Sentiment Analysis of Restaurant Review with Classification Approach in the Decision Tree-J48 Algorithm

Masrur Adnan
Department of Business and Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
masrurmoch@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Kelly Rossa Sungkono
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
kelly@its.ac.id

***Abstract*—In the increasingly fierce restaurant business competition, many restaurants compete to provide the best quality for consumers. The quality of the restaurant includes food and drink, environment, place and service. The condition affects the restaurant brand image that is marked whether consumers are satisfied or not. One of review restaurant website called TripAdvisor is chosen as the data source because it contains User-Generated Content features (UGC). The advantage of using UGC is to ensure the authenticity of consumer comment data. English texts classification is used in this study to determine dissatisfaction (negative) and satisfaction (positive) of consumers based on their comment or review. This paper utilises Decision Tree-J48 as the classification method in this study. Data retrieval is done by crawling data using WebHarvy. The overall performance of the Decision Tree-J48 method is the average value of Precision 48.7%, Recall 36.8%, F-Measure 41.4% and accuracy 45.6%. The benefit of this classification results is as a recommendation for consumers to choose the best restaurant.**

Keywords— decision tree-j48; f-measure.; precision; recall; sentiment analysis

I. INTRODUCTION

In this era, restaurant business competition is getting tighter. Several people like and choose to build a restaurant business because they think that foods and drinks are the basic needs of everyone. The number of restaurants is increasing, even though there are a lot of restaurant owners who close their businesses. Many restaurants find difficulties in selling their products, although they have good quality. According to Kharadi and Patel, an analysis of the quality of a restaurant considers its products and services [1]. Govindarajan explicitly explained in his research that the measurement of restaurant quality is from food, service, place atmosphere, price, and feasibility [2].

Consumer satisfaction is a concern because it is the most crucial thing. Consumer satisfaction determines the quality of a restaurant. Even though there are restaurants that have a well-known brand and have many branches, but each of those restaurants has its disadvantages. Although the satisfaction of each consumer is different, the comments of the customer portray the condition of restaurants. The kind comments can attract other people to come to the restaurant, so business owners get the benefit from those comments. The business owners also get benefit from bad comments, because the bad public comments are free checking of their products and free advice to make the restaurant better [3]. Consumer comments written on the website are varied but can be classified into two categories: negative and positive. These comments can be analysed and used as well as structured data using the right techniques.

This paper aims to look for positive or negative judgments and assess the performance of the method in the reviews or comments on the website, especially at restaurants in Surabaya. These comments or criticisms are data in the form of text or word data. The text data will then be classified into negative or positive judgments using Decision Tree-J48.

II. LITERATURE REVIEW

Based on the research that has been conducted [4], Schrauwen conducted Sentiment Analysis based on classification using the Naïve Bayes algorithm, Maximum Entropy and Decision Tree Classifier. Performance evaluation is done by measuring Accuracy, Precision and Recall with the N-fold Cross Validation approach.

In another study [5], researchers measured the value of accuracy before and after the addition of the feature selection method using the Naïve Bayes and Adaboost methods. By using a combination of these methods, the research produces better accuracy values than using only one method.

Research on sentiment analysis [6]–[10] has also been done with the Probabilistic Latent Semantic Analysis method [9]. Data is taken from the title of the review, not the whole comment. The results of his research showed that the results of the identification reached 76% accurate.

In this study the review data or restaurant comments used are English data because it is a global language. This paper uses languages other than English has been done by using Cantonese using the Naïve Bayes method and Support Vector Machine [11]. In addition, there are also those that combine with the N-Grams adjectives using the Naïve Bayes classification method [12].

