Advances in Social Science, Education and Humanities Research, volume 436
1st Borobudur International Symposium on Humanities, Economics and Social Sciences (BIS-HESS 2019)

Twitter Data as Decision Tree Parameter for Analysis of Tourism Potential Policies

Edy Subowo[1*], Imam Rosyadi[1], Hadwitya Handayani Kusumawardhani[1]

[1] *Informatic Management, Faculty of Engineering and Computer Science, Universitas Muhammadiyah Pekajangan Pekalongan, Indonesia*
[*]*Corresponding author. Email: edy.subowo@gmail.com*

ABSTRACT
This research provides an analysis of tourism potential in Pekalongan Regency based on Twitter media so that it can provide input for related agencies to develop new potential tourism objects. Decision tree method with C4.5 is used to classify positive reviews where tourists will visit again and vice versa negative reviews with tweet data parameters related to tourism object, like location access, service satisfaction, conditions and functionality of existing facilities, and shopping experiences. The results of the review classification from the decision tree serve as input to the promotion strategy to be applied in the website media. Training data collection of 250 tweets related to the name of the tourism object in Pekalongan Regency was conducted and the experimental process was carried out in testing the model using RapidMiner 5.3. Stratified sampling with the C4.5 Decision Tree obtained the highest accuracy rate of 92% using fold = 6:4. 20 tweet sampling tests related to the Welo Asri Petungkriyono found 17 positive review tweets with the most words related to the condition and functionality of existing facilities and 3 negative reviews related to location access and service satisfaction.
Keywords: *Twitter data, decision tree parameter, tourism policies*

1. INTRODUCTION

Tourism object is one of the commodities that is growing rapidly in Indonesia, including in Pekalongan Regency. The development of a culture of self-existence among millennial young people through social media adds to the emergence of new tourist attractions, which are largely the result of community self-help. Through social media young people can share information such as pictures, videos, comments, ratings and take into account the opinions of their closest friends, so that social media-based data recommendations are needed to get recommendations that are more accurate and in accordance with user tastes [1]. Millennial young people tend to get bored quickly and always look for new places to show their existence, so that less tourist attractions and promotions will lose visitors and close, therefore it is necessary to classify the parameters of visitor satisfaction based on satisfaction index based on service satisfaction [2].
Twitter API is used to get twitter data with the keyword tourist locations in Pekalongan district. The data text extraction can be used as a parameter of customer satisfaction parameters if the data has gone through pre-processing to get its basic words, literary library can be used because it has a time effectiveness of 97.72% [3].
Twitter's data mining process consists of information extraction, information summarization, and document classification [4]. Information extraction is carried out with a stemming process in which research on the Indonesian stemming process was first conducted by Asian [5], followed by Ciptaning [6] which uses the stemming process for document classification. The next process is information summarization where the extraction of features such as user data, data content, time, events as conducted by Zhou [7], while Kenta [8] extracts geotag features on Twitter data to display the accuracy of the location of the tweet created. These features are used to avoid the possibility of data errors like Hui [9] where the C4.5 method is used as an error data analysis in the data mining process. In the same data set, C4.5 method has an accuracy of 91.59% higher than ID3, fuzzy mathematic and pair analysis set [10]. Therefore, this study also tests the feasibility of tourist visitor satisfaction parameters such as access to locations, service satisfaction, conditions and functionality of existing facilities, as well as shopping experience from tweets obtained for the classification process using the C4.5 algorithm.

2. METHOD

The Process of text mining is divided into two processes, the first process is the process to get a classification rule, where training data is 250 twitter data with the keyword names of all attractions in Pekalongan collected for classification process using C4.5 algorithm so that the accuracy value is obtained. Accuracy standards in accordance with the results of Wang's research [10] that is equal or more than 91.59%. Continued visitor satisfaction parameter testing is performed until the desired accuracy level is reached so that the output obtained is rule classification based on visitor satisfaction parameters in the form of word classifications showing positive and negative reviews.
The second process starts from mining the testing data to then do the classification process with classification rules so

that we get the frequency of positive reviews and negative reviews along with the contents of the reviews. The results of the review classification from the decision tree are used as input to the promotion strategy for the relevant agencies to be applied in the media website. Figure 1 shows the method used starting from the text mining process to the classification with C4.5 in this study.

