THAMMASAT REVIEW
PRINT ISSN: 0859-5747 ONLINE ISSN: 2630-0303

Volume 23 No 2 (July-December) 2020 [Page 280-301]

Big Data and Social Media Qualitative Research Methodology

Papon Chongthanavanit [a,*], John M. Kennedy [b] and Jantima Kheokao [a]

[a] *School of Communication Arts, The University of the Thai Chamber of Commerce, Thailand*

[b] *College of Arts and Sciences, Indiana University, United States of America*

Received 2 June 2020; Received in revised form 23 September 2020

Accepted 5 November 2020; Available online 22 December 2020

Abstract

The paper demonstrates and evaluates a methodology for conducting big data and social media research using qualitative methods. Large-scale databases of customer-generated content in social media have captured scientific attention, producing an abundance of valuable information. However, only a few studies have used a qualitative approach to analyze big data. This paper presents the step by step process of managing big data and qualitative analyses integrating computational approaches. It is based on a study of dental tourism in Thailand as a case example to validate as well as to highlight the advantages and limitations of the methodology.

Keywords

Big data, Social media, Qualitative research

*** Corresponding author:** icemantop@gmail.com

DOI: 10.14456/tureview.2020.20

Introduction

An encompassing definition of big data was perhaps best coined by De Mauro, Greco, and Grimaldi (2016) as "the information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value" (De Mauro et al. ,2016, p.122). An ever increasingly high "volume" of data is supported by an improved "velocity" of data collection and processing through advancement in digital technologies as well as a "variety" of data formats/sources that can now be integrated into a unified management platform (Laney, 2001). With enhanced computer capacity scaling with Moore's law, the management and analysis of large databases, which traditionally required supercomputer capacity, can now be readily performed on a desktop computer, thus extending opportunities for applications of big data technologies in multiple disciplines including social science research (Lazer et al., 2009; Manovich, 2011).

For qualitative analysis researchers, an exponential increase of social networks, reviews, and booking websites has provided a nearly limitless avenue of data which could potentially be mined to analyze people's opinions from around the world on subject matters such as customers' perceptions of brand image, goods and services. However, qualitative methodologies to systematically analyze the vast volume and variety of formats from several sources of text-based information, e.g., customers' comments on digital websites, in an automated manner are still lacking.

As Thailand is ranked as the third most favored country in the world for medical and dental tourism with cost savings of 50-75% compared to the US, it is a significant selling point for dental services in Thailand that it can compete with other countries (Sutherland, 2017). To achieve a competitive edge, market specialists should understand customer satisfaction and capitalize on key advantages in developing a novel marketing communication strategy to establish a brand for dental health tourism.

This study describes a methodological framework for qualitative content analysis within the context of massive information from social media content, through an analysis of customers' perception of dental tourism in Thailand as a case study.

Literature Review

Social Media

Social media are interactive computer-mediated technologies that facilitate creation or sharing of information, ideas, career interests and other forms of expression via virtual communities and networks (Obar & Wildman, 2015). Social media have common features are as follows. (1) Social media are interactive web internet-based applications. (2) User-generated/shared contents (e.g., text posts or comments, digital photos or videos, and data generated through all online interactions) are the essential component of social media. (3) Users create service-specific profiles and identities for the website or applications that are designed and maintained by the social media organization. (4) Social media facilitate the development of online social networks by connecting a user's profile with those of other individuals or groups (Kaplan & Haenlein, 2010; Obar & Wildman, 2015). Owing to the fast-paced and constantly interactive nature of social media, tons of information are generated every second serving as a source of data that can be transformed into values data to be managed and analyzed insightfully.

Qualitative Research in an Era of Big Data

Qualitative research is a scientific methodology focusing on humanistic or idealistic information such as people's beliefs, experiences, attitudes, behavior, and interactions, etc. (Holsti, 1969). Qualitative research allows researchers to have the ability to explore a new dimension, compared to a quantitative approach, that cannot be obtained through numerical assessment of variables alone (Mills, 2018). However, there are some limitations to qualitative research which should be considered. First, a qualitative approach might take several months or years because this process probes into personal interaction for data collection. For example, it requires labor-intensive analysis processes such as categorization (coding), recoding, and manual transcription of interviews. Thus, there is no way for research to collect qualitative data with a large sample size. In years past, when content analysts faced the challenge of too much data, e.g., too many news articles or television transcripts to code, researchers often used random or stratified sampling methods to reduce the amount of information (Riff, Lacy, Fico, & Watson, 2019). In addition, it is difficult to identify logical causality (e.g., a specific action caused an event to happen) since qualitative research does not have control variables or confounders. It is impossible to analyze qualitative data with mathematics and statistics. The output from qualitative research is typically in the form of opinion and judgment rather than concrete results (Batrinca & Treleaven, 2015; Obar & Wildman, 2015). As a result, conclusions from one qualitative study may not generalize to others. However, in this digital era, these limitations can be overcome by adopting computerization in archiving, scraping, and analyzing qualitative data which allows researchers to readily perform qualitative studies on big data sources such as social media content (Gandomi & Haider, 2015; Lewis, Zamith, & Hermida, 2013).

Computational Methods as a New Tool for Scraping and Analysis of Qualitative Data

Some key technical terms related to the scraping and analysis of unstructured textual data are defined here (Batrinca & Treleaven, 2015):

Sample framing: Sample framing is a process where a list of all the items in population related to study is defined.

Opinion mining: Opinion mining (sentiment mining, opinion/sentiment extraction) is the area of research that utilizes computerized systems to analyze personal opinion from textual content written in natural language.

Scraping: Data scraping is a technique in which a computer program extracts data from another program and manages them to become human-readable data.

