Aspect-Based Sentiment Analysis on Natural Tourism in West Bandung Using Multinomial Logistic Regression Algorithm

Tammara Audina Putri, Faqih Hamami*, and Ekky Novriza Alam
School of Industrial and System Engineering, Telkom University, Bandung, Indonesia
*faqihhamami@telkomuniversity.ac.id

Abstract. Now is the moment for the tourism market to recover after being hit by the Covid disaster. This recovery is accompanied by improvements in various aspects related to the tourism sector by the government, including the Tourism and Culture Office of West Java Province. Given the immense potential of the tourism sector in West Java, there is a need for better utilization to restore the tourism sector, especially the natural tourist attractions in West Bandung Regency. To determine the aspects that need development, it is essential to listen to public opinions. This research aims to analyze aspect-based sentiment on the natural tourist attractions in West Bandung Regency based on reviews on Google Maps. The research process is based on the Knowledge Discovery in Database (KDD) data mining and utilizes the Multinomial Logistic Regression algorithm. Conducting this research enables the prediction of public reviews regarding natural tourist attractions from various aspects and sentiments. The aspects considered in this research include accessibility, facilities, and activities. The results of this research focus on the impact of pre-processing techniques and oversampling methods on the F-1 score performance of the model. The research shows that the stemming pre-processing technique (SM) and emoji processing (EP) yield the best performance in the Multinomial Logistic Regression algorithm. Additionally, the ROS method for oversampling significantly improves the model's performance.

Keywords: Natural Tourism, West Bandung Regency, Aspect-Based Sentiment Analysis, Multinomial Logistic Regression, Google Maps Reviews.

1 Introduction

The pandemic first occurred in Indonesia in 2020. One of the sectors affected by the COVID-19 pandemic in Indonesia was the tourism sector. The tourism sector is affected by total visits by foreign and domestic tourists, which has a fatal impact on the Indonesian economy. This decline in the number of tourist visits has had a huge impact on the Gross Domestic Product (GDP) of the tourism sector. With total losses achieved by Indonesia amounting to USD 3.4 billion. Tourism is considered the lo-

A. Putro Suryotomo and H. Cahya Rustamaji (eds.), *Proceedings of the 2023 1st International Conference on Advanced Informatics and Intelligent Information Systems (ICAI3S 2023)*,
Advances in Intelligent Systems Research 181,
https://doi.org/10.2991/978-94-6463-366-5_11

comotive of the nation's economic movement, and the government makes this sector a priority for continued recovery.

If a line is drawn, the tourism sector can be a source of income for GDP and foreign exchange. Quoting from iNews.id, Minister of Tourism and Creative Economy (Menparekraf) Sandiaga Salahuddin Uno said that 2022 will be the year of recovery for the tourist market. Seeing the number of domestic tourists in 2021 which has increased compared to the previous year, supports the government's statement that 2022 will be the year the tourism sector recovers after the pandemic. This recovery is of course followed by improvements by the government in various matters related to the tourism sector. Of course, we must pay attention to the behavior of domestic tourists from the pandemic to post-pandemic transition.

There is a new characteristic in tourist behavior after the pandemic, namely that tourists will prefer natural tourism with short travel times [1]. From this, local governments that manage tourist attractions can focus on natural tourist destinations and also other factors that attract tourists. The Central Statistics Agency (BPS) stated that the provinces with the most six types of tourism groups were led by West Java with 427 tourist attractions, then East Java with 420 tourist attractions and finally Central Java with 285 tourist attractions. If you look at the total tourist attractions, West Java has a lot of tourist attractions, but it is very unfortunate that natural tourism is in last place of the three provinces. In fact, seeing the huge potential that West Java has, natural tourism can be increased considering that after the pandemic the government wants to improve the quality of natural tourism.