978-1-7281-3832-9/19/$31.00 ©2019 IEEE

III. METHODOLOGY

Fig.1 describes the methodology of this paper. This paper uses data reviews or comments on the web page of triapdvisor with the objects is restaurants in Surabaya as the source. Data is collected using WebHarvy software (see Fig. 2), where the format of the results is Excel file. The information contains the name of customer, star rating, comment title, comments, and the name of the restaurant. Then, the data is carried out by the preprocessing process with the Python programming language using the NLTK (Natural Language Tool Kit) library. The preprocessing stage consists of the case folding step, symbol removal, tokenisation, slang word conversion, stop word removal, and stemming stage. The calculation of the appearance of each word in a document by using Decision Tree-J48 produces the result of pre-processing.

A. Data Collecting

The data needed in this study is comment text or review data on a restaurant reviews website. The data contains comments or reviews of 10 restaurants in Surabaya, such as Steak Hut Manyar Kertoarjo, Asian King, Layar Seafood, and Bromo Cafe. The selection of triapdvisor as the chosen reviews website because the site haves User-Generated Content (UGC). User-Generated Content is content on a website created by the user and published on the website.

The process of retrieving specific data from the internet can be called scrapping data. In this study, this study utilises WebHarvy software to do scrapping data. Fig. 3 depicts a screenshot or review on the triapdvisor website and Fig. 4 describes the flow of scrapping process.