Sentiment analysis: Sentiment evaluation refers to the use of natural language processing, computational linguistics, and textual content analytics to extract subjective information in source materials.

Text analytics: Text analytics involve information retrieval, lexical evaluation to determine word frequency distributions, sample recognition, tagging/annotation, statistics extraction, and information mining strategies.

With technological advances especially in data science, computational methods allow the scraping and analysis of massive and complicated data that could not be readily accomplished by manual labor. It saves time and reduces human error in data scraping (Fuchs, Höpken, & Lexhagen, 2014; Miah, Vu, Gammack, & McGrath, 2017; Schroeder, 2014). Gandomi and Haider (2015) suggested big data analytical processes be segmented in two phases: (1) data management (acquisition and recording; extraction, cleaning, and annotation; and integration, aggregation, and representation) and (2) analytics (modeling and analysis, and interpretation).

In the first phase, due to a massive data source, opinion mining and scraping are laborious to perform manually. Programing can be executed to more efficiently perform these tasks (Demšar et al., 2012; Smedt & Daelemans, 2012). Programing codes can be written in programming languages of choice such as Python, Java and C (Manovich, 2011; Marine-Roig & Clavé, 2015; Raun, Ahas, & Tiru, 2016; Zamawe, 2015).

In the second phase, using software to perform qualitative content analysis was challenging in the past due to the limitations of the software to understand complex text-based information. The approaches for automated textual analysis draw from two competing ideas: lexical and non-lexical text classification. The lexical approach starts with a set of words, for which a typical sentiment (positive, negative, or neutral) is predefined. This set may be created manually or semi-automatically subjected to linguistic rules. On the other hand, the non-lexical approach is based on machine learning, meaning that these classifiers require "training" on the sentiment expressed in manually classified texts from the same domain. Consequently, the machine learning classifiers are known for poor performance when used in a domain different from the one on which they were trained (Kirilenko, Stepchenkova, Kim, & Li, 2018; Schmunk, Höpken, Fuchs, & Lexhagen, 2013). Automated sentiment analysis of user-generated content has been applied on tourism and hospitality data. For example, Schmunk et al. (2013) applied automated sentiment analysis tools both lexical and machine learning approach to reviews posted by visitors to a ski resort on the TripAdvisor.com (127 reviews) and booking.com (81 reviews). They found good performance with approximately 70% accuracy. Moreover, García-Pablos, Cuadros, and Linaza (2016) applied machine learning sentiment analysis tools to textual hotels reviews and found results accurate to 70-80%.

Nowadays, improvements in software, programs, technology, and electronic techniques of data coding have allowed analysis to be performed more accurately (Hilal & Alabri, 2013). In particular, NVivo, a Qualitative Data Analysis (QDA) computer software package produced by QSR International, has become a handy tool for qualitative researchers owing to its many useful features. For example, NVivo is equipped with powerful query tools to help uncover subtle trends. Its automated analysis features with machine learning allow deeper exploration of datasets. The analysis of qualitative data has become easier and has yielded more comprehensive results (Mahrt & Scharkow, 2013; Miah et al., 2017; Riff et al., 2019; Simon, 2001). The objective of this article is to demonstrate and evaluate a methodology for conducting big data and social media research using qualitative methods by using computerized forms of data mining and content analysis.

Methodology

This paper demonstrates and evaluates a methodology for conducting big data and social media research using qualitative methods by using computerized forms of data mining and content analysis. The four steps of processing and managing big data and qualitative analysis are (1) Sample framing, (2) Data scraping and management, (3) Analysis and interpretation of qualitative data with NVivo, and (4) Process validation. The steps are explained in detail and the methodology is described to analyze dental tourism in Thailand as an example

1. Sample framing

Sample framing is a process where a list of all the items in a population related to the study is defined. It is a complete list of every data item relevant to the sample on which a study focuses (Groves et al., 2011). In regard to social media content, social media websites provide rich databases from which samples could be extracted. Here, the TripAdvisor website was chosen to study dental tourism because it is one of the primary review websites that tourists from around the world visit to read and share their experiences. Then, the data were narrowed by using inclusive criteria to select the topics which included the appropriate keywords ("Dental" and "Thailand"). In total, the sample for this study comprises 1,430 topics (10 topics per page, 143 pages) (Figure 1A and B).

