Classification of Healthcare Service Reviews with Sentiment Analysis to Refine User Satisfaction

Review Paper

Khai Herng Leong
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
khaiherngleong98@gmail.com

Dahlila Putri Dahnil
Centre for Software Technology and Management (SOFTAM),
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia
dahlilaputri@ukm.edu.my

Abstract *– In natural language processing, sentiment analysis determines the polarity of a message based on lexical emotion. This technique is utilized intensively in service sectors to study the level of consumer satisfaction. However, the healthcare service field lacks such practice to detail responses in existing feedback systems. A proposed application which implements sentiment analysis is developed for improvement. User reviews are classified according to their word influences, namely positive, negative and neutral states. In addition, topic modelling is included to organize them in several service themes. A graphical user interface, GUI which records the analytical results is presented to users for interaction. This approach does not only benefit patients to choose their desired medical centres, but also healthcare management who wish to enhance their service quality.*

Keywords*: Healthcare Service Review System, Natural Language Processing, Sentiment Analysis, Topic Modelling, Web Scraping*

1. INTRODUCTION

Healthcare preserves human health through prevention, detection, as well as treatment of disabilities, illness, injuries and mental. Medical tourism to Malaysia is becoming popular because of high quality healthcare service with low cost compared to other countries in Asia region [1].

We often hear about healthcare service issues of certain medical centres in our community. Most of the customers express their opinion about the medical services they received to closest people, while some may give feedback via online to share their experiences. These responses are very useful for patients to choose a medical centre based on the testimonies. Customers experiences and testimonies are able to help other patients in the selection of medical centres. Patients can compare the satisfaction details in choosing which medical centre they would like to seek treatment from. Furthermore, medical centre management can refer to this application for service improvement. Thus, a comprehensive healthcare service feedback system is required to classify the reviews.

A sentiment analysis algorithm is developed to provide classification to the user reviews. This technique emphasizes word emotion to determine the polarity of a sentence. Scattered responses can be arranged neatly based on evaluation parameters. All functionalities are implemented to facilitate the process of referral, comparison and selection of medical centres.

2. LITERATURE REVIEW

Internet feedback system collects ratings and reviews from users as references for the community. The former acts as a Likert scale which comprises five or ten points. The higher the rating, the better the reputation. Next, the latter states the aspects of service quality, such that we can comprehend its general description. The existing systems which support healthcare service are Google review, Lyfboat and Yelp.

Google review is a popular feedback system based on world location, where users can rate and leave responses for any services. The application extracts several keywords according to mentioned frequency to present information more effectively. All reviews can be arranged in four order types, namely most relevant, newest, highest and lowest rating. A language translation tool, Google Translate is utilized to interpret foreign comments. Fig. 1 shows the interface of the Google review.

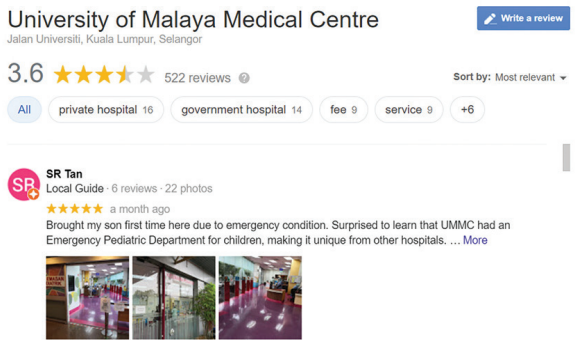

Fig. 1. The Interface of Google Review

Lyfboat is an international healthcare website which provides search and query functionalities for treatment procedures, doctor and hospital information. The latter has neat details including centre excellence, medical infrastructure and transportation facility. Nonetheless, at most 12 well known Malaysian hospitals are recorded in the database. Besides, star rating is the only feedback channel. The interface is shown in Fig. 2.

