A State College Customer Feedback Data Analysis using Machine Learning-Based Algorithm

Lorna T. Soriano
Bicol State College of Applied Sciences and Technology, Philippines
ltsoriano@biscast.edu.ph

Asia Pacific Journal of Multidisciplinary Research
Vol. 7 No.4, 59-64
November 2019 Part IV
P-ISSN 2350-7756
E-ISSN 2350-8442
www.apjmr.com
CHED Recognized Journal
ASEAN Citation Index

Date Received: October 5, 2019; Date Revised: November 26, 2019

Abstract *– Obtaining customers' view through text-based feedback in a survey is considered an important process for organizations including education sector since it provides an overview of the different relevant aspects which aid administrators in planning, policy making and decision making. Over the years, academic institutions have collected vast amounts of textual data through survey. However, analyzing voluminous amounts of unstructured customer feedbacks to gain a general view of their concerns and sentiments remains a challenge for the institution.*

This study conducted a text analysis of the feedbacks from the customer satisfaction survey of one of the State Colleges in Region V, Philippines for academic year 2018-2019. A machine learning-based algorithm such as topic modeling using Latent Dirichlet Allocation (LDA) was employed in the study for the automatic summarization of text and extraction of topics from these unstructured data. Moreover, it describes the text mining process performed to retrieve useful information from the huge amount of text-based data. The topmost concerns extracted from the customer feedbacks were then identified. In the result, specific concerns for offices were revealed such as staffing, environment, customer feedback system, and IT system. Furthermore, issues on security personnel and student assistants' attitude as well as library operation and management are notably highlighted in the feedbacks.

Keywords *– Latent Dirichlet Allocation (LDA), machine learning-based algorithm, topic modeling*

INTRODUCTION

Government institutions like State Universities and Colleges (SUCs) in particular, are employing customer satisfaction survey to gather information on clients' needs, issues, experiences, and expectations purposively to figure out potential problems and challenges. Usually, the survey questionnaire contains an open-ended question that solicits additional information on customer experiences. These will help on the formulation of new as well as on the improvement of existing policies and standards. Moreover, this is in line with the government's call for ensuring people-centered, clean, and efficient governance by institutionalizing response and feedback mechanisms [1].

Analyzing customer response data continues to be challenged in providing useful insights for management analysis despite of numerous attempts to do so. Data that are collected from fixed-point rating scale are known to suffer from multitude of problems such as yes-saying, no-saying and scale use measures of central tendency that challenge inference [2]. Unstructured feedbacks, on the other hand, highlight compliments as well as issues that the institution may not be aware of, thus giving them the opportunity to take appropriate action. However, academic institutions encounter difficulty in capturing the general view of their clients' sentiments from these qualitative feedbacks unlike quantitative data which are easier to interpret.

The analysis of unstructured customer feedback has drawn significant attention in the current marketing literature. Unstructured data are information that either do not have predefined data model or not organized in a predefined manner. According to [3], more than 80% of all potentially useful business information are unstructured data, in the kind of sensor readings, console logs, customer feedbacks, and the like. The study of [4] demonstrated how to organize and analyze text-based data for extracting customer insights from a huge collection of documents using text analytics tool to improve business operations and management. In the paper of [5], the authors presented a method that process and classify text files such as reviews and appraisals for opinion mining at sentence level using Natural Language Processing (NLP) and Opinion Mining algorithms. Meanwhile, the work of [6] introduced an automatic summarization technique that

discover, extract, and rank noteworthy topics from a set of online reviews. In general, word counts and frequencies are utilized as variables to identify words substantial in assessing customer behavior or in discriminating among outcomes such as satisfied versus unsatisfied experiences [2].

Alternatively, specific words in user-generated content are only indicators of latent topics which are a priori unknown [7]. These latent topics are group of words with relatively high probability of usage that can be generated using machine learning-based algorithms. Such algorithms are methods that define set of approaches to find patterns in data to be able to predict future data patterns [8] and are used in data mining. In particular, topic modeling technique, an unsupervised machine learning-based method, is one of the most popular techniques in text mining for data mining, latent data discovery, and finding relationships among text documents. It employs Latent Dirichlet Allocation (LDA) to uncover the underlying themes of text corpus and decompose its document based on those themes [9]. Topic modeling is significantly used in various research disciplines such as in software engineering, political science, medical/biomedical, computational linguistics, and geographical/locations [10].

This study aims to analyze the text-based customer feedbacks in a state college in Region V for academic year 2018-2019. This is based on the previous study that employ unsupervised learning algorithm to capture the general view of the school's clients. Specifically, it describes the text mining technique applied on the textual data obtained from the customer satisfaction surveys of the institution.