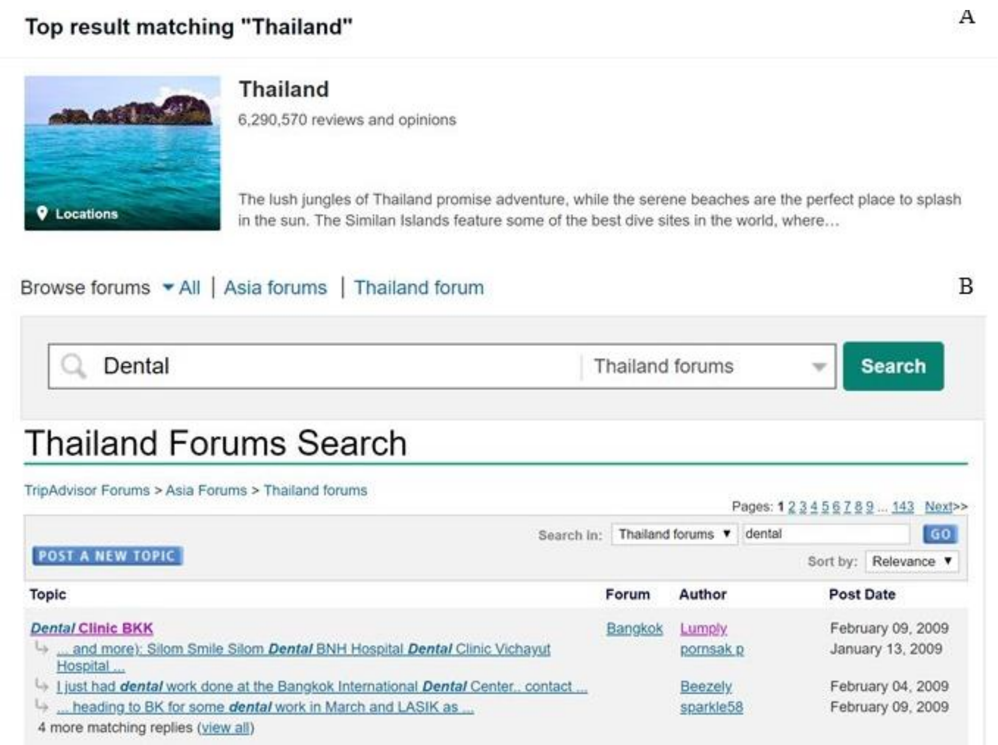

Figure 1 Sample Framing (A) A screenshot of the interface when searching "Thailand"; and (B) A screenshot of the interface when searching "Dental" in Thailand forums

2. Data scraping and management

Data scraping is a process of data gathering from the media of choice such as websites, social media, or news. Here, the focus is on web scraping since websites constitute a dominant source of social media content from which data can be readily extracted via programming. In this study, programming was written in Python to facilitate the data scraping process.

A coding form was designed to define and categorize the data of interest into a column format which is required for NVivo qualitative analysis. In this example, data on username, location, and opinions were collected and tabulated into an Excel file. To complete this task, a Python script was written to generate URLs of every sample topic in the scope of the "Dental" + "Thailand" sample frame within a defined period (from December 2011 to

March 2019) (Figure 2A). Then, a second script was used to perform a syntax search to locate and extract data on username, location, and reviews and opinions (Figure 2B). These data were cleaned to trim accessory syntax (Figure 2C). The process was completed in 47,553.2 seconds. The tabulated data was further polished via an Excel filter function to remove irrelevant content (e.g., TripAdvisor auto-generated messages) and to cluster by country providing a final table ready for NVivo analysis (Figure 2D). In total, 7,365 reviews and opinions were collected with Australia (3,096 comments), United States (939 comments), and United Kingdom (817 comments) as the three most predominant sources of comments, respectively.

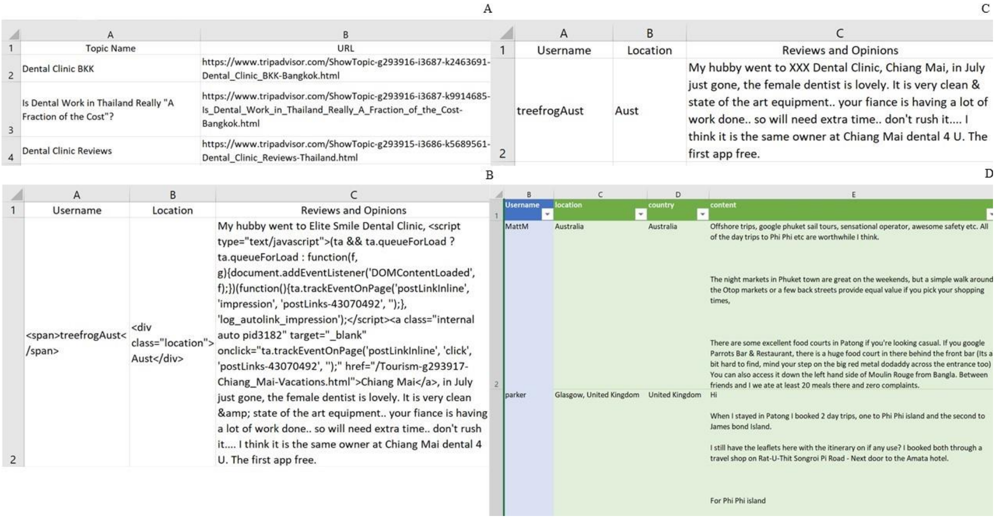

Figure 2 Data Scraping and Management (A) URLs of every topic sample; (B) Data with extraneous syntax (e.g., <, >, /, p, ", =); (C) Trimmed data; and (D) Final data clustered by country

3. Analysis and interpretation of qualitative data with NVivo

This section describes the methodology used to perform a qualitative analysis of social media content using the NVivo software. There are multiple aspects (e.g., finding keywords from the interview manuscript) at which text-based data could be mined and analyzed with the NVivo software. Here, the focus is on a generalized methodology to extract the overall impression of selected populations towards particular topics. Specifically, an example was developed of how to analyze the perception of international customers towards dental tourism in Thailand from different angles such as quality of service, price, and place. The Excel file was first imported to the NVivo program following the step-by-step software autosuggestion. The data were rechecked again to make sure they fit into appropriate columns and were correctly identified as either closed or open-ended questions (Figure 3). Close-ended questions have definite answers and are used for creating attributes (e.g., gender, age, and country) while open-ended questions are more descriptive in responses and are used for creating nodes (e.g., factors of interests or variables). In other words, close-ended questions were used for independent variables, but open-ended questions were used for dependent variables. In this case, "Content" was the only open-ended question used to create nodes (Figure 4). To evaluate the emotional tone of the "Content", an automated sentiment analysis was executed. The result was categorized into 4 levels, namely, "Very positive", "Moderately positive", "Moderately negative", and "Very negative" (Figure 5). The software created the sentiment nodes automatically.

