Knowledge Engineering and Data Science (KEDS)
Vol 5, No 2, December 2022, pp. 143–149

pISSN 2597-4602
eISSN 2597-4637

Social Media Mining with Fuzzy Text Matching: A Knowledge Extraction on Tourism After COVID-19 Pandemic

Ida Bagus Putra Manuaba[a,1,*], I Wayan Budi Sentana[b,2], I Nyoman Gede Arya Astawa [a,3], I Wayan Suasnawa [a,4], I Putu Bagus Arya Pradnyana [a,5]

[a] *Electrical Engineering Department, Politeknik Negeri Bali, Kampus Jimbaran, Badung, Bali, 80361 Indonesia*
[b] *School of Computing, Macquarie University, 4 Research Park Dr, Macquarie Park NSW 2113, Australia*
[1] *manuabaputra@pnb.ac.id*;* [2] *i-wayan-budi.sentana@students.mq.edu.au, i-wayan-budi.sentana@hdr.mq.edu.au;*
[3] *arya_kmg@pnb.ac.id ;* [4] *suasnawa@pnb.ac.id;* [5] *bagusarya12@pnb.ac.id*
** corresponding author*

ARTICLE INFO

Article history:
Received 26 October 2022
Revised 3 November 2022
Accepted 4 December 2022
Published online 30 December 2022

Keywords:
Social Media
Text Mining
Fuzzy Matching
COVID-19
Tourism

ABSTRACT

Social media mining is an emerging technique for analyzing data to extract valuable knowledge related to various domains. However, traditional text matching techniques, such as exact matching, are not always suitable for social media data, which can contain spelling mistakes, abbreviations, and variations in the use of words. Fuzzy matching is a text matching technique that can handle such variations and identify similarities between two texts, even if there are differences in spelling or phrasing. The gap in existing research is the limited use of fuzzy matching in social media mining for tourism recovery analysis. By applying fuzzy matching to social media data related to COVID-19 and tourism recovery, this research seeks to bridge this gap and extract valuable insights related to the impact of the pandemic on tourism recovery. We manually retrieved 19,462 Twitter records and differentiated the data sources using four diver parameters to indicate data related to the impact of COVID-19 on the tourism industry, such as the economy, restrictions, government policies, and vaccination. We conducted text mining analysis on the collected 7,352 words and identified 25 highly recommended words that indicated COVID-19 recovery from a tourism perspective. We separated the four words representing the tourism perspective to perform fuzzy matching as a dataset. We then used the inbound dataset on the fuzzy matching process, with the 7,352-word data collected from the text mining process. The matching process resulted in 18 words representing COVID-19 recovery from a tourism perspective.

I. Introduction

The COVID-19 pandemic has forced governments worldwide to restrict the movement of populations, bringing economic activity to a total standstill [1]. Governmental policies restricting mobility, such as bans, lockdowns, and social distancing, have obstructed the tourism industry [2]. According to the United Nations World Tourism Organization, global international tourist arrivals have fallen by 74%, resulting in a loss of international tourism receipts of about US$1.3 trillion [3]. While social media is a leading source of instant data representing human expression, the methods for analyzing data retrieved from social media about COVID-19 recovery from a tourism perspective are limited.

The primary objectives of this study were (i) to extract knowledge about COVID-19 recovery from a tourism perspective, as reported on social media, and (ii) to design a text mining and fuzzy matching approach for collecting data on this topic. Twitter was used as the social media platform for data collection due to its availability to the public and ease of collecting massive amounts of data as a dataset [4].

This study mined tweets using four different parameters related to COVID-19 recovery from the tourism perspective. The parameters were a vaccine, travel, restriction, and work. Text mining was used to analyze the collected dataset. This method has been previously used to investigate diseases

https://doi.org/10.17977/um018v5i22022p143-149

and chemicals related to COVID-19 [5], the impact of the COVID-19 pandemic on business [6], and public attention about COVID-19 on social media [7].

The contribution of this research is to provide a better understanding of how COVID-19 recovery impacts the tourism industry and to demonstrate the potential of text mining and fuzzy matching techniques in extracting insights from social media data. The research also presents a methodology for data cleaning, tokenization, filtering, and n-gram generation that can be applied to other social media mining studies. Additionally, the fuzzy matching process presented in this research can help identify similar words and phrases that may not have been captured in single-word data collection, thereby providing a more comprehensive understanding of the studied topic. Overall, the research aims to contribute to the knowledge of COVID-19 recovery in the tourism industry and provide a practical methodology for extracting insights from social media data.