The tool used in this study is a computer with an Intel i3 processor and 5GB RAM. The software used in this study is Python 2.7.15 while the media website uses PHP language with data processing using MySQL. Google Cloud is used as a media for automatically retrieving Twitter data every 12 hours. Python libraries used are numpy, scikit-learn, tweepy, literary, xlsxwriter, and pymysql.

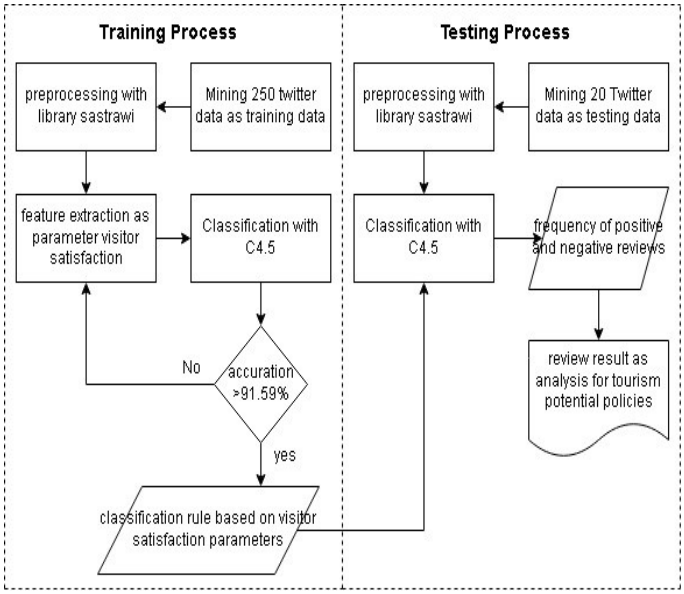

Figure 1 Research process

Processing starts from the stemming process by breaking down each word and eliminating symbols and numbers, the process of stopping word removal to eliminate words, conjunctions, feature extraction to get the results of the data needed for classification, and the labelling process for class determination. Calculation of the appearance of words in 250 training data tweets is then grouped according to class expressions. Table 1 shows how word expressions are formed from pre-processing along with the conclusion words from a tweet. Words that show expressions based on the class will be carried out a classification process with C4.5, Cross-Fold Validation is used to get the accuracy value of the system by matching the C4.5 result label with the description label of the tweet sentence as described in Table 1.

Table 1 Pre-processing process on a tweet

No	Real Words	Stemming	Stop Word	Feature Extract	Label (result)
1	welo the coolest, the path is smoothest, the service is okay, the place is good, eating and drinking is also delicious, especially the coffee, it's tastefully	welo the cool, the path is smooth, the service is okay, the place is good, eat and drink is also delicious, special the coffee, it's taste	Welo **cool** path **smooth** service **okay** place **good** eat drink also **delicious** special coffee taste	**cool** ->labelling result R(+) **smooth** -> nice access P1(+) **okay** -> satisfied service P2(+) **good** -> good condition P3(+) **delicious** -> satisfied shopping P4(+)	Positive Review (normal term)
2	Go to linggo is easy, service is good, the snacks are cheap, the facilities are complete, but there is nothing new, lazy to go there now	Go to linggo is easy, service is good, the snack are cheap, the facility are complete, but there is nothing new, lazy to go there now	Go linggo **easy** service **good** snack **cheap** facility **complete lazy** go	**Lazy** -> labelling result R(-) **easy** -> nice access P1(+) **good**-> satisfied service P2(+) **complete** -> facility P3(+) **cheap** -> satisfied shopping P4(+)	Negative Review (not normal term)

Twitter data in the training process is 250 tweets that have at least 2 parameters (P) and labelling adverbs (result) R. In normal rules, if P1 (+) and P2 (+) and P3 (+) and P4 (+) then R (+) according to example number 1 From Table 1, while in number 2 an abnormal condition occurs so that the rule cannot be used. Because tweets like in example number 2

will not be used as input data, other than that research on the causal rules of parameter-results will also be studied as rules of the decision tree with C4.5 to be made.C4.5 Algorithm.