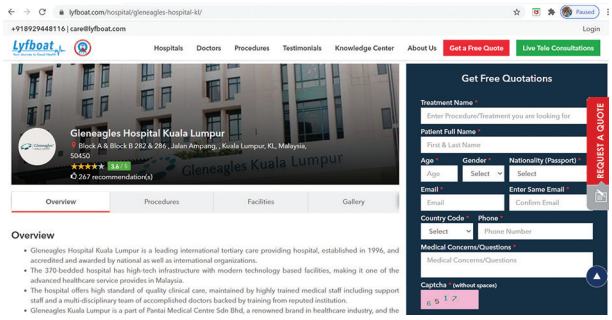

Fig. 2. Lyfboat interface

Yelp is a mobile application and business service website which covers Malaysia and 31 other countries. It emphasizes comprehensive search and comparison between premises in terms of distance, evaluation, price and others. Medical centres are included to provide healthcare service references. However, there are less reviews available because it is not a common practice among Malaysians. The interface of the Yelp website is shown in Fig. 3.

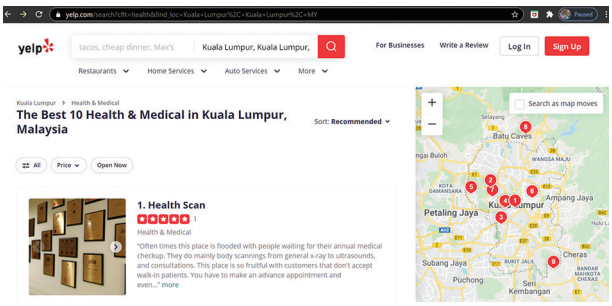

Fig. 3. Yelp interface

Sentiment analysis is used for biometrics, computational linguistics, natural language processing, NLP, as well as text analysis in distinguishing, extracting, quantifying and studying subjective information. It is divided to two approaches, namely lexicon and machine learning. The latter consists of classification algorithm which trains dataset samples to predict the polarity of other documents [2]. Linear regression, Naive Bayes and support vector machine are some of the models which are often applied in this field. On the other hand, lexicon method estimates polarity scores based on word matches and relevant sentimental lexicons [3]. The score ranges from -1 to 1, where below shows how to label it:

- Equal or more than 0.05 is positive state.
- Equal or less than -0.05 is negative state.
- Between -0.05 and 0.05 is neutral state.

Hence, sentiment analysis identifies whether a sentence is positive, negative or neutral state. The accuracy of this system depends on the efficiency of human judgement. An algorithm which achieves a 70% validity level is considered to have good machine intelligence as we agree about any events in an average of 80% only.

In terms of these approaches, lexicon method is more superior than statistically trained classifiers [8]. Lexicon extension with linguistic information improves system durability. The correlation between frequency of keyword and overall rating of text is stated clearly, which guarantees the quality in generating lexicons.

3. THE PROPOSED SYSTEM

Five hospitals and clinics selected in the Google review feedback obtained from the website for each hospital [9-13] The processing from the feedback to the GUI is presented in Fig. 4. Initially, the data is scraped and stored in an integrated development environment, IDE. Topic modelling was conducted in each response to identify service areas in the information classification. Next, sentiment analysis was implemented to obtain polarity scores based on the probability of the appearance of character emoticons [4]. This index is capable of calculating the cumulative average of the ratings and determining the nature of the polarity to classify the responses.

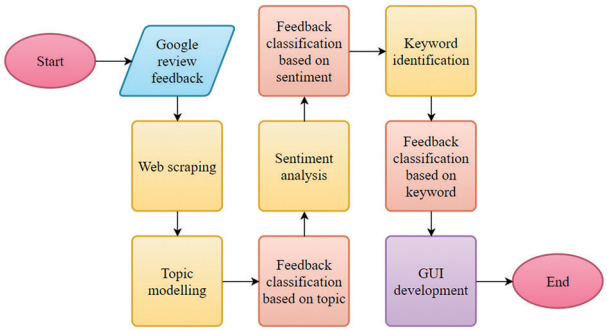

Fig. 4. Architectural design of algorithm

Frequently used words are also identified to indicate reviews that have these words. Finally, a GUI is built to deliver all processed health service feedback information to users. The approach and text examples are shown in Fig. 5.