Figure 3 A Screenshot of Rechecking the Data and Format

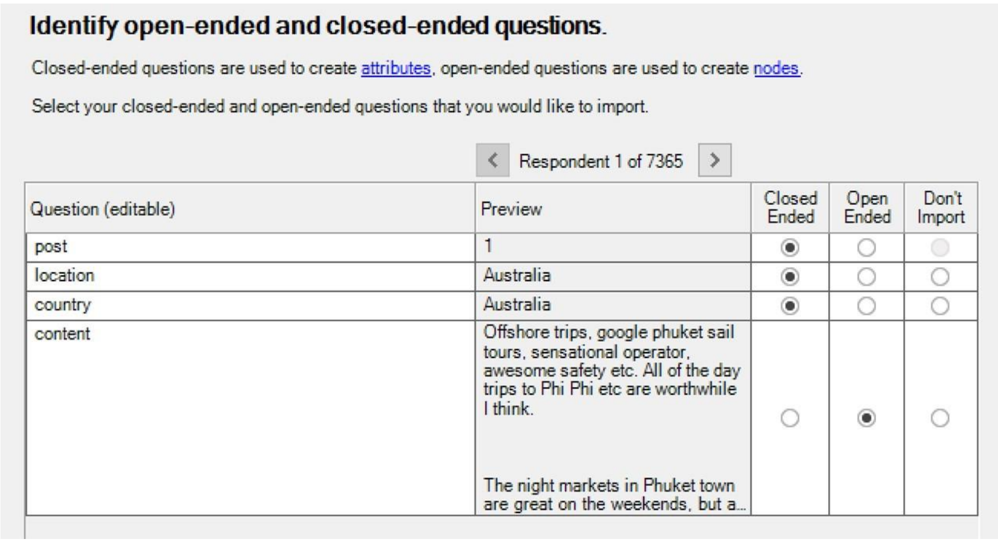

Figure 4 A Screenshot of Identifying Open-ended and Closed-ended Questions

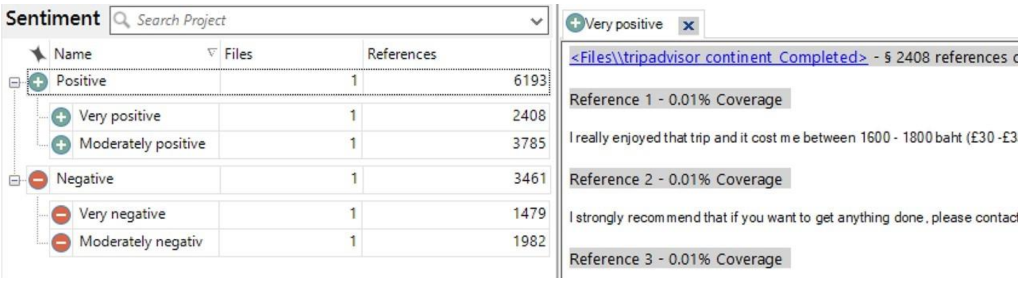

Figure 5 A Screenshot of Automated Sentiment Analysis

Study objectives must be clearly defined to perform content analysis appropriately. For example, some researchers may study how some factors influence other variables, or some may focus on keywords in their data. In this paper, frequency word analysis was used as a tool to manage the data and explore keywords. Then, frequency word count was used to identify keywords and matched with factors of interest. A frequency word count was performed using a query tool. The software generates a list of most frequent words in every review and opinion. There are more than a thousand words used in this data set. Only 200 most frequent words were selected. Then, each frequent word was classified into each factor. Some words that are not relevant to the factors were excluded (e.g., Bangkok, Thailand, looking) (Figure 6).

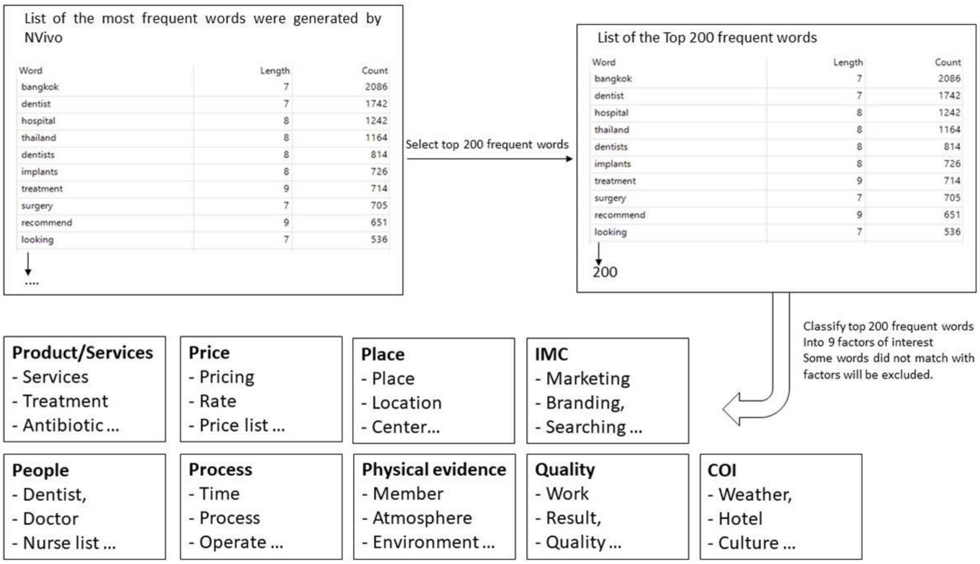

Figure 6 Diagram Showing Frequency Word Count Workflow

For an accurate analysis, it is important to carefully define the level of word grouping and the minimum length of the letters in the words. There are five levels of word grouping in order of stringency: (1) Exact matches (talk), (2) Stemmed-word (talking), (3) Synonyms (speak), (4) Specialization (whisper), and (5) Generalization (communication) (Figure 7). Choosing a broader level of word grouping could result in error? where the content is overinterpreted in an attempt to assign appropriate sentiment. Choosing a narrower level of word grouping could reduce coverage in the content analysis as some key words are missed. For this analysis on dental tourism, the word grouping at the stemmed-word level was found to be the right balance. Regarding the latter consideration, the minimum length of the words was selected at 7 letters to exclude commonly used words related to sentence structures such as is/am/are, a/an/the, and he/she/it, etc.