As for the changes to the West Java Province Tourism and Culture Department's Strategic Plan for 2018-2023, it is also stated that the attraction for West Java Tourism enthusiasts is to build experiences of active community involvement (responsible practices). Public opinion is often found in everyday life. For example, word of mouth, where people around you give their opinions. A good opinion is one that is based on information that comes from trusted sources [2]. Data mining methods can be carried out by the West Java Province Tourism and Culture Office to analyze public opinion in order to improve the tourism sector again. Implementing this method will also make the process automatic and does not need to be touched much by human hands because it uses a computer to look for patterns.

It is necessary to carry out sentiment analysis which has aspects in it so that the Department of Tourism and Culture knows the aspects that need to be improved and maintained, so that the focus on improving the quality of natural tourist attrac-tions can be more specific. It is very necessary to carry out research related to as-pect-based sentiment analysis so that the Department of Tourism and Culture can know which of the five aspects need to be prioritized for improvement and also which ones need to be maintained. There is previous research that is similar to this research. For example, research entitled "Aspect-Based Sentiment Analysis and Topic Modeling in Tourism Destinations Based on Google Maps and Tripadvisor User Reviews: Case Study of Borobudur and Prambanan Temples" is a thesis written by Arianto & Budi (2021). The overall result of this research is that the Logistic Regression model can predict data well in almost all scenarios in every aspect of this research [7].

Regarding aspects that need to be paid attention to by the government, this is confirmed by the statement from the Head of the West Java Tourism and Culture Department in 2022, Benny Bachtiar, that there are 5 aspects that need to be improved. The five aspects are accessibility, accommodation, attractions, activities and amenities. Next, it will be combined into 3 aspects, namely accessibility, facilities and activities. Then, created data modelling with *multinomial logistic regression* because this research used more than 2 sentiments (multiclass). It is hoped that this research will provide a solution to the West Java government, especially for the West Bandung Regency, so that it can improve the quality of natural tourist attractions based on aspects tested based on tourist review data from Google maps reviews. This will really support the West Java Tourism and Culture Department's desire to focus on consumer centric, namely consumer-based and listening to what tourists want.

2 Method

2.1 Theoretical Review

Aspect-Based Sentiment Analysis (ABSA) is a sub-area of sentiment analysis that can allow someone to gain deeper insight into the features of items that interest users by mining reviews [3]. Different from usual sentiment analysis, by carrying out this aspect-based sentiment analysis, there are labels for the aspects that will be tested in a problem. Sentiment analysis is usually only categorized into negative or positive opinion categories. By labeling this aspect, broader insights will be obtained. With that, ABSA (Aspect-Based Sentiment Analysis) is a text mining method for classifying text data that can be taken from reviews into several sentiment classes and aspect labels.

In this study, multinomial logistic regression will be used because it can produce more than one class to classify the aspects that will be used in this research. In multinomial logistic regression there are more than two dependent variables [4]. An example is when a customer makes a decision to buy or not buy an item, then a product is considered to pass or fail quality control. This algorithm can be used to perform aspect-based sentiment analysis because it can check several classes, not just binary. The difference with binary logistic regression which is often used to carry out classification is the use of a function to model probability relationships between classes. Multinomial logistic regression uses a function called softmax, while binary logistic regression uses the sigmoid function [5]. The way multinomial logistic regression works can be represented in the following image.

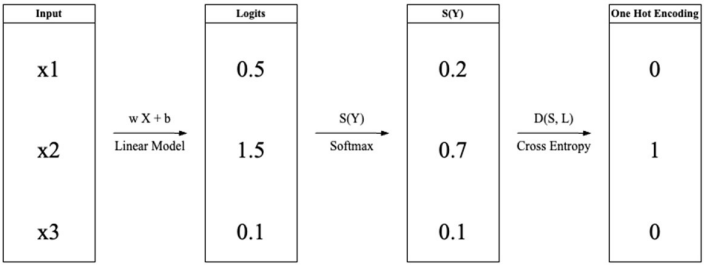

Fig. 1. How the Multinomial Logistic Regression Model Works

The workings of the multinomial logistic regression model are divided into input, linear model, logits, softmax function, cross entropy, and one-hot encoding. Starting from the input which is the features we have in the dataset. For example, when there is a dataset about Iris flowers, the features taken are sepal length, width and petal length and width [6]. Next, these inputs are entered into the linear model shown in the figure, where X is a set of inputs. X can be described as a matrix which contains all the features x2*w2, and x3*w3. Next is logits which are the output of the linear model and can be called scores. Then, put it into the softmax function.