II. Method

A. Data Collecting

The primary goal of this study was to retrieve data from Twitter to create a dataset related to COVID-19 recovery from a tourism perspective. The study used a keyword-based approach to retrieve data from Twitter based on parameters related to the tourism perspective, such as a vaccine, travel, restriction, and work. Data were retrieved from Twitter for all parameters, resulting in 19,462 records. Data retrieval for each parameter was limited to a maximum of 5,000 records to optimize the performance of the text-mining analysis process. The details of the data retrieved at the beginning of 2021 for each parameter are presented in Table 1.

After retrieving the data, the collected data from each parameter were combined into one dataset. Before conducting text mining analysis, the dataset was cleaned using several methods to produce a

Table 1 Data parameter on Twitter

Parameter Name	Number Data Retrieve on Twitter
Vaccine	4607 record
Work	5000 record
Restriction	4.877 record
Tourism	4.978 record

dataset related to the COVID-19 recovery in the tourism industry. Removing HTML links from the dataset records was a part of the dataset cleansing process. It was essential to remove duplicate records to prevent redundant data in the dataset. At the end of the dataset-cleaning process, there were approximately 7,352 records.

B. Social Media Text Mining

In this study, social media text mining refers to text mining on datasets retrieved from social media. Twitter was chosen as the social media platform to be used as a data source due to its public availability and ease of collecting massive amounts of data as a dataset. Text mining is generally used on social media to identify public perception [8], mine information [9], or investigate problems on social media [10].

Several text mining processes were applied to extract knowledge from the data retrieved on Twitter, such as tokenization, case transformation, stop word filtering, token filtering, and generating n-grams. The tokenization process divides the text into specific parts or tokens, transforming all characters into lower cases. The stop word filtering process selects important words using a stop list algorithm. N-gram items can sequence contiguous items, which group items within a sequence of text most frequently extracted from text or speech corpora. N-Gram can generate patterns of word occurrences related to COVID-19, which helps classify tweets as positive or negative for COVID-19 [11].

C. *Fuzzy Matching*

In the mining process, fuzzy logic is generally used to enrich and treat uncertainty [12], setting association rules mining for generating classifiers [13] and bridging the gap between the ambiguities of different understandings [14]. The fuzzy matching approach was slightly better than the weighted approach [15]. Fuzzy logic is also a promising approach that can significantly improve accuracy [16]. Fuzzy matching matches entities with entities contained in the provided entity dictionary. The study shows that the proposed method results in an entity recognition accuracy of 86.69% and an entity disambiguation accuracy of 88.69% [17]. Fuzzy matching can help improve large-scale data integration efficiently and accurately [18]. Fuzzy string matching algorithms are applied in various applications, including information retrieval, data cleaning, and natural language processing. Fuzzy string matching techniques include Levenshtein distance, Jaro-Winkler distance, and n-grams [19]. Fuzzy string matching also can be a valuable tool in automating the assessment of listener transcripts in speech intelligibility studies, reducing the time and effort required for manual assessment and improving the accuracy of the assessment process [20].

Based on the description above, this research uses fuzzy matching using the Jaccard Similarity approach to improve data quality and increase the accuracy of information extraction. The steps of fuzzy matching using Jaccard similarity are:

1. Tokenization: Convert the text into tokens (words and phrases).
2. Preprocessing: Apply various preprocessing techniques such as removing stopwords, converting text to lowercase, and removing special characters.
3. Creating n-grams: Create n-grams (a sequence of n tokens) from the tokens.
4. Calculating Jaccard similarity: Calculate the Jaccard similarity between the n-grams of two strings. The Jaccard similarity is the size of the intersection divided by the size of the union of two sets.
5. Fuzzy matching: Compare the Jaccard similarity scores between the n-grams of two strings and determine if they match based on a predefined threshold.

The sample dataset for matching was limited to 25 records of data, and during the matching process, only 10 similar data were allowed to be matched for each sample dataset.

III. Result and Discussion

The social media text mining process with fuzzy matching involves four main steps: retrieving data from the data source, text mining, creating a dataset sample for fuzzy matching, and performing the fuzzy matching process. The raw data source is in an Excel format, combining data from four different parameters related to COVID-19 and tourism perspectives. The data source is cleaned and transformed into a usable dataset for the data retrieval process. Before the text mining process, the dataset is converted from nominal to text. Text mining involves several steps: tokenization, case transformation, stop word filtering, token filtering, and n-gram generation. The output of the text mining process is then used to create a dataset sample for fuzzy matching. Simple algorithms such as data sorting, attribute selection, example filtering with range, and example filtering are used to collect dataset samples for fuzzy matching. These sample datasets are related to parameters based on COVID-19 recovery and tourism perspectives. The dataset samples for matching are limited to 25 records, and the number of matches is limited to 10 similar data for each dataset sample. The primary process is illustrated in Figure 1.