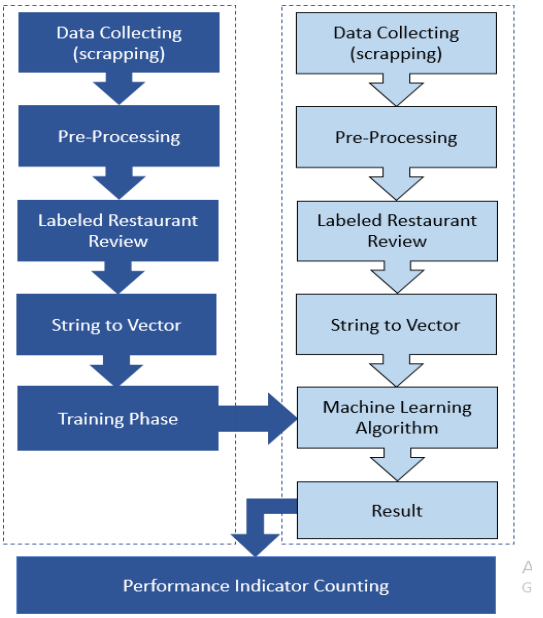

Fig. 1 Stage of research

Fig. 2 Comment view on triapdvisor website

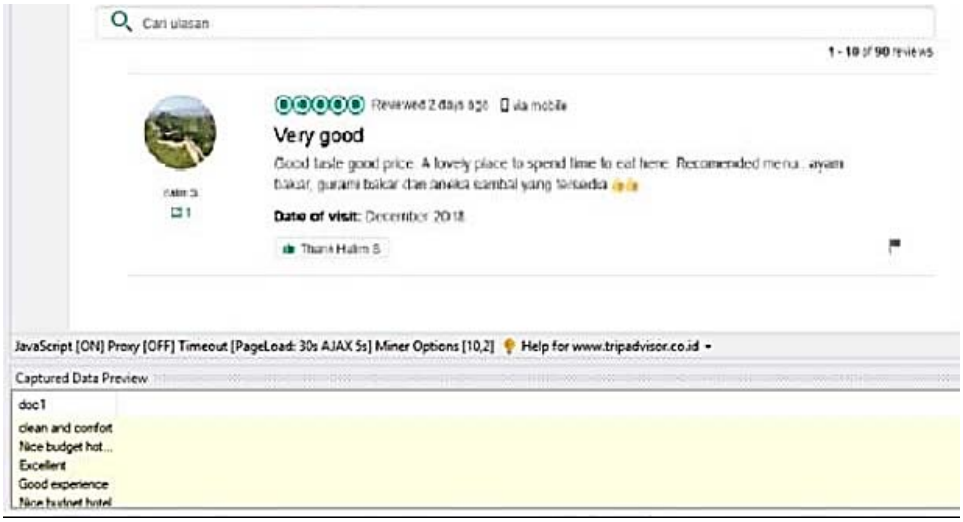

Fig. 3 WebHarvy software screen shoot

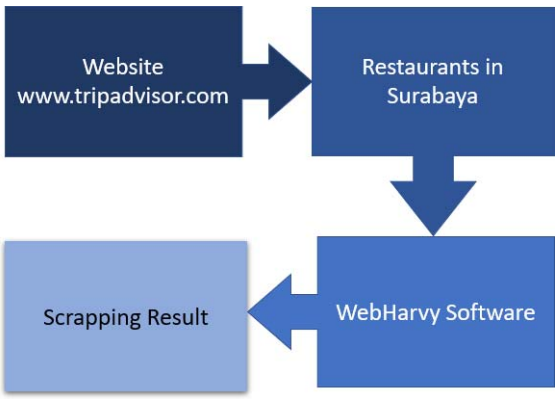

Fig. 4 scrapping data process using WebHarvy

B. *Preprocessing*

Preprocessing step is a crucial step to get clean and specific data so that later, the results of classification determination can be more accurate. In the preprocessing process, there are several parts of the process in it [13].

1. Case Folding is the step where the process of changing all words containing uppercase letters into lowercase letters.
2. Symbol Removal is the stage where the process of punctuation (comma (,), period (!), exclamation point (!), question mark (?), Etc.) is carried out, specific characters ($, %, #, &, and others), and numbers (0 to 9).
3. Tokenisation is the process in which a sentence is broken down into words. Usually done by spacing as a separator
4. Slang Word Convert is the process to change the existing non-standard words, into the official word, for example, the word of "woles" is replaced with the word of "relaxed", the word of "mbois" is replaced with the word of "cool", and others. This process can be done with the help of the slang word dictionary.
5. Stop word removal is a process where the non-essential words are removed in the classification process, for example: "at", "in", "for", "above", etc.
6. Stemming is a process where words are changed to the original form of the original word, for example, the words "technical" and "technically" are changed to "technic".

The Preprocessing process is using the IDLE Python framework software with using the Python programming language

C. *Decision Tree-J48*

Decision tree-J48, as shown in Fig. 5, is a data mining classification method in WEKA where the process is by converting a data into a decision tree -J48 with a decision rule.

Fig. 5 Decision Tree concept

The process starts with data at the root node, which then continues to select attributes that are formulated by the logic test on that attribute. The concept of data in the Decision Tree-J48 includes:

1. Data can be expressed into a table form with inside it is containing characteristics and records.
2. Features show parameters that are made as criteria in forming trees. One feature that states data per-item data solutions is called the target attribute.
3. Features have values called instances

The relationship between the preprocessing process and Decision Tree-J48 is shown in Fig. 6.

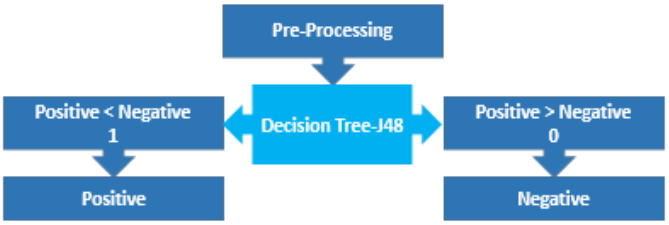

Fig. 6 Relations of preprocessing with Decision Tree-J48

Iterative Dichotomizer 3 (ID3) algorithm is one of the data mining models to produce decision tree -J48 based on available data. In this model, the concept used is the entropy of information, along with its working steps:

1. The information gain value is obtained using Equation (1) and Equation (2).

$$Gain(S,A) = Entropy(S) - \sum_{veNilai(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

where:

$$Entropy(S) = (-P_+ log_2 P_+) - (-P_- log_2 P_-) \quad (2)$$

Entropy(S) = The number of bits needed to extract the class (+ or -) from random data on S sample

S = Sample data used for training

P_+ = Number of positive solutions in the sample data

P_- = Number of negative solutions in the sample data

$log_2 P_+$ = Code length to state optimal information that has P probability.

