168/JSMS.2022.0634.

[7] N. M. S. Iswari, N. Afriliana, and Suryasari, "User-Generated Content Extraction: A Bibliometric Analysis of the Research Literature (2007–2022)," *Journal of Hunan University Natural Sciences*, vol. 49, no. 11, pp. 120–126, Nov. 2022, doi: 10.55463/issn.1674-2974.49.11.14.

[8] H. Wen, E. Park, C. W. Tao, B. Chae, X. Li, and J. Kwon, "Exploring user-generated content related to dining experiences of consumers with food allergies," *Int J Hosp Manag*, vol. 85, no. August, 2020, doi: 10.1016/j.ijhm.2019.102357.

[9] Y. Liu, A. Soroka, L. Han, J. Jian, and M. Tang, "Cloud-based big data analytics for customer insight-driven design innovation in SMEs," *Int J Inf Manage*, vol. 51, 2020, doi: 10.1016/j.ijinfomgt.2019.11.002.

[10] D. Kaplan, "What Does Semantic Mean In Computer Science." https://enjoymachinelearning.com/blog/finding-semantic-similarity-between-sentences-in-python/ (accessed Apr. 17, 2023).

[11] Baeldung, "Semantic Similarity of Two Phrases." https://www.baeldung.com/cs/semantic-similarity-of-two-phrases (accessed Apr. 17, 2023).

[12] adminlp2m, "Analisis Semantik - Definisi, Cara Kerja dan Contohnya," 2023. https://lp2m.uma.ac.id/2023/02/11/analisis-semantik-definisi-cara-kerja-dan-contohnya/#:~:text=Analisis%20Semantik%2C%20diungkapkan%2C%20adalah%20proses,paragraf%2C%20atau%20bahkan%20seluruh%20manuskrip. (accessed May 04, 2023).

[13] RN Is Here, "Analisa Sentimen Ulasan di Tokopedia - 1," 2021. https://www.kaggle.com/code/raniaprillyap/analisa-sentimen-ulasan-di-tokopedia-1/data (accessed Oct. 12, 2022).

[14] A. A. Chaudhri, S. S. Saranya, and S. Dubey, "Implementation Paper on Analyzing COVID-19 Vaccines on Twitter Dataset Using Tweepy and Text Blob," 2021. [Online]. Available: http://annalsofrscb.ro.