Visitors Needs Analysis in Mall XYZ with Text Mining Analysis

Faurizal Limansyah, Mokh. Suef, and Vita Ratnasari
Department of Technology Management, Institut Teknologi Sepuluh Nopember, Surabaya
e-mail: faurizallilansyah@gmail.com

***Abstract*— Mall is a shopping center consisting of the main tenants (anchors), retail, shops, services, and public facilities that are built, managed, and arranged by the manager with the aim of conducting interaction between visitors and sellers. In 2018 Surabaya City had 33 malls or 50.77% of the total number of malls in East Java, this shows that the mall population in Surabaya City is very high. Mall XYZ is one of the malls that stood in the city of Surabaya and has been operating for one year. This mall still has not shown a development that is not optimal because of the many complaints from tenants to building management related to the sale of products that do not reach the target, as a result tenants often violate the rules (house rules) that have been set by managers to break the cooperation contract between building management and tenants unilaterally . To find out the cause of these conditions, text mining methods are used to filter reviews from visitors filtered from the Mall XYZ Google page. of the existing reviews classification will be based on reviews that contain positive and negative sentiment. review with positive sentiment becomes an appreciation that must be maintained and improved while a review containing negative sentiment becomes a complaint that must be addressed by the building management as the mall manager.**

***Keywords*— Needs, Mall, Visitors, Text Mining.**

I. INTRODUCTION

MALL has a sense as a forum in the community that enlivens the city or the local environment in addition to functioning as a place for shopping activities or buying and selling transactions, as well as a place to gather or relax [1]. The number of malls in Indonesia is increasing every year, accompanied by an increase in the population of Indonesia from year to year. The Central Statistics Agency noted that at least in 2018 there were 708 malls already established in Indonesia and 33 of them were in the city of Surabaya [2]. Mall XYZ is one of the malls that have been operating in the city of Surabaya. These conditions can be interpreted as lack of buying and selling transactions within Mall XYZ. This has a major impact on building management as the manager of Mall XYZ, starting from opening and closing shops arbitrarily, inserting or removing goods without confirming building management, to terminating the cooperation contract between building management and tenants unilaterally.

The perception of visitors through social media is a form of concern from visitors to the condition of the mall which will later build the mall to be even better, so that an opinion about the condition of the mall is very much needed for management improvement. The opinions given by visitors are positive and negative, This positive opinion includes visitor's appreciation of the condition of the mall which is satisfactory, while many negative opinions come from complaints about the condition of the mall that is not optimal.

An evaluation of the condition of the mall is needed, so that the mall can compete and still have an appeal to visitors so that it can run according to its function. Many methods can be used in classifying data, but this classification focuses on text data where the appropriate method is used is text mining. This study will examine what the needs of visitors, so that building management / mall managers can connect the needs of visitors with the condition of the mall.

II. METHOD

Text mining is a branch of data mining that aims to analyze data in the form of text. Text mining is a step of text analysis done automatically by computer software to dig up some quality information from a series of texts summarized in a document. The initial idea of making text mining is to find patterns of information that can be extracted from unstructured text [3]

The data used in this study is a collection of reviews obtained from Google users in Indonesia taken on March 1, 2019 to December 21, 2019. Data is obtained by manually updating data on every review on the Google Mall XYZ page.

Based on the data source obtained in the form of data review, then the data must be filtered by taking data that only contains sentiment. Explanation of the research variables can be seen in table 1, where the research variable consists of two different data where the x variable is visitor sentiment and the y variable is the i-word frequency that appears on the visitor review.

In this sentiment analysis, the form of the predictor variable used is the value of the weighted result so that it is denoted by w. Data structure before data pre-processing can be seen in Table 2.