METHODS

The research design utilized in this study is descriptive since the researcher intended to analyze customer concerns obtained from their text-based feedback and suggestions from the survey form. Moreover, this research utilized text mining process. Text mining is an integral part of data mining that is aimed at automatic extraction of interesting and non-trivial patterns from the unstructured textual data [11],[12]. Figure 1 shows the text mining process flow which begins with the collection of data from various resources followed by preprocessing step, application of text mining technique on the cleaned data, analyzing the result of process, and lastly, evaluating and interpreting the result.

Text Extraction

In this study, data for school year 2018-2019 were retrieved from the database of the customer satisfaction survey system of the state college. Such survey is being accomplished by the different stakeholders of the school composed of students, employees, and external clients. Specifically, their comments, suggestions for improvements, and feedbacks were extracted from the database.

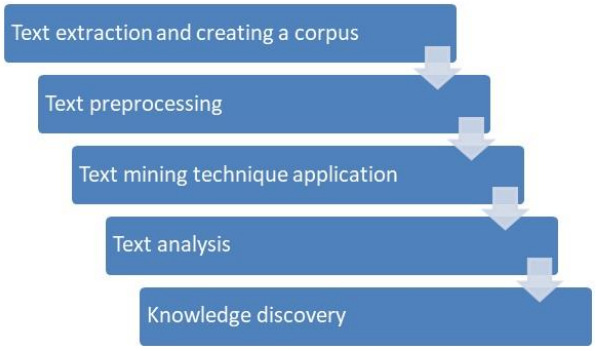

Fig. 1. Text data mining process flow.

Text Preprocessing

Preprocessing is an important and critical task during text mining process before applying any text mining technique [13]. In this phase, cleaning to minimize, if not to remove inflected words and inconsistency contained in the document was performed on the corpus. Data cleaning is absolutely vital for generating a useful topic model. Document normalization such as trimming of leading and trailing whitespaces, converting the text to lowercase, collapsing consecutive whitespaces, removing punctuations and stop words; and stemming of words were carried out.

Initially cleaning of data was done manually. Misspelled words were corrected; symbols like happy face were removed; expressions such as "ok" and "ty" were deleted; and words in local dialect were translated into standard English. Afterwards, OpenRefine, an open-source tool for data cleanup and transformation to other format, was utilized to complete the cleaning process.

Text Mining Technique

Topic modeling is the text mining technique employed in this study. Specifically, an open-source topic modeling toolkit using Latent Dirichlet Allocation (LDA) was utilized. LDA is the leading-edge unsupervised technique for extracting thematic information from a pool of documents [14]. The basic

idea of the process is that documents are produced from a mixture of latent topics, where each topic is a probability distribution over words [8]. These topic probabilities give a concise representation of a document.

Machine Learning for Language Toolkit (MALLET), a Java-based package for machine learning applications to text that uses LDA, was utilized to analyze and identify the underlying topics in the customer feedbacks.

Text Analysis

The generated topic models were presented to field experts to identify the theme for each set of words. Word Cloud application was utilized to view the words according to its weight or frequency. It is a visualization tool that uses the size of text to represent the frequency of each word in the source sample. This will help in the analysis of generated result.

RESULTS AND DISCUSSION

Data Processing Technique

The datasets were retrieved from the state college customer satisfaction survey (CSS) database for school year 2018-2019.

The dataset was primarily composed of 472 records. After extracting the data, preprocessing tasks were performed on these to remove noise and inconsistencies. These include removal of leading and trailing whitespaces, punctuations, symbols, non-important but highly repeated words and stop words; clustering terms and word stemming. This prepares the dataset for further manipulation. The corpus was reduced to 428 records after cleaning.

Topic Models

The Mallet with LDA algorithm was used to generate topics or themes that occurred in the client feedbacks.

The parameters such as the number of iterations, number of topics, and number of top words were initially set to 1000, 15, and 15, respectively. Upon reviewing the topic composition break down, it was found out that related words fall under different topics. This means that the setting is too broad and need to narrow it down by reducing the number of topics and increasing the number of top words. Hence, the LDA model was run again with decreasing topics from 10 to 7 and increasing the top words to 20. The list of generated topics is displayed in Figure 2.

On each line in the figure, the first element is the topic number, the second element indicates the weight of that topic and the words that follow are the most frequently occurring words that fall into that topic.

Validation and Evaluation Procedures

The validation procedure for the generated topic models involves consultation with field experts. The experts were composed of the Vice President for Academic Affairs, Student Development Services director and Institutional Planning officer. They were requested to examine the pool of words in each topic and identify the main theme of each set. A snapshot of 3 topics from the dataset randomly chosen with 10 top words along its weight is listed in Table 1.