2. The selected attributes are attributes that have the greatest information gain value.
3. Form a node containing the attribute.

Repeating the calculation process to get the information gain value until all data entered in the same class.

D. Confusion Matrix

A confusion matrix is a method that has the primary function to calculate the performance of a classification model according to the calculation of testing data, where the data that is the result of predictions are in two classes, namely positive class and negative class [14]. In the evaluation process using confusion matrix, the values of precision, recall value, and accuracy values are obtained from Equation (3) – (5).

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where:

TP = The number of positive cases is classified as positive

TN = The number of negative cases is classified as positive

FP = The number of negative cases is classified as negative

FN = The number of positive cases is classified as negative

E. Waikato Environment for Knowledge Analysis (WEKA)

WEKA is a workbench that makes it easy for users to implement Machine Learning techniques, including the implementation of Text Mining. WEKA makes it easy to use Preprocessing, Classification, Clustering, Select Attributes and Visualization of Results. The GUI of WEKA is shown in Fig. 7.

In this study, the WEKA feature used is the feature of the Decision Tree-J48 Machine Learning because it fits the needs. J.48 in WEKA is the development of ID3, which has a new feature that handles missing value, pruning, continuous attribute value range and derivation of the rule [15].

Fig. 7 The view of GUI on WEKA

IV. Experiment Result

The data source of this paper is derived from online reviews using scrapping or crawling — data taken from the page of triapdvisor with respondents in restaurants in Surabaya. The data received is a text that uses English. The entire document is 68. TABLE I shows the data from the labelling process.

TABLE I shows the results of labelling document data using machine learning and manually by the user. The data consists of 34 positive review data and 34 negative review data, which are then carried out using WEKA using Decision Tree-J48 method to produce data that some of them change. TABLE I shows the differences when manually labelling users with the results of using machine learning. 1 means the label of document is positive, -1 means the label is negative, and "+" behind the result of Decision Tree-J48 means the result of Decision Tree-J48 is different with the result of the expert.

TABLE II shows the classification of the review data using Decision Tree-J48, resulting in a precision level of data reaching 44% for negative data and 46.5% for positive data so that it obtained an average of 45.3% overall.

Whereas from TABLE III, it can be seen that the initial data that was processed contained 45.5882% data that was correct according to machine learning. Then as much as 54.4118% of the information is corrected because it is considered wrong by machine learning.

TABLE I. DATA FROM LABELING PROCESS

No	Document	Experts	J.48
1	about the steak well the taste was nice	1	-1+
2	about the place felt comfortable	1	-1+
3	overall worth to try this restaurant	1	1
4	best luck	1	-1+
5	affordable price good ambience recommended for family event	1	1
6	so many promotion from any credit card	1	-1+
7	great service affordable price for steak	1	1
8	usually I visited this restaurant once in a month	1	1
9	recommended steak restaurant in surabaya	1	1
10	not too crowded	1	1
11	cosy place and decent food	1	1
12	nice environment and atmosphere	1	1
13	nice meal but if it is a larger portion will be perfect	1	1
14	cozy place good service	1	-1+
15	good food and perfect salad	1	1
16	chicken cordon blue so delicious	1	1
17	love it	1	-1+
18	thumbs up	1	-1+
19	lunch with family at surabaya	1	-1+
20	good taste and good service at steak hut manyar ketoarjo	1	1
21	recommended	1	-1+
22	having a great lunch with fam	1	1
23	nice steak hut salad with nz sirloin steak	1	-1+
24	should come here next	1	-1+
25	convenience environment with great taste steak	1	1
26	also love the burger and the new chicken schnitzel	1	1