After manually calcifying the data review, the next step taken is to process the data,

1. Remove a review that does not contain sentiment (positive or negative). [4]
2. Doing case folding, which changes all the text with lowercase letters (non-capital) and eliminating punctuation.[3]
3. Perform cleansing, which is the process of cleaning a review of noise. Words omitted in the review are punctuation characters, emoticons, hashtags (#), usernames (@username), and URL links. [4]
4. Perform tokenizing to break up tweets into word for word.[5]

Tabel 1.
Research Variable

Variable	Information	Scale of Data
x	Frekuensi kata ke-i yang muncul pada objek *(google review)*	Ratio
y	Sentiment (Positif/Negatif) 1 = Sentiment Negatif 2 = Sentimen Positif	Nominal

Tabel 2.
Research Data Structure Before Data Preprocessing

No	Pages	*Review (y)*	Sentiment Classification
1	*Google Pages Mall XYZ*	Akses menuju mall sangat kurang, security dan cs nya kurang ramah	Negatif
.		.	.
.		.	.
n		Mall paling sepi yg ada disurabaya, gimana nasib tenant tenantnya yah?	Negatif

Tabel 3.
Confusion Of Logistic Regression Matrix

Actual	Prediction	
	Negative	Positive
Negative	18	2
Positive	9	9

5. Perform stemming using the confix-stripping stemmer algorithm to get basic words. The word maker is basically based on a prepared list.[6]
6. Stopping process based on a stoplist that contains predetermined stopwords. The words contained in the review will be compared with the stopwords list, if there are words contained in the stopwords then the word will be deleted from the review so that identical keywords are found.
7. Change the review data into the frequency of occurrence of words using TF-IDF[6]

Word Cloud can represent a text data by plotting words that often appear. The more often the word appears, the bigger the letter of the word, as well as if a word rarely appears, the word size will be smaller than the others. Data visualization in this study will be carried out using the help of the website https://www.wordclouds.com/ with the results of the previous stage of the data processing.

After that to classify positive and negative sentiments using binary logistic regression analysis, with several stages including:

1. Divide data into two, the first is 80% training data and the second is 20% testing data.
2. Conduct an independence test using training data.
3. Form a Logistic Regression model using training data.
4. Test the significance of parameters individually and as a whole
5. Validate the accuracy of predictions from the model with data testing.

Calculating the value of classification accuracy.

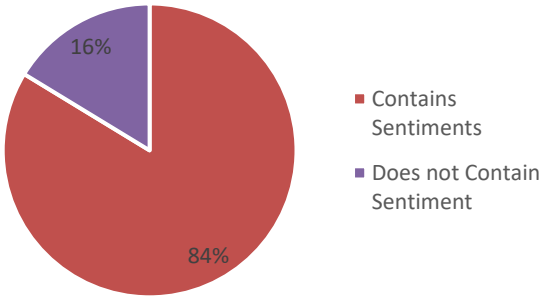

Figure 1. Amount of Data Containing Sentiment.

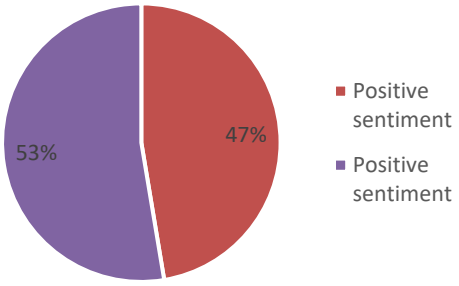

Figure 2. Data Characteristics Review.

III. RESULT AND DISCUSSION

All data that can be used (227 reviews) from the Google Mall XYZ page will then be selected, only reviews containing sentiments will be carried out further analysis. Rivew which contains sentiments is a review that is the expression or opinion of visitors, both negative and positive. While the review that does not contain sentiment is a review that is informative. Comparison of the total number of reviews that do not contain sentiment with the total number of reviews that contain sentiment can be seen in Figure 1.

Based on Figure 1 it can be seen that only 84% or equal to 190 reviews of the total reviews containing sentiment. While 16% does not contain sentiment. From the results of the classification of review data that has been selected, it can produce a percentage comparison of reviews that contain negative sentiments and reviews that contain positive in Figure 2.

Figure 2 shows that out of a total of 190 reviews containing sentiments, 53% of public reviews of Mall XYZ contain negative sentiments or complaints that are complaints, while the remaining 47% contain positive sentiments or visitor perceptions that are appreciative.

Mall XYZ visitor review data that has been obtained and classified, will then be filtered. Where every review sentence will be taken from the essence. the filtering is done by data processing. There are several steps that must be done in the data preprocessing, including the stages of cleansing, case folding, stemming, stopwords, and tokenizing.