Table 1. Top 10 Words from 3 Topics

Topic 5		Topic 0		Topic 4	
staff	57.11	Office	26.11	survey	12.11
approachable	38.11	Aircondition	16.11	online	9.11
accommodating	30.11	Ventilation	7.11	provide	5.11
personnel	28.11	Location	6.11	system	5.11
additional	11.11	Poor	5.11	confidentiality	4.11
improve	8.11	Restructure	3.11	area	4.11
employee	8.11	Change	3.11	improve	4.11
helpful	6.11	Additional	3.11	waiting	3.11
regular	5.11	Fine	2.11	facilities	3.11
permanent	5.11	Appliance	2.11	additional	3.11

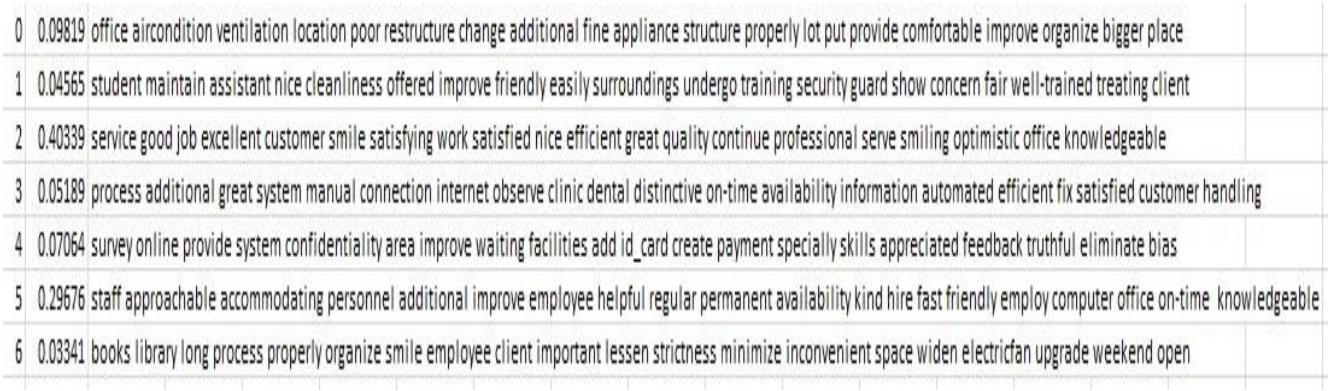

Fig. 2. list of generated topics

Fig. 3. Topic 5 word cloud representation.

Fig. 4. Topic 0 word cloud representation

Fig. 5. Topic 4 word cloud representation

To help the experts to easily see the words and their relative weights, individual topic was visualized through word cloud using Wordle application. The word cloud representations of the top words in selected topics are presented in Figure 3 to Figure 5.

After thorough examination on groups of words of each topic, the experts have labeled the topic models extracted from the feedbacks. The identified main themes are listed in Table 2.

Table 2. Labeled Topic Models

Topic No.	Main Topic/ Theme	Top Words
0	Office Environment	office aircondition ventilation location poor restructure change additional fine appliance structure properly lot put provide comfortable improve organize bigger place
1	Security Personnel and Student Assistant Attitude	student maintain assistant nice cleanliness offered improve friendly easily surroundings undergo training security guard show concern fair well-trained treating client
2	Customer Service Quality	service good job excellent customer smile satisfying work satisfied nice efficient great quality continue professional serve smiling optimistic office knowledgeable
3	Office IT System	process additional great system manual connection internet observe clinic dental distinctive on-time availability information automated efficient fix satisfied customer handling
4	Customer Feedback System	survey online provide system confidentiality area improve waiting facilities additional id_card create payment specially skills appreciated feedback truthful eliminate bias
5	Office Staffing	staff approachable accommodating personnel additional improve employee helpful regular permanent availability kind hire fast friendly employ computer office on-time knowledgeable
6	Library Operations and Management	books library long process properly organize smile employee client important lessen strictness minimize inconvenient space widen electricfan upgrade weekend open

Table 3. Probabilistic topic distribution over new document

# doc	New Text Data	Office Environment	Security Personnel and Student Assistant Attitude	Customer Service Quality	Office IT System	Customer Feedback System	Office Staffing	Library Operations Management
0	maintain office cleanliness	96%	0%	0%	1%	0%	0%	2%
1	lack of employee	21%	0%	0%	1%	0%	73%	4%
2	improve building facilities	69%	0%	0%	2%	0%	6%	23%
3	continue what you are doing, excellent service	0%	0%	98%	1%	0%	0%	0%
4	very accommodating	1%	0%	83%	0%	0%	11%	4%

Table 3 (cont.) Probabilistic topic distribution over new document

# doc	New Text Data	Office Environment	Security Personnel and Student Assistant Attitude	Customer Service Quality	Office IT System	Customer Feedback System	Office Staffing	Library Operations Management
5	some monitor are not working well	10%	2%	3%	31%	3%	6%	44%
6	install aircondition	91%	0%	0%	3%	0%	1%	4%
7	Add staff	1%	0%	0%	0%	0%	96%	2%
8	Online survey	1%	0%	0%	2%	95%	0%	2%
9	Employ regular staff	0%	0%	0%	0%	0%	97%	2%

The evaluation procedure, on the other hand, was done by using the trained model to infer new sets of document and manually inspect the topic assignments if the model correctly predicts the topic. Table 3 shows the result of the inference method on the new documents.