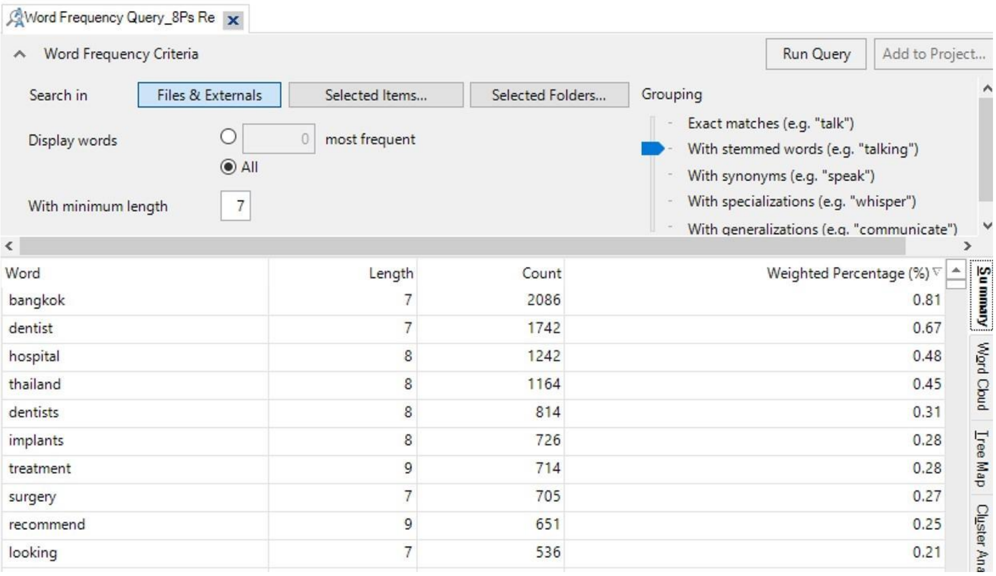

Figure 7 A Screenshot of Frequency Word Count

Factors of interest are listed which are country-of-origin image (COI), integrated marketing communication, people, physical evidence, place, price, process, physical evidence, product/service, and productivity/quality. Frequency words were analyzed and matched into each factor (Table 1). Next, a code for each factor was created by using "Text search query in the data set" in the query tool. Frequency words of factors were typed in the "Search for" tab (see the red square in Figure 8). Then the program explored all content containing the frequency words relevant to factors of interest and created codes (nodes).

Table 1 Coding by Using Frequent Keywords and Categorizing Them into Each Factor

Factors	Keywords
Product/Services	Services, Treatment, Medication, Insurance, Full Alternative, Antibiotics, Choice
Price	Pricing, Rate, Price list, Promotion, expensive, Saving, Money, Cheaper, Cheap, Affordable, Bargain, Budget, Cheapest, Discount, Worth
Place	Place, Location, Center, Convenient, Hard to go Inconvenient, shop, Branch, Building, Distance
Integrated marketing communication (IMC)	Reputation, Report, Feedback, Marketing, Branding, Searching, Websites, Research, Email, Advertising, Agent, Experienced, Google, Message, Offer, Online, Popular, Review, TripAdvisor
People	Dentist, Doctor, English, Friendly, Staff, Professional, Nurse, Periodontist, Endodontist, Skilled, Specialist, Surgeon
Process	Time, Process, Operate, Call, Telephone, Procedures, Inform, Afterward, Appointment, Arrange, Guide, Management, Prepare, Phone, Proceed, Quick, Schedule, System, Wait
Physical Evidence	Member, Atmosphere, Environment, Decoration, Relaxing, Facility, equipment, Technology, Modern, Uncomforted, Unclean, Training Room, Caring, Cleaning, Advance, Believe, Comfortable, Garden, Hygiene, Standard
Productivity & Quality	Work, Result, Quality, crown, implant, whitening, braces, Invisalign, Veneer, Surgery, Root, Filling, Bridge, Amalgam, Denture, Extraction, Outcome, Scaling, Whiting, Zoom
Country-of-origin image (COI)	Weather, Hotel, Culture, Politics, Social, Economic, Transportation, Attraction, Sea bleach, Mountain, Nightlife, Market, Food, Travel, Trip, Road, People, Local, Airport, Accommodation, Accommodate, Airline Bank, Breakfast, Buffet, City, Driving, Island, Massage, Mosquito, Passport, Pool, Resort, Ride, Restaurant, River, Shuttle, Skytrain, Snorkeling, Street, Sunscreen, Supermarket, Swimming, Taxi, Traffic, Transport, Visa, Market

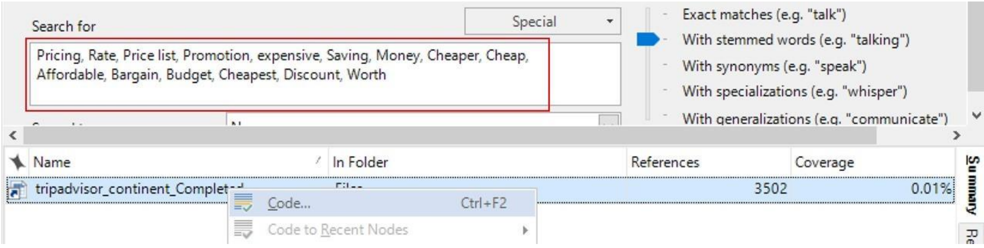

Figure 8 A Screenshot of Using "Text Search Query in the Data Set" to Create Codes (Nodes)