In general, the C4.5 algorithm for constructing a decision tree is to select attributes as the root, create branches for each value, divide cases in branches, and repeat the process for each branch until all cases in the branch have the same class. To select the root attribute, it is based on the highest gain value of the existing attributes. To calculate the gain used the formula as stated in equation 1 below:

$$Gain\ (S.\ A) = Entropy\ (S) - \sum_{i=1} \frac{}{|S|} * |S^5| \quad Entropy \quad (S_5)$$

*Gain S.A=Entropy S-$_N$ _i=1^N?S_i/S*Entropy (S_i)* (1)

$$Entropy\ (S) = \sum_{i=1} -pi * log_2 \quad pi$$

Entropy S= _i=1^ N? *-pi*log_2 pi* (2)

Where :

S : Case Set; N: Number of attribute A; |*S*=|: Number of cases in S; A: Attribute; |*S*5|: Number of cases in the –i partition

Next step, labelling experiment scenarios are made as shown in Table 2 using the causal rule of the four parameters using the C4.5 algorithm to obtain decision tree rules as the output of the training process. The output of the training process with the accuracy of the results of the cross-fold validation is used as a rule classification algorithm for C4.5 to be used as a rule for the testing process.

Table 2 Trial rule of if-else scenario for labelling

Sce.	Rule
S1	*if P1(+) or P2(+) or P3(+) or P4(+) then R(+)*
S2	*if P1(+) and P2(+) or P3(+) or P4(+) then R(+)*
S3	*if P1(+) and P2(+) and P3(+) or P4(+) then R(+)*
S4	*if P1(+) and P2(+) and P3(+) and P4(+) then R(+)*

Cross-fold validation to optimize the cut-off value in stepwise regression is very helpful in adjusting an accurate predictive regression model [11]. The purpose of validation is to obtain the composition of training data and test data that provide the most optimal accuracy results called fold values. In Table 3, true positive (TP) is the number of positive records classified as positive, false positive (FP) is the number of negative records classified as positive, false negatives (FN) is the number of positive records classified as negative, true negatives (TN) is the number of negative records classified as negative. After the test data is classified, a confusion matrix will be obtained so that the amount of accuracy can be calculated.

Table 3 Validation table

Classification	Predicted Class	
	Class = Yes	Class = No
Class = Yes	a (true positive)	b (false negative)
Class = No	c (false positive)	d (true negative)

3. RESULTS AND DISCUSSION

To get 250 Twitter data as training data, data mining is done by keyword list of tourist objects in www.pekalongankab.go.id and only taking the user name, time, content and location of the tweet created. Data twitter is restricted from 2019 and the location of the tweet is Pekalongan Id. The Twitter geotag feature is used to filter the location of tweets created. Table 4 shows the frequency of words that appear.

Table 4 Word Frequency based on 250 tweets as training data

Parameter	Class	Word Content
Access to tourism object (P1)	bad (-) nice (+)	broken (3), hole (7), repair (2), stray (11), jammed (33) smooth (84), glide (17), fast (3), easy (2), safe (13)
Service satisfaction (P2)	disappointed (-) satisfied (+)	fierce (2), sour (87), angry (1), impatient (29) friendly (39), smile (51), kind (3), okay (37), cute (5)
Conditions, functionality and facilities (P3)	poor (-) good (+)	Ugly (94), broken (87), scary (2), dull (44) Function (6), good (17), well (7), complete (5)
Shopping Experiences (P4)	sad (-) happy (+)	Expensive (137), stale (8), disgust (2) Tasty (23), cheap (3), full (53), delicious (32)
Result /adverb (R)	negative (-) positive (+)	Lazy (2), no more (7), give-up (3) Cool (87), again (15), remember (1), happy (39)

The word content data in Table 4 is used as a determination of labelling with the if-else rule in Table 2 with the criteria for obtaining an accuracy value with the validation rules as shown in Table 3. The determination of the if-else rule uses the division of training data and test data called the value fold of 250 twitter data from the training process. Manual labelling as shown in Table 1 is used as a comparison of the predictive value of labelling with the C4.5 algorithm as shown in Figure 2. The 6: 4-fold value in scenario 3 has the highest accuracy value with 100% manual labelling accuracy and C4.5 accuracy is 92%. The accuracy value of C4.5 is in accordance with the accuracy standard according

to Wang's research results [10] which is equal to or more than 91.59%. The if-else rule in scenario 3 with 6: 4-fold accuracy is used as the C4.5 labelling rule in the testing process. Table 5 shows the if-else rules.