The dataset sample for fuzzy matching is compared with the dataset from the text mining process using fuzzy matching. The Jaccard similarity threshold is the minimum similarity value between two tokens to be considered a match. A threshold of 0.6 means that strings with a Jaccard similarity of at least 0.6 will be considered a match. Here is a pseudocode for the fuzzy matching process.

```
#fuzzy matching using Jaccard similarity:
Input:
  - String s1
  - String s2
  - Integer k (the size of the k-grams)
```

```
Output:
  - Jaccard similarity score (a value between 0 and 1)

Algorithm:
1. Create sets of k-grams for both s1 and s2
   - Split s1 into k-grams and store them in set A
   - Split s2 into k-grams and store them in set B
2. Calculate the intersection of A and B (i.e., the number of k-grams that
appear in both sets)
3. Calculate the union of A and B (i.e., the number of k-grams that appear
in either set)
4. Calculate the Jaccard similarity score as the ratio of the intersection
to the union:
      J(A,B) = |A ∩ B| / |A ∪ B|
5. Return the Jaccard similarity score
```

In this pseudocode, |A| and |B| denotes the size of sets A and B, respectively. The k parameter determines the size of the k-grams, which are contiguous sequences of k characters from the input strings. The Jaccard similarity score is between 0 and 1, where 1 indicates a perfect match between the two strings, and 0 indicates no similarity.

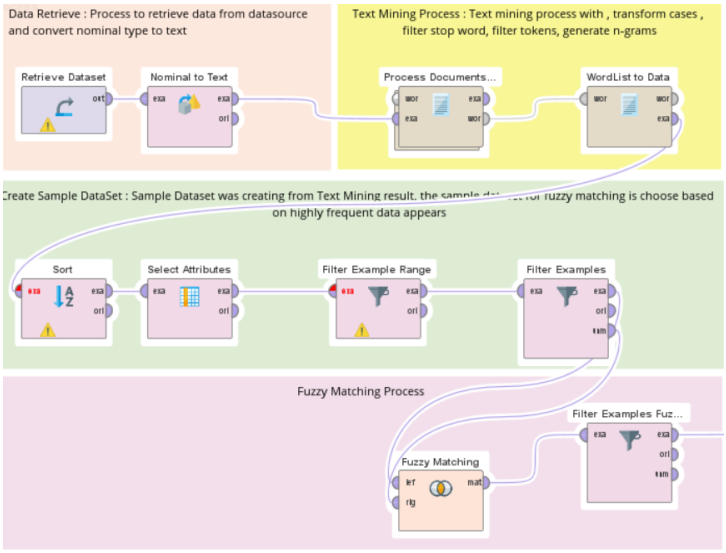

Fig. 1. Social media text mining analysis with fuzzy matching

Before cleaning the data source, it consisted of 19,462 records, but after the cleaning process, it was reduced to 7,352 records. Various techniques were used to clean the data source, including removing HTML links, removing Twitter tags, removing words similar to COVID-19, and eliminating duplicate data records. The text mining process extracted 10,173 words with varying frequency levels from the data source.

Several processes involve tokenization, case transformation, stop word filtering, token filtering, and generating n-grams to extract knowledge from Twitter data using text mining. The first step, tokenization, is dividing a text into specific parts or tokens. The next step is case transformation, where all characters are transformed into lowercase. In token filtering, additional rules are added, such as a minimum of 4 characters and a maximum of 25 characters. N-grams are used to sequence contiguous items. Persistent words are extracted using a word extraction process to analyze the data. The resulting data is shown in Figure 2.