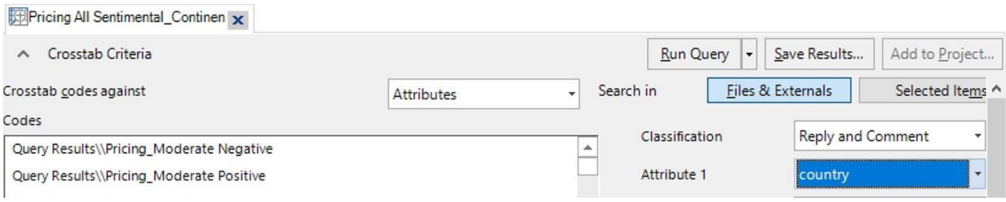

Figure 9 A Screenshot of Country Selected in Attribute.

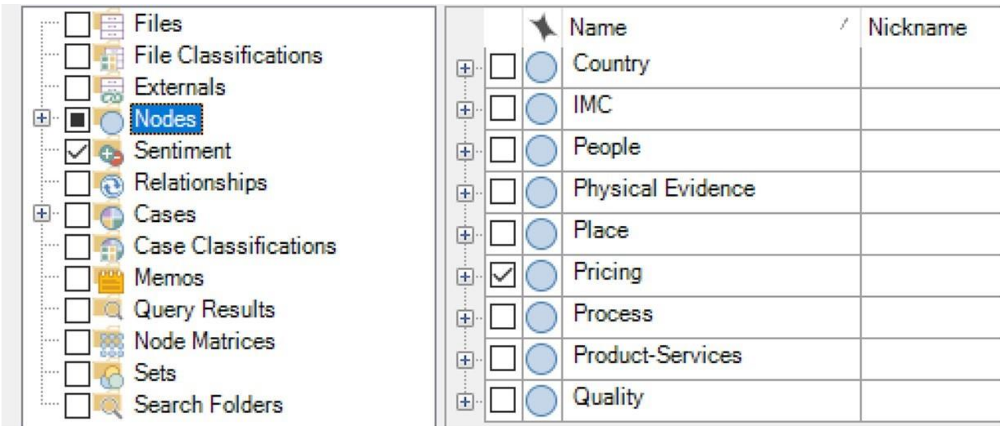

Figure 10 A Screenshot of Selecting Nodes, Sentiment

Reply and Comment	country = Australia (3096)	country = United Kingdom (817)	country = Thailand (882)
Pricing_Moderate Negative	94	18	26
Pricing_Moderate Positive	168	47	35
Pricing_Very Negative	48	17	20
Pricing_Very Positive	115	45	21
Total (unique)	356	110	82

Figure 11 A Screenshot of the Crosstab Results Presented in a Tabulated Data

Table 2 Example of Reported Crosstab Results

Reply and comment	*Australia (n = 3096)*		*United Kingdom (n = 817)*		*Thailand (n = 882)*	
	n	%	n	%	n	%
Pricing Very Negative	48	1.55%	17	2.08%	20	2.27%
Pricing Moderately Negative	94	3.04%	18	2.2%	26	2.95%
Pricing Moderately Positive	168	5.43%	47	5.75%	35	3.97%
Pricing Very Positive	115	3.71%	45	5.51%	21	2.38%
Total (unique)	356	11.5%	110	13.46%	82	9.3%

According to the objective of this study, the aim was to determine which factors influence foreign customers' satisfaction with Thailand's dental tourism. In order to achieve this, the crosstab function was used. The crosstab is a table showing the relationship between two or more variables. It can display the same data, but groups the data both horizontally and vertically so that the datasheet can be more compact and easier to read. In this example, the crosstab function was used to simultaneously analyze three variables (country, sentiment, and factors). The country was selected as an attribute (Figure 9) while sentiment and each factor were selected as nodes ("Selected Items") (Figure 10). Finally, the crosstab results are presented as tabulated data showing the list of sentiment as rows, countries as columns, and a count of the number of reviews and opinions as a readout. Either reviews or opinions can be presented as a count of number or percentage that corresponds to each classified sentiment (Figure 11).

An example of the crosstab results is presented in Table 2. The number of comments related to pricing in the sentiment analysis was counted and calculated into percentages. For example, of the 3,096 comments from Australia, 48 or 1.55% had a very negative opinion about pricing. The total (unique) indicates the total number of comments about pricing removing any redundant comments made by a reviewer. For example, a total of 425 comments were made about pricing by the reviewers from Australia, but the reviewers made 69 redundant comments.

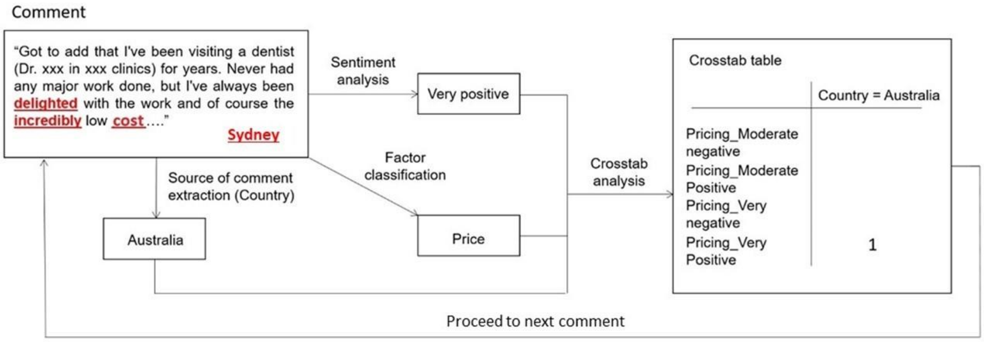

Figure 12 Diagram Showing Analytical Framework

The analytical framework is summarized in Figure 12. The schematic demonstrates how the extracted data were managed and analyzed. First, an attribute of interest (countries) was selected. Next, an automated sentiment analysis was performed to create nodes of positive and negative keywords classifying them into four levels. For example, the sample comment contains "delighted" and "incredibly" and was classified as a very positive node. Separately, keywords were identified and matched with appropriate factors of interest previously defined from the frequency word count analysis. In the same example (Figure 9), the comment contains "cost" and was matched into the factor "price". The process is reiterated for each comment in the database. Finally, a crosstab function was used to compile the analysis together for the final readout.