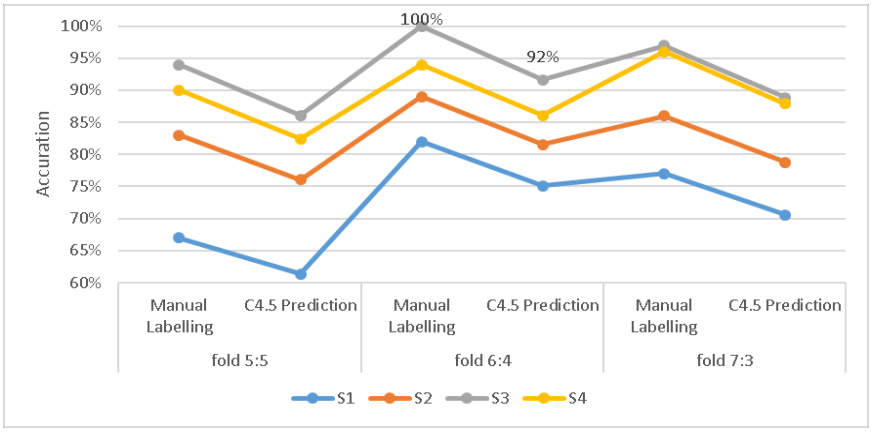

Figure 2 Accurate result with Cross-fold validation on manual labelling vs C4.5 labelling prediction

Table 5 Labelling rule on testing process

Access to tourism object (P1)	Services Satisfaction (P2)	Conditions, function, facilities (P3)	Shopping Experiences (P4)	Labelling Result (R)
Bad (-)	Disappointed (-)	Poor (-)	Sad (-)	negative review
Bad (-)	Disappointed (-)	Poor (-)	Happy (+)	negative review
Bad (-)	Disappointed (-)	Good (+)	Sad (-)	negative review
Bad (-)	Disappointed (-)	Good (+)	Happy (+)	negative review
Bad (-)	Satisfied (+)	Poor (-)	Sad (-)	negative review
Bad (-)	Satisfied (+)	Poor (-)	Happy (+)	negative review
Bad (-)	Satisfied (+)	Good (+)	Sad (-)	negative review
Bad (-)	Satisfied (+)	Good (+)	Happy (+)	positive review
Nice (+)	Disappointed (-)	Poor (-)	Sad (-)	negative review
Nice (+)	Disappointed (-)	Poor (-)	Happy (+)	negative review
Nice (+)	Disappointed (-)	Good (+)	Sad (-)	negative review
Nice (+)	Disappointed (-)	Good (+)	Happy (+)	positive review
Nice (+)	Satisfied (+)	Poor (-)	Sad (-)	negative review
Nice (+)	Satisfied (+)	Poor (-)	Happy (+)	positive review
Nice (+)	Satisfied (+)	Good (+)	Sad (-)	positive review
Nice (+)	Satisfied (+)	Good (+)	Happy (+)	positive review

Twitter data for the testing process is not filtered like normal not notes like example number 2 in Table 1 also filtering the word parameter (P) and result description (R) that must be in the training data. Figure 3 shows the interface of the application.