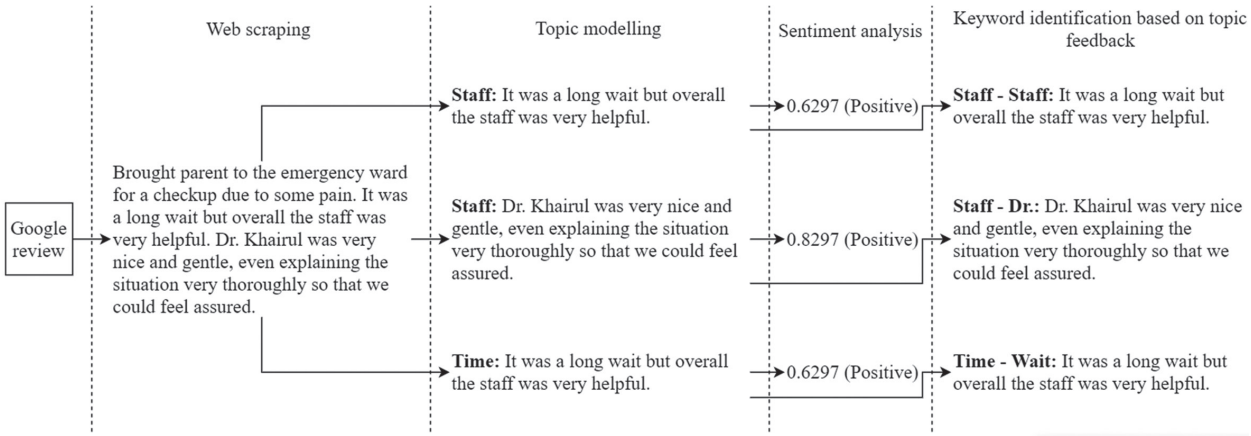

Fig. 5. Algorithm and text processing steps

3.1. WEB SCRAPING

A list of feedback, star ratings and premise information in a Google review is required to run this project. In addition to copy and paste, web scraping is an alternative and efficient method of collecting duplicate data. The mechanism of the technique is to crawl hypertext markup language, HTML and extract the desired information through programming. The HTML parser activates the application programming interface, API to access website content [6]. The Beautiful Soup and Selenium WebDriver modules were used throughout this procedure. The flow of data from a Google review to be recorded in a spreadsheet and text file is presented in Fig. 6.

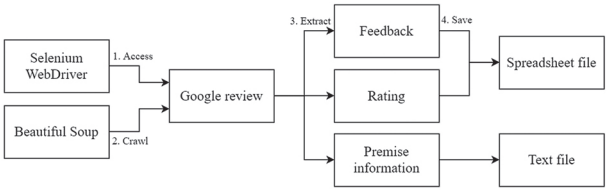

Fig. 6. Process of web scraping

Web scraping involves four steps, namely access, crawl, extract and save, with the following essence:

1. **Access**: WebDriver is programmed to emulate users browsing Google reviews in reading medical centre information.
2. **Crawl**: Beautiful Soup crawls Google review HTML to find the data it needs.
3. **Extract**: Feedback lists, star ratings and premise information were extracted.
4. **Save**: This information is stored in a spreadsheet and text file for use in the next stage.

The five hospitals and clinics surveyed are as follows:

- Subang Jaya Medical Centre, SJMC.
- Ara Damansara Medical Centre, ADMC.
- Tung Shin Hospital.
- Universiti Kebangsaan Malaysia Medical Centre, UKMMC.
- The KL Sky Clinic.