4. Process validation

The results were reviewed in order to validate the analysis methodology. The comments, reviews, and opinions from Australia were chosen. Based on manually reading the comments and replies from Australian reviewers, four types of errors were discovered which are false positive, false negative, irrelevant, and redundant.

4.1 False positive

False positive refers to an error where a negative result is interpreted as positive. In this paper, for example, a false positive happened when a comment mentioned a good experience with dental treatment in the past but not in Thailand or when a reviewer commented with sarcasm (Table 3, Comment number 50).

4.2 False negative

False negative refers to an error where a positive result is interpreted as negative. In this paper, for example, a false negative occurred when a reviewer had a bad experience with dental treatment in the past but not in Thailand (Table 3, Comment number 10).

4.3 Irrelevant

The comment is not related to the main topic. In this study, a reviewer may review or comment on something else which is not related to dental tourism (Table 3, Comment number 135).

4.4 Redundant

This is the same comment or reply posted several times. This may happen because of an error in data collection or because reviewers sent a comment or replied to posts twice (Table 3, Comment number 100&101).

Table 3 shows summaries and examples of each type of error.

Table 3 Example of each type of error

#	*Review and opinion*	*Sentiment classification*	*Type of error*	*Note*
50	"…Dentists who care little or who may not really be aware of infection control will be only too happy to take your tourist dollars. Remember, too, the dentists may be trained overseas but where are their staff trained...& it's the staff that take care of the instruments. **Be aware...don't bring home a beautiful new smile with a deadly souvenir attached**…."	Positive	False Positive	"Be aware...don't bring home a beautiful new smile with a deadly souvenir attached." Most of words are positive such as beautiful, smile, souvenir so software did not know the real meaning of this sentence so the automatic sentiment analyzed wrong.
10	"…I am more inclined to give people **warnings about Aussie dentists**! I had major mouth restorative work done over 3 weeks in June and 3 weeks in November 2007 with Bangkok Smile on Asoke Road. Would not hesitate to recommend them to everyone - they were wonderful. One thing your boyfriend might not be aware of - depending on how much work he is having done - his medical expenses in Bangkok are claimable on his tax return in Australia if his total expenses in the fin/yr are more than $1500. So make sure he keeps all his receipts	Positive	False Negative	This comment shows a bad experience in Australia not in Thailand.

#	*Review and opinion*	*Sentiment classification*	*Type of error*	*Note*
	and his bank statements that will have the conversion rate, etc. My dental advice is limited to restorative work like implants, etc., so if that's what he is having done, feel free to send me a private message here. If it's things like crowns, etc., then there are quite a few people here who have had stuff like that done over there who can help you out.….”			
135	"…I don't have an issue with **weight loss surgery,** but I have a very close friend who had had it and I have to agree, it's not something you have done and that's it. Getting a gastric band is safer as it's reversible and you can alter the amount of liquid and therefore restriction of the stomach via a port just under the skin. Liquid can be taken in and out with a small needle very easily. She has to see her specialist regularly; it's not a once off thing. Gastric stapling might involve less after care, but it's permanent and I believe they don't recommend it anymore in Australia….”	Positive	Irrelevance	This reviewer discussed about weight loss surgery which is not relevant to our topic.

#	*Review and opinion*	*Sentiment classification*	*Type of error*	*Note*
100 & 101	"Hi all, I know this topic may have been posted quite a bit in the past but I am interested to know anyone who has recently got dental treatment in Thailand specifically the XXX Dental Center and can recommend the dentist that did their treatment. I will be requiring about 8 crowns and it is way too expensive to have done locally here in Australia so am seriously considering getting my dental treatment done overseas. If anyone has recently visited the BIDC and can share their story of any of the dentists and provide the names, it will be greatly appreciated. Thank you all Perth Boy"	Positive	Redundancy	Reference 100 and 101 provide exact information.

'All Australians' comments which constitute the largest number of opinions were analyzed to get a more quantitative picture of methodological validity. Three factors (nodes) with the largest number of comments (quality, country, and people) were chosen for evaluation. The selected dataset was read and manually assigned and the types of error were tallied when identified. The error percentage was calculated by dividing the number of errors in each type by a total number of datasets (Table 4).

Table 4 Percentage of error in each category.

Factors	**n**	**False Positive**		**False Negative**		**Irrelevance**		**Redundancy**	
		n	%	n	%	n	%	n	%
Quality	736	24	3.26%	93	12.64%	10	1.36%	34	4.62%
Country	512	12	2.34%	41	8.01%	8	1.56%	27	5.27%
People	509	11	2.16%	47	9.23%	3	0.59%	19	3.73%
Average	1,757	47	2.67%	181	10.30%	21	1.20%	80	4.55%

Discussion & Conclusion

A methodology is described that applies computational methods to mine and analyze vast databases of consumer-generated content from social networks or review websites for qualitative research. The process started with sample framing where selected keywords were browsed via search tabs that are a common feature on many social network websites. Next, a programming code was written to scrape relevant data (e.g., username, location, reviews, and opinion) from the framed sample. The scraped data were then cleaned up and managed into an appropriate table format. Finally, the organized data were exported to NVivo to perform qualitative analysis encompassing sentiment analyses in multiple categories of interest.