The result shows that the model predicted the topics of new documents correctly as seen in the first document which characterized mostly by the office environment. Also the fourth and fifth documents which refer to quality of customer service.

Topmost Customer Concerns

After analyzing the feedbacks from the customer satisfaction survey system database, the topmost customer concerns were identified and ranked by composition as shown in Table 3.

Table 3. Top 7 Customer Concerns

Rank	Main Topic/Theme	Weight
1	Customer Service Quality	0.403
2	Office Staffing	0.297
3	Office Environment	0.098
4	Customer Feedback System	0.071
5	Office IT System	0.052
6	Security Personnel and Student Assistant Attitude	0.046
7	Library Operations and Management	0.033

The result reveals that customer service quality is the topmost concern of the respondents. This is expected since the survey is with respect to customer satisfaction. Subsequently, additional concerns for offices were identified such as staffing, environment, customer feedback system, and IT system. Moreover, issues on security personnel and student assistants' attitude as well as library operation and management are notably highlighted in the feedbacks.

These were then presented to the administrators as inputs for decision and policy making since one of the recommendations included in the ISO audit findings for the school is the consolidation and clustering of customer feedbacks to be able to see their overall perception as well as for prioritization and focus of action needed.

CONCLUSION AND RECOMMENDATION

In this paper, the researcher presented how machine learning-based algorithm was utilized to analyze and determine the topmost concerns from the unstructured feedbacks of the customers. The study reveals that topic modeling using Latent Dirichlet Allocation (LDA) was effective in identifying the underlying themes in the customer feedbacks for the state college. The results were submitted to school administrators which are considered useful in making decisions and formulating school policies. Furthermore, it addressed the suggestions and recommendations cited in the ISO audit findings for the state college on the utilization of the customer feedbacks.

Hence, the researcher is recommending to employ the same method to other text-based survey conducted by the school to gain clear insights of the customers' sentiments and concerns. Furthermore, the school should consider an online survey system to encourage more respondents to write down their truthful feedback.

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude for all those who contributed in the realization of this study. Special thanks to my family for the support and inspiration and above all, to our Almighty Father, my source of strength, knowledge, and wisdom.

REFERENCES

[1] National Economic and Development Authority. (2017). *Philippine Development Plan* (Vol. 313). https://doi.org/10 July 2012

[2] Büschken, J., & Allenby, G. M. (2016). Sentence-based text analysis for customer reviews. *Marketing Science*. https://doi.org/10.1287/mksc.2016.0993

[3] Das, T. K., & Mohan Kumar, P. (2013). Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering and Technology*

[4] Chakraborty, G. (2014). Analysis of Unstructured Data : Applications of Text Analytics and Sentiment Mining. *SAS Global Forum, Washington D.C.* https://doi.org/10.1.1.456.2621

[5] Shahbaz, M., Guergachi, A., & Ur Rehman, R. T. (2014). Sentiment miner: A prototype for sentiment analysis of unstructured data and text. *Canadian Conference on Electrical and Computer Engineering*. https://doi.org/10.1109/CCECE.2014.6901087

[6] Zhan, J., Loh, H. T., & Liu, Y. (2009). Gather customer concerns from online product reviews - A text summarization approach. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2007.12.039

[7] Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*. https://doi.org/10.1509/jmr.12.0106

[8] Allahyari, M., Trippe, E. D., & Gutierrez, J. B. (2017). A Brief Survey of Text Mining : Classification , Clustering and Extraction Techniques. *ArXiv*, *abs/1707.0*.

[9] Blei, D. (2012). Introduction to Probabilistic Topic Modeling. *Communications of the ACM*. https://doi.org/10.1145/2133806.2133826

[10] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-018-6894-4

[11] Alsumait, L., Wang, P., Domeniconi, C., & Barbará, D. (2010). Embedding Semantics in LDA Topic Models. In *Text Mining: Applications and Theory*. https://doi.org/10.1002/9780470689646.ch10

[12] VijayGaikwad, S., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications*. https://doi.org/10.5120/14937-3507

[13] Katariya, N. P., & Chaudhari, M. S. (2015). Text Preprocessing for Text Mining Using Side Information. *International Journal of Computer Science and Mobile Applications*.

[14] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

COPYRIGHTS