3.2. TOPIC MODELLING

Topic modelling identifies latent themes that potentially describe a piece of text [7]. Latent Dirichlet allocation, LDA are among the popular unsupervised machine learning algorithms in this approach. The technique detracts from previous Dirichlets for distributing topics and words, as well as avoiding overfitting effects [5]. This model assumes all documents are generated through a statistical generation process, meaning they contain a number of speculative titles from a list of related keywords. The flow of topic modeling in formulating the feedback theme, in which Subang Jaya Medical Centre (SJMC) serves as a modeling example is presented in Fig. 7.

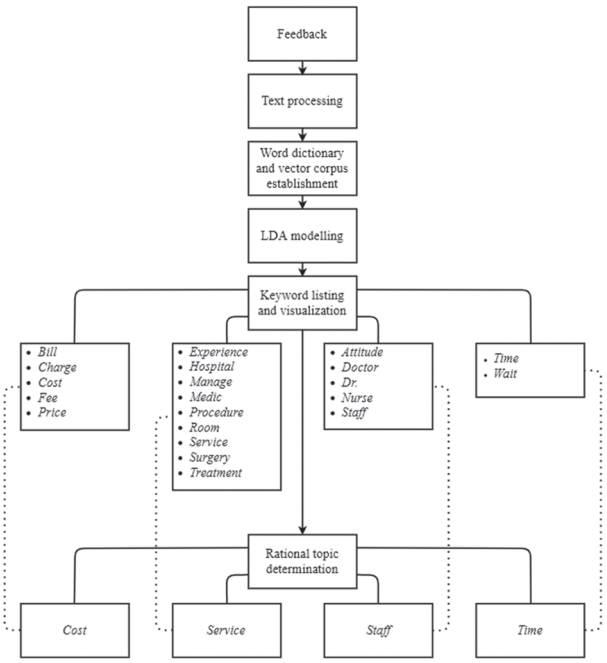

Fig. 7. Process of topic modelling with SJMC as example

This procedure was also carried out in response to ADMC, Tung Shin Hospital, Universiti Kebangsaan Malaysia Medical Centre (UKMMC) and The KL Sky Clinic.

First, text processing is implemented to restore word structure in constructing vector dictionaries and corpora. Both of these matrices were modelled with the LDA algorithm to generate interactive lists and graphs of keywords representing various types of themes.

The role of the developer is to identify as many as four service topics from a particular word that have a semantic relationship based on the above framework. For example, 'bill', 'charge', 'cost', 'fee' and 'price' can interpret 'Cost'. The list of titles in each medical centre is as follows:

- **SJMC**: 'Cost', 'Service', 'Staff' and 'Time'.
- **ADMC**: 'Park', 'Service', 'Staff' and 'Time'.
- **Tung Shin Hospital**: 'Cost', 'Service', 'Staff' and 'Time'.
- **UKMMC**: 'Park', 'Service', 'Staff' and 'Time'.
- **The KL Sky Clinic**: 'Service' and 'Staff'.

Each feedback may contain at least one theme that can be classified. The method of classification according to keywords and topics is presented in Fig. 8.

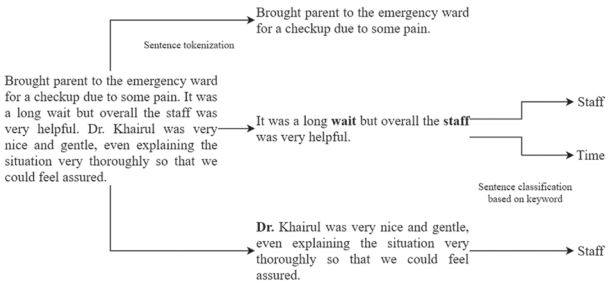

Fig. 8. Feedback classification based on keyword and topic

Response tokenization is performed to isolate verses. If it has a themed word, the sentence is categorized in a related topic, as in the illustration above. 'Dr.' and 'staff' represent 'Staff', while 'wait' is 'Time'. The process was also conducted on other reviews in five medical centres.