Data visualization with Word Cloud is used to find out predictor variables (words) that are often used or often appear in data reviews. The data used in Word Cloud is review data that has been differentiated based on its classification, which

Review	*Review*	Masih sepi, stand cuma sedikit yang buka.. Hiburannya kurang nih...
Cleansing	Remove the username and URL link	Masih sepi, stand cuma sedikit yang buka.. Hiburannya kurang nih...
Case folding	Equalize letters and eliminate punctuation	masih sepi stand cuma sedikit yang buka hiburannya kurang nih
Stemming	Eliminating the word affix	masih sepi stand cuma sedikit yang buka hiburan kurang
Stopwords	Remove conjunctions	masih, sepi, stand, sedikit, buka, hiburan

Figure 3. Pre-Process Data Review Simulation

Figure 4. Visualisasi Word Cloud Data Review Mall XYZ

is positive or negative. General analysis, word cloud visualization in Mall XYZ visitor review can be seen in Figure 4. Data that can be processed in this study collected 190 data, of which 190 of these data contained negative and positive sentiments. Of course the data to be processed is data that has passed the preprocessing stage first.

The results of the Mall XYZ review data cloud show that the word "quiet" seems to dominate, this means that the word has a high frequency of appearing (often mentioned) in visitor reviews on the Mall XYZ google page. The results of this analysis are in accordance with the characteristics of the data that had been done in the previous stage, where there were 53% of reviews from visitors that contained negative sentiment. This also illustrates that the condition of the quiet mall into the spotlight for visitors. In addition, the words "parking", "clean", "open", "comfortable", "tidar", "road" are also clearly seen in the overall visualization of the word cloud review of visitors.

The next step is to analyze any positive words that often appear in Mall XYZ visitor reviews. So that the visualization using Word Cloud is done to find out the appreciation of

Figure 5. Visualization of Word Cloud Positive Data Review Mall XYZ

Figure 6. Visualization of Word Cloud Negative Data Review Mall XYZ

visitors about the condition of the Mall XYZ. The results of visualization can be seen in Figure 5.

Based on Figure 5 it can be seen that the word "good" has the most number of conversations. The next word is "comfortable", "clean", "shopping", "complete", "strategic", and other words. This shows that visitors really appreciate the condition of Mall XYZ. Most of the appreciation is aimed at Mall XYZ in terms of physical conditions (appearance, design, interior, landscape) of the Mall within the scope of the Mall XYZ area.

Appreciation of visitors to the good condition of this mall will certainly be a strength in doing business or still survive in the future. in accordance with the theory of understanding and purpose of the mall which has been described in Chapter II. Where the mall is a place for people to bring the city or the local environment aside from functioning as a place for shopping or buying and selling activities, as well as a place to gather or relax. There are also several words that refer to the names of tenants in Mall XYZ, such as miniso, supermarkets (lion superindo), guardian, karaoke (asang karaoke), and the existence of a food court managed by aiola group. An application that was offered to Mall XYZ which has invited tenants with national to international brands.

This condition is a positive value for Mall XYZ which can already read the needs of existing visitors, some tenants have been able to accommodate the needs of visitors and become

an attraction for other visitors. If seen from the results of the analysis that has been done. it can be concluded that the apartment located above Mall XYZ is a positive trend going forward. The concept of Mix-ed Use Building which is applied at the time of this building can provide convenience for its residents. So that residents do not need to spend a lot of time to find their needs.

Furthermore, using data from the same source, the Google Mall XYZ page, a visualization for review data is included in the negative category. From the visualization using Word Cloud this is done to find out the complaints of visitors about the condition of Mall XYZ. The results of visualization can be seen in Figure 6.

Based on Figure 4.5 it can be seen that the word "quiet" has the most number of conversations. These results are the same as the results in the previous stage, namely the visualization of data review in general and the results of the classification of reviews that contain negative sentiment. This shows that people still complain about the condition of the mall which is very quiet. The word lonely is interpreted in two different perspectives where it is devoid of visitors and lonely of tenants at Mall XYZ.

The condition of the deserted mall is a negative image for the mall, many reviews describe that the mall is deserted. this has become a condition that really needs to be handled well by building management to improve the positive image that has existed in the results of previous analyzes where the mall has a comfortable atmosphere to visit.

The next word is "tenant", "blm" (not yet), "crowded", "variation", "empty", "closed" if you look at the current condition of the mall, the tenant's existence is still considered lacking. many tenants came out because of the quiet mall conditions. this is one of the complaints of visitors. where the need for visitors can not be accommodated at this mall. this is supported by the words "electronics", "shoes", "foodcourt", and "eating" on the results of negative word cloud visualization. some tenants are still not available at Mall XYZ, this is an input for building management to invite tenants who sell some of these items. In addition, the types of food available either in the food court or not considered less varied by visitors.