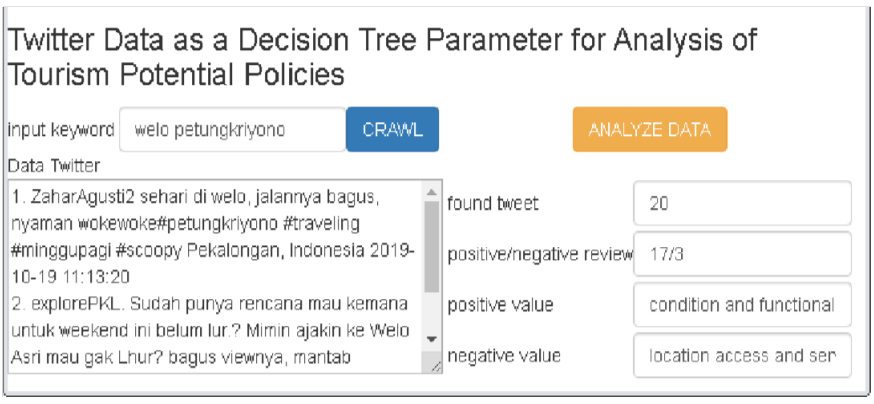

Figure 3 Testing process user interface

By using labelling rules as shown in Table 5, the decision tree C4.5 algorithm is used to obtain a review of a tourist attraction. Figure 5 shows the interface of the testing process with input of 20 tweet sampling tests related to Welo Asri Petungkriyono found 17 positive review tweets with the most words related to the condition and functionality of existing facilities and 3 negative reviews related to location access and service satisfaction.

4. CONCLUSION AND FUTURE WORK

This research provides an analysis of tourism potential in Pekalongan Regency based on Twitter media so that it can provide input for related agencies to develop potential new tourism objects. Twitter data is broken down based on the basic words and then grouped according to the frequency of occurrence of words.

The results of the review classification from the decision tree serve as input to the promotion strategy to be applied in the website media. Training data was collected on 250 tweets related to the name of the attraction in Pekalongan Regency. Stratified sampling with the Decision Tree method obtained the highest level of accuracy by 92% using fold = 6: 4. 20 tweets sampling test related to the Welo Asri Petungkriyono tourism object obtained 17 positive review tweets with the most words related to the condition and functionality of existing facilities and 3 negative reviews related to location access and service satisfaction. This can be used as a reference for related agencies to develop these attractions. In the future, further research is needed regarding deep learning methods with cloudbased paid Twitter Warehouse API.

REFERENCES

[1] J. Borràs, A. Moreno, and A. Valls, “rt Syst. Appl., vol. 41, no. 16, pp. 7370–7389, 2014.

[2] Y. Jiang, J. Shang, and Y. Liu, “Maximizing customer satisfaction through an online recommendation system: A novel associative classification model,” Decis. Support Syst., vol. 48, no. 3, pp. 470–479, 2010.

[3] N. Yusliani, R. Primartha, and M. Diana, “Multiprocessing Stemming: A Case Study of Indonesian Stemming,” Int. J. Comput. Appl., vol. 182, no. 40, pp. 15–19, 2019.

[4] B. Pang and L. Lee, Opinion Mining and Sentiment Analysis: Foundations and Trends in Information Retrieval, vol. 2, no. 1–2. 2008.

[5] J. Asian, H. E. Williams, and S. M. M. Tahaghoghi, “Stemming Indonesian,” Conf. Res. Pract. Inf. Technol. Ser., vol. 38, no. 4, pp. 307–314, 2005.

[6] H. T. Ciptaningtyas, “Enhanced confix stripping stemmer and ants algorithm for classifying news document in Representation of Textual,” Technology, pp. 149–158, 2007.

[7] Y. Zhou, S. De, and K. Moessner, “Real world city event extraction from Twitter data streams,” Procedia - Procedia Comput. Sci., vol. 98, no. DaMIS, pp. 443–448, 2016.

[8] K. Oku, F. Hattori, and K. Kawagoe, “Tweet-mapping method for tourist spots based on nowtweets and spot-photos,” Procedia Comput. Sci., vol. 60, no. 1, pp. 1318–1327, 2015.

[9] H. L. Han, H. Y. Ma, and Y. Yang, “Study on the Test Data Fault Mining Technology Based on Decision Tree,” Procedia Comput. Sci., vol. 154, pp. 232–237, 2019.

[10] X. Wang, C. Zhou, and X. Xu, “Application of C4.5 decision tree for scholarship evaluations,” Procedia Comput. Sci., vol. 151, no. 2018, pp. 179–184, 2019.

[11] Z. Mahmood, “On the Use of K-Fold Cross-Validation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise Regression On the Use of K-Fold CrossValidation to Choose Cutoff Values and Assess the Performance of Predictive Models in Stepwise Regression,” vol. 5, no. 1, 2009.