3.3. SENTIMENT ANALYSIS

Sentiment analysis is the emotional research of texts in positive, negative and neutral, where this status is known as the nature of polarity. Essentially, the three traits signify cheerful, hateful and moderate feelings respectively.

In this project, the Valence Aware Dictionary for Sentiment Reasoning model, VADER is implemented to determine sentence sentiment and calculate the average polarity of service topics through scoring. The score range is from -1 to 1, where equal to or greater than 0.05 signifies positive, equal to or minus -0.05 signifies negative, and between -0.05 and 0.05 signifies neutral. Table 1 shows the nature of topic polarity in each medical centre.

According to the table, the reputations of ADMC and The KL Sky Clinic are satisfactory, while SJMC, Tung Shin Hospital and UKMMC are modest.

Table 1. Polarity of service topics

Medical centre	Service topic	Average mark	Polarity	Rating
SJMC	Cost	-0.03	Neutral	3
	Service	0.03	Neutral	3
	Staff	0.11	Positive	3
	Time	-0.03	Neutral	3
ADMC	Park	0.39	Positive	4
	Service	0.15	Positive	3
	Staff	0.21	Positive	4
	Time	0.03	Neutral	3
Tung Shin Hospital	Cost	0	Neutral	3
	Service	0	Neutral	3
	Staff	-0.04	Neutral	3
	Time	-0.11	Negative	3
UKMMC	Park	-0.04	Neutral	3
	Service	-0.03	Neutral	3
	Staff	0.01	Neutral	3
	Time	-0.05	Negative	3
The KL Sky Clinic	Service	0.44	Positive	4
	Staff	0.46	Positive	4

3.4. KEYWORD IDENTIFICATION

Before calculating word frequency, text processing should be performed to restore formatting and get rid of less meaningful keywords. Examples of sentences used in the demonstration are as follows:

The 3 nurses are good,
can better!

A description of this procedure is attached with the text output in Table 2.

Table 2. Text processing

Step	Process	Output
1	Replace line breaks with spaces.	The 3 nurses are good, can better!
2	Remove tabs.	The 3 nurses are good, can better!
3	Replace '&amp' with '&'.	The 3 nurses are good, can better!
4	Remove '(Translated by Google)', if any.	The 3 nurses are good, can better!
5	Remove the original review not in English, if any.	The 3 nurses are good, can better!
6	Remove accented letters.	The 3 nurses are good, can better!
7	Remove digits.	The nurses are good, can better!
8	Remove punctuation.	The nurses are good can better
9	Convert uppercase to lowercase.	the nurses are good can better
10	Remove stopwords.	nurses good better
11	Perform word lemmatization.	nurse good good

12	Perform word stemming.	nurs good good
13	Perform word tokenization.	nurs, good, good
14	Calculate the cumulative frequency of words.	nurs (1), good (2)

A word cloud consist of a collection of keywords and the size of each keyword determines their dominance in a content. Fig. 9 shows the collection of keywords and the significant keyword is determined based on its size.

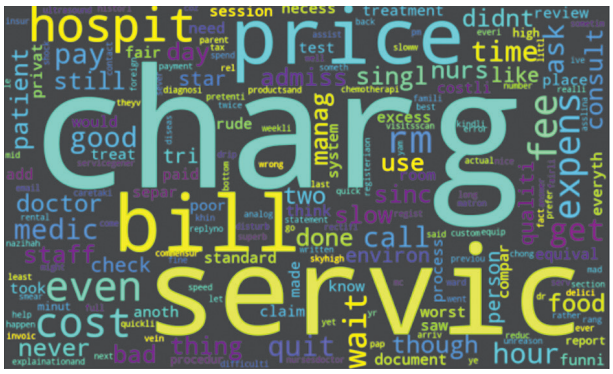

Fig. 9. Word cloud

The cumulative number of words was calculated to determine the five key keywords that were frequently specified in the response to each service topic. All reviews that have the word are organized in their respective groups. Table 3 presents this list of contents.