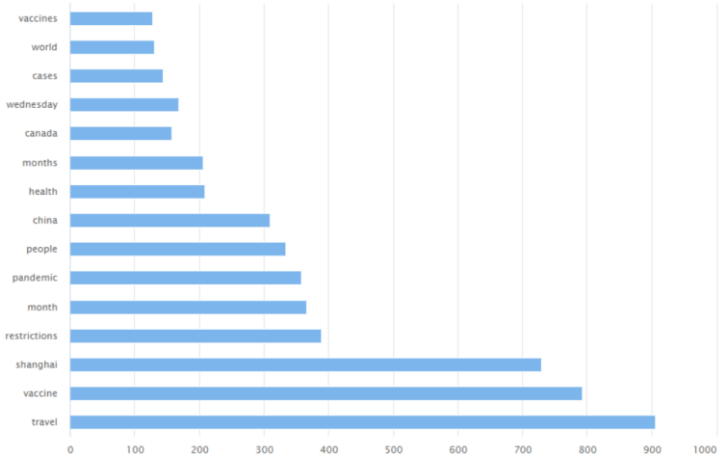

Fig. 2. Result of text mining process

Figure 2 shows that the most frequent words are "travel", "vaccine", "Shanghai", "restrictions", "month", "pandemic", "people", "China", and "health". These words are then collected into a dataset sample for fuzzy matching. Additional rules were added to the fuzzy matching process, such as a maximum of 10-word results for each sample dataset and a minimum similarity value of 50. The results of the fuzzy matching process are shown in Table 2.

Table 2. Result of Fuzzy Matching

Sample Dataset	Word Similarity	Similarity	Sample Dataset	Word Similarity	Similarity
travel	take	60.0	restrictions	travel_restrictions	100.0
travel	state	55.0	month	months	91.0
travel	travel_restrictions	100.0	month	ending_month	100.0
travel	travellers	75.0	month	month_lockdown	100.0
vaccine	chinese	57.0	month	north	60.0
vaccine	vaccines	93.0	people	reopening	53.0
vaccine	vaccinated	82.0	china	vaccinated	53.0
vaccine	vaccination	67.0	china	coming	55.0
vaccine	airlines	53.0	china	chinese	67.0
restrictions	authorities	52.0	china	children	62.0
restrictions	getting	53.0	china	china_largest	100.0
restrictions	residents	57.0	china	canada	55.0
restrictions	return	56.0	china	india	60.0
restrictions	testing	53.0	health	deaths	67.0

The data from Table 2 is visually represented in a word cloud, where words with greater prominence appear more frequently. The words with a similarity value greater than 75 are travel_restrictions, china_largest, ending_month, month_lockdown, vaccines, vaccinated, month, and travelers. Among these, travel_restrictions appears twice with a similarity of 100. When visualized as a word cloud, the data is shown in Figure 3.

Table 3 shows the data comparison result without fuzzy matching and with the fuzzy matching process. Without fuzzy matching, the process only has to collect single-word data, and with fuzzy matching, the process can be explored more deeply. The process data, with fuzzy matching, appears more than single-word data result such as travel restriction, month lockdown, and ending months. Fuzzy matching can appear similarity words on highly frequent appears, such as on travel word, this process shows the similarity words travelers and travel restriction, on vaccine word this process shows the similarity words: vaccines, vaccinated, and vaccination.

Fig. 3. Word cloud of text mining with fuzzy matching

Table 3. Text mining with or without fuzzy matching

Without Fuzzy Matching		With Fuzzy Matching	
travel	china	travel_restriction	authorities
vaccine	health	travelers	month_lockdown
sanghai	month	chinese	reopening
restrictions	canada	vaccines	getting
month	wednesday	vaccinated	ending_months
pandemic	cases	vaccination	resident
people	world	airlines	testing

Table 3 compares the results obtained without the fuzzy matching process and those obtained using it. Only single-word data was collected when the fuzzy matching process was not used. However, a deeper data exploration was made possible by utilizing fuzzy matching. The fuzzy matching process revealed more than just single-word data, including insights on travel restrictions, month-long lockdowns, and ending months. Fuzzy matching also identified similar words that frequently appeared. For example, the word "travel" is similar to the words "travelers" and "travel restriction," while "vaccine" is similar to "vaccines", "vaccinated", and "vaccination".

IV. Conclusion

The study highlights the importance of analyzing social media data to gain insights into the impact of the pandemic on the tourism industry. The study uses text mining and fuzzy matching techniques to extract relevant data from Twitter and analyze it. The results show that the recovery of the tourism industry is a topic of discussion on Twitter and that certain words are frequently mentioned in this context, such as travel restrictions, vaccines, and lockdowns. This study contributes to the growing literature on social media mining and its applications in the tourism industry. It provides insights that can be useful for policymakers and businesses to make informed decisions regarding the recovery of the tourism industry in the context of the ongoing pandemic. The study can be extended to other social media platforms, such as Instagram or Facebook, to gain a more comprehensive understanding of how COVID-19 recovery is affecting the tourism industry. The use of more advanced natural languages processing techniques, such as sentiment analysis or topic modeling, can provide more nuanced insights into the attitudes and opinions of social media users toward COVID-19 recovery in the tourism industry.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

Publisher's Note: Department of Electrical Engineering - Universitas Negeri Malang remains neutral with regard to jurisdictional claims and institutional affiliations.