The use of computational tools enables extraction, management, and analysis of large databases of social media content in a short period of time at the scale which could not be manually performed within a reasonable time frame. While an initial learning curve is relatively high to write programming code for data extraction, the extent at which the code could extract data is far greater than what could be achieved by human input, and the written code could be readily modified for extraction of different datasets of interest, saving time in the long run.

While the analysis of large social media databases is traditionally limited to those of numerical formats (e.g., review scores and number of mentions), an incorporation of NVivo software into the framework allows qualitative analysis of social media contents which are mostly in text-based formats (e.g., posts and comments). As a specific example, customers' review posts from TripAdvisor were extracted and processed to analyze international customers' perception of dental tourism in Thailand. Sentiment analysis via NVivo allows quantification on relative percentage of positive versus negative comments according to the defined categories of interest (e.g., service, price, and place). Sentiment analysis may be more susceptible to error than other types of qualitative research that use different types of data. For example, if big data are compared with qualitative approaches and traditional methodologies, such as interviews, traditional one is mostly based on interview, the interpretation is for a limited sample size. On the other hand, larger information from big data is good for the results can be generalized, but the software also generates more mistakes that scale with data. This is some compromise that should be considered.

Using the example mentioned above, the accuracy of the analytical framework was evaluated by manually analyzing a subset of the extracted database focusing on the comments from Australian customers which comprise the majority of the total comments (42%). Previous studies show the accuracy of automated textual analysis approximately 70-80% (García-Pablos et al., 2016; Kirilenko et al., 2018; Schmunk et al., 2013). However, few studies mention an error of around 20-30%. Four types of error in the analysis were identified in this investigation, namely, (1) False positive, (2) False negative, (3) Irrelevant, and (4) Redundant. False negative was the most frequently encountered error comprising about 10% of total comments verified. Overall, this method has an accuracy of about 80-85%. Furthermore, the computational methodology of this study can analyze the whole database which might allow to better find valuable insights that could otherwise have been missed or inaccurately captured when researchers randomly choose a selected pool of comments to analyze manually. Therefore, the proposed methodology of this research serves as alternative framework to enable qualitative analysis of social media content at scale.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

I gratefully acknowledge the funding received towards my doctoral degree from Royal Golden Jubilee Ph.D. (RGJPHD)Thailand.

References

Batrinca, B., & Treleaven, P. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai Society, 30*(1), 89-116.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review, 65*(3), 122-135.

De Mauro, A., Greco, M., & Grimaldi., M. (2016). A formal definition of Big Data based on its essential features. *Library Review,, 65*(3), 122-135.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., . . . Starič, A. (2012). Orange: data mining toolbox in Python. *The Journal of Machine Learning Research, 14*(1), 2349-2353.

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations–A case from Sweden. *Journal of Destination Marketing Management, 3*(4), 198-209.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management, 35*(2), 137-144.

García-Pablos, A., Cuadros, M., & Linaza, M. T. (2016). Automatic analysis of textual hotel reviews. *Information Technology & Tourism, 16*(1), 45-69.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561): John Wiley & Sons.

Hilal, A. H., & Alabri, S. S. (2013). Using NVivo for data analysis in qualitative research. *International interdisciplinary journal of education, 2*(2), 181-186.

Holsti, O. R. (1969). Content analysis for the social sciences and humanities. *Reading. MA: Addison-Wesley*.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons, 53*(1), 59-68.

Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. *Journal of Travel Research, 57*(8), 1012-1025.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note, 6*(70), 1.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., . . . Gutmann, M. (2009). Computational social science. *Science, 323*(5915), 721-723.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting Electronic Media, 57*(1), 34-52.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting Electronic Media, 57*(1), 20-33.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities, 2*, 460-475.

Marine-Roig, E., & Clavé, S. A. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. *Journal of Destination Marketing Management, 4*(3), 162-172.

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2017). A big data analytics method for tourist behaviour analysis. *Information Management, 54*(6), 771-785.

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? *Qualitative Research, 18*(6), 591-603.

Obar, J. A., & Wildman, S. S. (2015). Social media definition the governance challenge: An introduction to the special issue. *Telecommunications policy, 39*(9), 745-750.

Raun, J., Ahas, R., & Tiru, M. (2016). Measuring tourism destinations using mobile tracking data. *Tourism Management, 57*, 202-212.

Riff, D., Lacy, S., Fico, F., & Watson, B. (2019). *Analyzing media messages: Using quantitative content analysis in research*: Routledge.

Schmunk, S., Höpken, W., Fuchs, M., & Lexhagen, M. (2013). Sentiment analysis: Extracting decision-relevant knowledge from UGC. In *Information and Communication Technologies in Tourism 2014* (pp. 253-265): Springer.

Schroeder, R. (2014). Big Data and the brave new world of social media research. *Big Data Society, 1*(2), 2053951714563194.

Simon, A. F. (2001). A unified method for analyzing media framing. *Communication in US elections: New agendas*, 75-89.

Smedt, T. D., & Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research, 13*(Jun), 2063-2067.

Sutherland, A. (2017). The cost of medical treatment in Thailand. Retrieved from https://thaiexpathealth.com/the-cost-of-medical-treatment-in-thailand/

Zamawe, F. C. (2015). The implication of using NVivo software in qualitative data analysis: Evidence-based reflections. *Malawi Medical Journal, 27*(1), 13-15.