Table 3. Most mentioned keywords of service topics

Medical centre	Service topic	Keyword
SJMC	Cost	Charge
		Price
		Service
		Bill
		Hospital
	Service	Medic
		Service
		Hospital
		Doctor
		Experience
	Staff	Staff
		Dr.
		Nurse
		Doctor
		Hospital
	Time	Wait
		Time
		Hospital
		Doctor
		Hour
ADMC	Park	Park
		Easy
		Spacious
		Clean
		Hospital
	Service	Service
		Medic
		Hospital
		Doctor
		Room
	Staff	Nurse
		Friendly
		Doctor
		Dr.
		Staff
	Time	Time
		Wait
		Hour
		Doctor
		Appoint
Tung Shin Hospital	Cost	Service
		Price
		Hospital
		Doctor
		Ask
	Service	Good
		Service
		Hospital
		Medic
		Doctor
	Staff	Doctor
		Nurse
		Dr.
		Staff
		Hospital
	Time	Ask
		Wait
		Time
		Doctor
		Hour
UKMMC	Park	Park
		Hospital
		Fee
		Rate
		Expensive
	Service	Hospital
		Service
		Patient
		Doctor
		Good

UKMMC	Staff	Nurse
		Staff
		Patient
		Doctor
		Time
	Time	Time
		Wait
		Doctor
		Patient
		Hospital
The KL Sky Clinic	Service	Treatment
		Clinic
		Service
		Medic
		Dr.
	Staff	Dr.
		Doctor
		Roland
		Treatment
		Friendly

4. MODEL VALIDATION

In terms of sentiment analysis, the motive of this section is to investigate how effective the VADER model is in classifying the nature of polarity in line with human feelings. The sentiment evaluation process of the technique as well as Google review users is presented in Fig. 10 and Fig. 11.

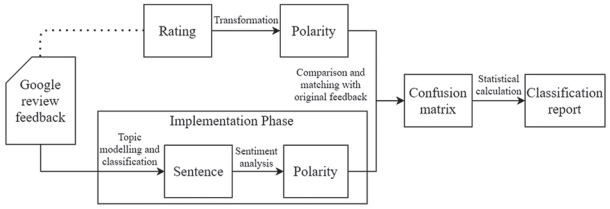

Fig. 10. Testing process of sentiment analysis

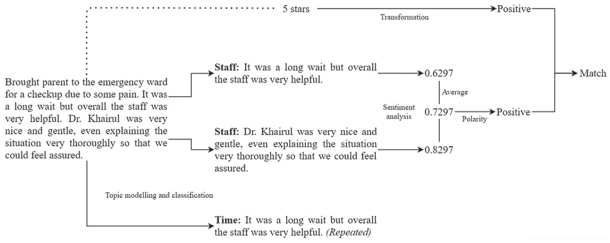

Fig. 11. Comparison and matching process of polarity

Google review star feedback and ratings are provided by users with a good visit experience. Star points range from 1 to 5, where data transformations are also carried out to categorize them in positive, negative and neutral properties as comparative benchmarks. The details of the classification are as follows:

- 1 and 2 stars are made as negative state.
- 3 stars is made as neutral state.
- 4 and 5 stars are made as positive state.

Each response may contain themed sentences of several topics that can be identified through topic analysis. In the implementation phase, they were extracted and classified in related topics, then conducted sentiment analysis to obtain polarity scores. Further, the mean score was calculated to determine the nature of the overall polarity which also consisted of positive, negative and neutral. It was observed that sentences repeated in other topics were ignored to avoid biased decisions.

In the example above, the Google review feedback is rated 5 stars, which is a positive attribute. After conducting topic analysis and classification, the sentences were divided into the themes which are 'Staff' and 'Time' respectively. The result of the process is presented in a standalone system with a GUI to enable users to view the feedback shown in Figure 12. Sentiment analysis was performed to obtain their polarity scores, namely 0.6297 and 0.8297. Average scores were also calculated, where repetitive sentences were excluded. The result is 0.7297 which indicates a positive status, which is similar to the original sentiment.