References

[1] M. Nicola et al., "The socio-economic implications of the coronavirus pandemic (COVID-19): A review," Int. J. Surg., vol. 78, pp. 185–193, Jun. 2020.
[2] M. Sigala, "Tourism and COVID-19: Impacts and implications for advancing and resetting industry and research," J. Bus. Res., vol. 117, pp. 312–321, Sep. 2020.
[3] UNWTO, "2020: A year in review," World Tourism Organization, 2020. (Access on 29 October 2022)
[4] J. X. Koh and T. M. Liew, "How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds," J. Psychiatr. Res., vol. 145, pp. 317–324, Jan. 2022.
[5] A. Karami, B. Bookstaver, M. Nolan, and P. Bozorgi, "Investigating diseases and chemicals in COVID-19 literature with text mining," Int. J. Inf. Manag. Data Insights, vol. 1, no. 2, p. 100016, Nov. 2021.
[6] P. Carracedo, R. Puertas, and L. Marti, "Research lines on the impact of the COVID-19 pandemic on business. A text mining analysis," J. Bus. Res., vol. 132, pp. 586–593, Aug. 2021.
[7] K. Hou, T. Hou, and L. Cai, "Public attention about COVID-19 on social media: An investigation based on data mining and text analysis," Pers. Individ. Dif., vol. 175, p. 110701, Jun. 2021.
[8] J. Y. Park, E. Mistur, D. Kim, Y. Mo, and R. Hoefer, "Toward human-centric urban infrastructure: Text mining for social media data to identify the public perception of COVID-19 policy in transportation hubs," Sustain. Cities Soc., vol. 76, p. 103524, Jan. 2022.
[9] A. Kang et al., "Environmental management strategy in response to COVID-19 in China: Based on text mining of government open information," Sci. Total Environ., vol. 769, p. 145158, May 2021.
[10] S. Luo and S. Y. He, "Understanding gender difference in perceptions toward transit services across space and time: A social media mining approach," Transp. Policy, vol. 111, pp. 63–73, Sep. 2021.
[11] N. Nasser, L. Karim, A. El Ouadrhiri, A. Ali, and N. Khan, "n-Gram based language processing using Twitter dataset to identify COVID-19 patients," Sustain. Cities Soc., vol. 72, p. 103048, Sep. 2021.
[12] C. Fernandez-Basso, K. Gutiérrez-Batista, R. Morcillo-Jiménez, M.-A. Vila, and M. J. Martin-Bautista, "A fuzzy-based medical system for pattern mining in a distributed environment: Application to diagnostic and co-morbidity," Appl. Soft Comput., vol. 122, p. 108870, Jun. 2022.
[13] D. Rohidin, N. A. Samsudin, and M. M. Deris, "Association rules of fuzzy soft set based classification for text classification problem," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 3, pp. 801–812, Mar. 2022.
[14] S. Rameem Zahra, M. Ahsan Chishti, A. Iqbal Baba, and F. Wu, "Detecting Covid-19 chaos driven phishing/malicious URL attacks by a fuzzy logic and data mining based intelligence system," Egypt. Informatics J., vol. 23, no. 2, pp. 197–214, Jul. 2022.
[15] C. Peng, P. Goswami, and G. Bai, "Fuzzy Matching of OpenAPI Described REST Services," Procedia Comput. Sci., vol. 126, pp. 1313–1322, 2018.
[16] Ida Bagus Putra Manuaba, Komang Ayu Triana Indah, Muhammad Fahmi, and Irma Nuraeni Salsabila, "An Improvement Object Detection Method Findcontour with Fuzzy Logic for Detect Balinese Script Object," Aptisi Trans. Technopreneursh., vol. 4, no. 3, pp. 257–262, Oct. 2022.
[17] M. Singh, M. Kumar, and J. Malhotra, "Energy efficient cognitive body area network (CBAN) using lookup table and energy harvesting," J. Intell. Fuzzy Syst., vol. 35, no. 2, pp. 1253–1265, Aug. 2018.
[18] L. Guan-Feng and M. Zong-Min, "An efficient matching algorithm for fuzzy RDF graph," J. Inf. Sci. Eng., vol. 34, no. 2, pp. 519–534, 2018.
[19] M. Pikies and J. Ali, "Analysis and safety engineering of fuzzy string matching algorithms," ISA Trans., vol. 113, pp. 1–8, Jul. 2021.
[20] H. R. Bosker, "Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies," Behav. Res. Methods, vol. 53, no. 5, pp. 1945–1953, Oct. 2021.