From the preprocessing data, it can be continued with the analysis using the logistic regression method. Logistic Regression classification method is a statistical method that can be used to classify Mall XYZ review data into positive and negative categories. From the Mall XYZ review data that has been carried out analysis using the logistic regression method, it can be obtained the results of positive and negative classification. From the results of the classification using logistic regression, then it will be compared with the manual classification that has been done before so that accuracy can be obtained from the classification of the team. Based on the tests that have been done, a confusion matrix can be made as in Table 3.

After binary logistic regression analysis, the data will be evaluated for the performance of the classification method. Actual data and predictive data from the classification model are presented using a cross tabulation (Confusion matrix), which contains information about the actual data class represented in the matrix row and the predicted data class in column [7].

Based on the confusion matrix table above, it can be seen that the accuracy of classification for negative data that is classified as negative is 18 reviews, while for positive data that is predicted to be positive is 9 reviews. For inaccurate classification the total is 11 reviews. Furthermore, based on the confusion matrix the accuracy of logistic regression analysis can be calculated with the following calculation.

$$Classification\ accuracy = \frac{jumlah\ prediksi\ benar}{jumlah\ total\ prediksi} \times 100\%$$

$$Classification\ accuracy = \frac{TP+TN}{TP+FN+TN+FN} \times 100\%$$

$$Classification\ accuracy = \frac{9+18}{9+18+9+2} \times 100\%$$

$$Classification\ accuracy = \mathbf{71}\%$$

Based on the above calculations, it can be seen that the classification analysis using logistic regression can achieve a classification accuracy value of 71% in the Mall XYZ visitor review data.

IV. CONCLUSION

Visitors more often make complaints compared to appreciation for Mall XYZ. the percentage of complaints given by visitors is greater than the percentage of appreciation that exists for Mall XYZ, 53% of reviews obtained were complaints of visitors while the rest of 47% are rivals who are appreciative of visitors to the condition of Mall XYZ,

Judging from the needs of the items sought by visitors, not only goods are indeed sought after by visitors, but the need for services is also sought by visitors. The need for food, drinks, shoes, electronics, and medicines is a need that is often sought by visitors. In addition, the need for fulfilling one's mental, mental, and insight needs also cannot be separated when visitors are in the mall such as recreation, leisure, security, and worship. Mall XYZ visitors put more emphasis on the lively atmosphere in the Mall and the existence of varied tenants. The existence of a sense of comfort and security is also one of the needs of visitors when they are in Mall XYZ. The holding of the event at Mall XYZ is also something that visitors need as a place for recreation. The mall comfort level and the visual condition of Mall XYZ are positive values for visitors, this needs to be maintained in order to provide a positive image of Mall XYZ.

Visitors have appreciated the existence of tenants that have existed in Mall XYZ, therefore building management as the manager must maintain these tenants, some tenants that must indeed be maintained by the building management include lion superindo as anchor, miniso, guardian, and karaoke host . As for the tenants needed by visitors including electronic tenants and gadgets (gadgets), tenant sneakers (shoes), tenant garments (clothes), and adding variants of existing tenant food and baverage (food and beverages).

REFERENCES

[1]. B. Maitland, Shopping Malls: Planning and Design. New York, 1985.
[2]. Badan Pusat Statistik, “Pusat Perbelanjaan Menurut Provinsi 2018,” 2019.

[3]. S. M. Weiss, N. Indurkhya, T. Zhang, dan F. Damerau, Text Mining Predictive Methods for Analyzing Unstructures Information. New York: Spinger Science+Business Media. Inc., 2005.
[4]. G. A. Buntoro, T. B. Adji, dan A. E. Permanasari, "Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation," 2014.
[5]. L. Bing, Handbook of Natural Language Processing Second Edition. Boca Raton: CRC Press, 2010.
[6]. D. Ariadi dan K. Fithriasari, "Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine Dengan Confix Stripping Stemmer," J. Sains dan Seni ITS, vol. 4, 2015.
[7]. D. D. Hosmer dan S. Lemeshow, Applied Logistic Regression, 2nd ed. New York: John Wiley & Sons, Inc, 2000.