The purpose of this process is to compare machine sentiment with that of humans to investigate the level of accuracy. Matches between these two data were recorded in a confusion matrix, as in Table 4.

684 feedback fractures were compared. Of the 227 negative traits, 116 neutrals and 341 positive of origin, VADER attempted to determine 204, 17 and 249 respectively. In addition, of the 338 negative traits, 51 were neutral and 295 positive predictions, 204, 17 and 249 were corresponding to the original sentiments, respectively. Indeed, a classification report can be tabulated with this data.

The report discusses four criteria that describe the classification statistics, namely accuracy, retrieval, F-score and accuracy. Table 5 presents the calculation results according to the confusion matrix.

Table 5. Classification report

	Precision	Recall	F-score
Negative	0.9	0.6	0.72
Neutral	0.15	0.33	0.2
Positive	0.73	0.84	0.78
Accuracy		0.69	

Precision checks the efficiency of the classifier in predicting the original valuation. Recall emphasizes the accuracy of the classifier in predicting the original valuation. Next, F-score tests the accuracy and recall performance. Finally, accuracy determines the overall percentage of sentiment that is correctly predicted by the comparison volume.

Table 4. Confusion matrix of polarity

		Original			Total predicted sentiment
		Negative	Neutral	Positive	
Prediction	Negative	204	63	71	338
	Neutral	13	17	21	51
	Positive	10	36	249	295
Total original sentiment		227	116	341	684

A level of accuracy that reaches 69% means that the VADER model is good at identifying user feedback feelings. In terms of the F-score, the neutral status underperformed because the response may contain a mixture of positive and negative sentences with extreme polarity. In contrast, the positive and negative traits showed very good results throughout the classification procedure. Negative sentiments with the highest percentage of precision indicated that the model was able to label nine out of ten reviews as bad, while positive sentiments with the highest recall values referred to the algorithm performing in determining good responses with minimum error rates.

5. CONCLUSION

Health services are a primary need because of their role in healing and saving human beings. Therefore, the quality of services must be taken care of to guarantee the universal interest. A comprehensive feedback system has been built to try to sustain this mission.

Compared to existing systems, this new application recognizes sentiment analysis techniques that classify feedback into three polarities, namely positive, negative and neutral. In addition, topic analysis and key word determination were also implemented to further detail user feelings. Such advantages make it easy for the patient to peruse the details of the review and select the desired medical centre. Healthcare management can also refer to this system to improve the quality of their services.

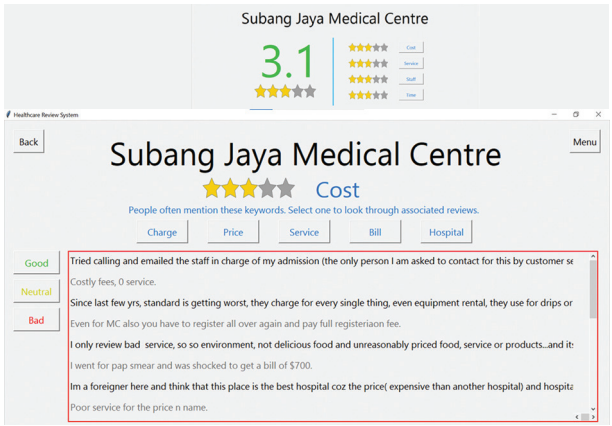

Fig. 12. The GUI to show the result after topic analysis and classification.

6. ACKNOWLEDGMENT

The research is funded by Universiti Kebangsaan Malaysia (UKM) research project DCP-2018-001/2 under Research Centre for Software Technology and Management (SOFTAM), www.ftsm.ukm.my/softam, Faculty of Information Science and Technology.