27	thanks for your good service and delicious beef steak ill be back here	1	1
28	thanks a lot	1	-1+
29	steak hut is restaurant that specialties on steak menu	1	1
30	usually have a promotion price cooperation with credit card from certain bank	1	-1+
31	good and delicious food clean and spacious place great for large groups good and friendly service	1	1
32	good food good price comfort place	1	1
33	must visit food in surabaya	1	1
34	grilled fish shrimp etc	1	-1+
35	I think they need more innovation for their menu	-1	1+
36	but to crowded if visit in weekend	-1	1+
37	area non smoking smoking small space	-1	-1
38	when I come here I thought it is only a small restaurant but I am totally wrong	-1	1+
39	have not been here for a long time and tried it for dinner the fish is a bit small tempe also small portion nasi goreng merah amazed us with the super tiny bit portion a small bowl and its	-1	1+
40	food price is expensive quality is average portion is small	-1	1+
41	sorry we are not coming back here again	-1	-1
42	but disappointed with their female staff	-1	-1
43	I just want to ask for spoon they serve their food without spoon to take up the food n no one coming	-1	1+
44	n no response	-1	-1
45	then she look back n yell at me with louder voice wait	-1	-1
46	at the end she is not coming to my table at all	-1	1+
47	n after sometime I ask to other staff when he take d other food and again without the spoon for the food	-1	1+
48	I know its just a staff but this just makes bad experience for coming here	-1	1+
49	but ac is not working well here	-1	1+
50	need more cool	-1	-1
51	now on a little bit expensive for the crabs prices and beside that for the parking area is very we can not imagine if there are 5 season	-1	1+
52	but very noisy lots of loud people in here	-1	1+
53	no privacy so I come here for a romantic dinner	-1	1+
54	if its good seafood you want go here but be prepared for a noisy evening	-1	1+
55	the restaurant is quite big and very crowded	-1	1+
56	but if u starving I go there	-1	-1
57	its quite take time especially when their restaurant very crowded	-1	-1
58	lobster are too expensive	-1	1+
59	very large restaurant crowded with local foreign	-1	-1
60	be prepared to be serve late menu	-1	1+
61	it's always crowded	-1	-1
62	the only thing that is the parking fee are not flat rate so the longer were there the more expensive the ticket is	-1	1+
63	this place always crowds when dinner time	-1	-1
64	variety of foods are many but nothing special	-1	1+
65	I only enjoyed ice cream as a solace	-1	1+
66	The food is terribly overrated	-1	1+
67	The chicken is dry and tough the marinade taste just like sweat ketchup	-1	1+
68	The price is not expensive, but the quality is poor for such a rating	-1	1+

TABLE II. QUALITY MEASUREMENT OF REVIEWS DATA USING DECISION TREE J.48

Class	True Rate	False Rate	Precision	Recall	F-Measure
Positive	0.324	0.412	0.440	0.412	0.425
Negative	0.588	0.676	0.535	0.324	0.403
Average	0.456	0.544	0.487	0.368	0.414

TABLE III. QUALITY MEASUREMENT BASED ON SEVERAL PARAMETER

Parameter	Value
Correctly Classified Instances	45.58 %
Incorrectly Classified Instances	54.42 %
Kappa Statistic	-0.0882
Mean Absolute Error	0.5372
Root Mean Squared Error	0.679

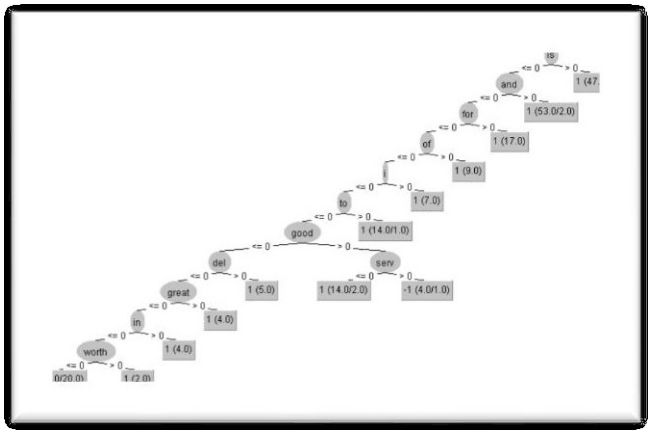

Fig. 8 Tree visualization

The results of the restaurant dataset classification using J.48 and filtering using StringToWordVector in WEKA are 25 Model Tree and 13 leaves. Fig. 8 is a visualisation of the data obtained using WEKA. TABLE IV, which is the result of the process in WEKA, is a Confusion Matrix. The accuracy of the Confusion Matrix is 45.6%.