7. REFERENCE

[1] International Living, Healthcare in Malaysia. https://internationalliving.com/countries/malaysia/healthcare-in-malaysia (accessed: 2021)

[2] M. Birjali, A. Beni-Hssane, M. Erritali, "Machine Learning and Semantic Sentiment Analysis Based Algorithms for Suicide Sentiment Prediction in Social Networks", Procedia Computer Science, Vol. 113, 2017, pp. 65-72.

[3] O. Araque, G. Zhu, C. A. Iglesias, "A Semantic Similarity-based Perspective of Affect Lexicons for Sentiment Analysis", Knowledge-Based Systems, Vol. 165, 2019, pp. 346-359.

[4] K. Utsu, J. Saito, O. Uchida, "Sentiment Polarity Estimation of Emoticons by Polarity Scoring of Character Components", Proceedings of the IEEE Region Ten Symposium, Sydney, NSW, Australia, 4-6 July 2018.

[5] R. Annisa, I. Surjandari, Zulkarnain, "Opinion Mining on Mandalika Hotel Reviews Using Latent Dirichlet Allocation", Proceedia Computer Science, Vol. 161, 2019, pp. 739-746.

[6] V. Singrodia, A. Mitra, S. Paul, "A Review on Web Scrapping and Its Applications", 2019 Proceedings of the International Conference on Computer Communication and Informatics, Coimbatore, India, 23-25 January 2019.

[7] S. Kim, H. Park, J. Lee, "Word2vec-based Latent Semantic Analysis (W2V-LSA) for Topic Modeling: A Study on Blockchain Technology Trend Analysis", Expert Systems With Applications, Vol. 152, 2020, p. 113401.

[8] D. Grabner, M. Zanker, G. Fliedl, M. Fuchs, "Classification of Customer Reviews Based on Sentiment Analysis", Proceedings of the 19th Conference on Information and Communication Technologies in Tourism, Helsingborg, Sweden, 25-27 January 2012, pp. 460-470.

[9] Pusat Perubatan Subang Jaya, Subang Jaya Medical Centre, https://www.google.com/maps/place/Subang+Jaya+Medical+Centre+(SJMC)/@3.0765703,101.5909797,15.92z/data=!4m7!3m6!1s0x31cc4c60badceb4f:0xa2a80452021765a6!8m2!3d3.079771!4d101.5938418!9m1!1b1

[10] Pusat Perubatan Ara Damansara, Ara Damansara Medical Centre, https://www.google.com/maps/place/Ara+Damansara+Medical+Centre/@3.1151723,101.5627106,17z/data=!4m7!3m6!1s0x31cc4e7f4b90f90d:0x34c9559e5d246762!8m2!3d3.1151669!4d101.5648993!9m1!1b1 (accessed: 2021)

[11] Hospital Tung Shin, https://www.google.com/maps/place/Tung+Shin+Hospital/@3.1459617,101.7016377,17z/data=!4m11!1m2!2m1!1stung+shin!3m7!1s0x31cc49d42ad29b1d:0xcbbb8b9f09c732e6!8m2!3d3.1464757!4d101.7040118!9m1!1b1!15sCgl0dW5nIHNoaW5aCyIJdHVuZyBzaGlukgEIaG9zcGl0YWywAQA (accessed: 2021)

[12] Pusat Perubatan Universiti Kebangsaan, Malaysia, https://www.google.com/maps/place/Pusat+Perubatan+Universiti+Kebangsaan+Malaysia/@3.0992724,101.7232131,17z/data=!4m7!3m6!1s0x31cc35e90db3a4ad:0xaf6721771e83594a!8m2!3d3.099267!4d101.7254018!9m1!1b1 (accessed: 2021)

[13] The KL Sky Clinic, https://www.google.com/maps/place/The+KL+Sky+Clinic/@3.1536606,101.7079711,17z/data=!4m7!3m6!1s0x31cc37d4514100eb:0xb390c52a4cc315e3!8m2!3d3.1536552!4d101.7101598!9m1!1b1 (accessed: 2021)