TABLE IV. ACCURACY OF THE METHOD

Accuracy = 45.6%				
	Actual (Expert)			
	Negative		Positive	Precision
Decision Tree-J48	Negative	11	14	44%
	Positive	23	20	53.5%
	Recall	32.4%	41.2%	

V. CONCLUSION

Decision Tree-J48 is one method in WEKA for classifying a text. Decision tree-J48 is very simple and efficient. But in this study using restaurant review text data, a matrix configuration was produced, which had an accuracy of 45.6% and was an unfortunate result. The average precision value and recall value are respectively, 48.7% and 36.8%. In the future, it must be done by using other methods for comparison material. Or maybe a hybrid method might be used, namely the combination of the Decision Tree-J48 method used in this study combined with other methods.

ACKNOWLEDGMENT

Authors give a deep thank to Institut Teknologi Sepuluh Nopember, The Ministry of Research, Technology and Higher Education of Indonesia, *Direktorat Riset dan Pengabdian Masyarakat*, and *Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia* for supporting the research.

REFERENCES

[1] G. Vinodhini and R. M. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," *International Journal*, vol. 2, no. 6, pp. 282–292, 2012.

[2] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of Internet restaurant reviews written in Cantonese," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7674–7682, 2011.

[3] A. Reyes and P. Rosso, "Making objective decisions from subjective data: Detecting irony in customer reviews," *Decision support systems*, vol. 53, no. 4, pp. 754–760, 2012.

[4] R. Karsi, M. Zaim, and J. El Alami, "Impact of corpus domain for sentiment classification: An evaluation study using supervised machine learning techniques," in *Journal of Physics: Conference Series*, 2017, vol. 870, no. 1, p. 12005.

[5] G. A. A. J. Alkubaisi, S. S. Kamaruddin, and H. Husni, "Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers.," *Computer and Information Science*, vol. 11, no. 1, pp. 52–64, 2018.

[6] M. Fikri and R. Sarno, "A Comparative Study of Sentiment Analysis using SVM and SentiWordNet," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* vol. 13, no. 3, 2019.

[7] N. M. Elfajr and R. Sarno, "Sentiment Analysis Using Weighted Emoticons and SentiWordNet for Indonesian Language," in *International Seminar on Application for Technology of Information and Communication*, 2018, pp. 234--238.

[8] A. Firmanto and R. Sarno, "Prediction of Movie Sentiment Based on Reviews and Score on Rotten Tomatoes Using SentiWordnet," in *International Seminar on Application for Technology of Information and Communication*, 2018, pp. 202--206.

[9] D. A. K. Khotimah and R. Sarno, "Sentiment Detection of Comment Titles in Booking. com Using Probabilistic Latent Semantic Analysis," in *2018 6th International Conference on Information and Communication Technology (ICoICT)*, 2018, pp. 514–519.

[10] U. W. Wijayanto and R. Sarno, "An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naive Bayes," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 476--481.

[11] T. Pranckevičius and V. Marcinkevičius, "Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification," *Baltic Journal of Modern Computing*, vol. 5, no. 2, pp. 221–232, 2017.

[12] H. Kang, S. J. Yoo, and D. Han, "Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews," *Expert Systems with Applications*, vol. 39, no. 5, pp. 6000–6010, 2012.

[13] Y. A. Ahmed and M. Kumari, "Twitter Sentiment Visualization Using Deep Learning," 2018.

[14] P. Erdogmus, "Introductory Chapter: Swarm Intelligence and Particle Swarm Optimization," in *Particle Swarm Optimization with Applications*, IntechOpen, 2018.

[15] G. Kaur and A. Chhabra, "Improved J48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, 2014.