Evaluation metrics aim to evaluate the performance of the classification model. The evaluation metric used is the confusion matrix which consists of actual and predicted. The use of a confusion matrix is important for models that produce two or more classes.

2.2 Knowledge Discovery in Databases (KDD).

Research framework was established to ensure clear planning of the research steps. This study is structured into six categories based on KDD (Knowledge Discovery in Database) which include initialization, data collection, data preprocessing, data transformation, data mining, and evaluation.

a. Initialization

The initialization of this research started from identifying problems that were occurring in the surrounding environment. Identification of this problem starts from observing the current situation in Indonesia which is currently in a pandemic period, then reduces the problem to look at West Java and ends at West Bandung Regency, then looks for problems related to the pandemic.

b. Data Collection

After initialization is complete, we proceed to the data retrieval stage. Starting with data scraping, namely collecting data from Google Maps reviews, especially on selected natural tourist attractions in West Bandung Regency. The selected natural tourist attractions are based on official website of West Bandung Regency and references from researcher that recommends the placesm also we sorted the reviews

with 'relevant'. Next, a dataset was obtained containing reviews from tourists at the selected natural tourist attractions. The total of scraped raw data was 13.587. The dataset will be labeled according to the needs of the research, namely the aspects to be researched. There was challenge in labelling, the dataset was imbalance.

c. Data Pre-Processing

After the data collection stage is carried out, it continues with data pre-processing which consists of several procedures, as follows.

(1) Data selection and data filtering and, in this section, we will filter the data and display data that is considered valid. The goal is to select relevant data for processing.
(2) Data cleaning, after the data has been filtered it is cleaned in several stages. Spelling correction, Whitespace &\n removal, Emoji processing, Punctuation removal, and Case folding.
(3) Stopword removal, which is the process of deleting a large number of common words that do not have meaning. For example, the word 'yang'.
(4) Stemming, which is the process of removing front and back affixes, so that only the base words are displayed.
(5) Tokenization, namely the process of changing a sentence into a smaller form, usually called a token.

d. Data Transformation

After data pre-processing is carried out, the next stage is split data or dividing the dataset into training data and test data. Enter the feature extraction stage using TF-IDF (Term Frequency - Inverse Document Frequency) which is an algorithm method for calculating the weight of each word used and converting it into a vector. Next, the data is divided into training data and testing data. The final stage at this stage is oversampling the training data to solve the imbalanced data problem.

e. Data Mining

The next stage is the implementation of the dataset that has been oversampled into a classification model using the multinomial logistic regression algorithm. By carrying out this classification, a model is formed based on the experiments that will be carried out and the modeling results are tried to be predicted using the test data that has been prepared.

f. Evaluation

At the evaluation stage, we will calculate the value from the model performance evaluation and proceed to cross validation to find out whether the model created falls into the overfitting category. Next, the results of these values are analyzed to gain insight into each experiment carried out.

3 Result and Discussion

3.1 First Testing Scenario

The first test scenario was carried out experimentally based on the use of the original dataset and pre-processing only. There are 5 scenarios that will use the original dataset and pre-processing stop word (SR), stemming (SM), and emoji processing (EP). This test scenario refers to a reference that uses these 3 pre-processing techniques [7]. Pre-processing will be combined to get the best results for the algorithm model used. The stages carried out are split data with a ratio of 70% training data and 30% testing data, converting tokens into vectors (TF-IDF), multiclass & multilabel modeling, model evaluation, and data validation using repeated K-Fold cross validation. The following is the first test scenario. The first scenarios for the model can be seen in **Error! Reference source not found.**.

Table 1. Research Testing Scenarios

Scenario	Dataset		Pre-Processing		
	Original	ROS	SR	SM	EP
Scenario 1	✓	×	✓	✓	✓
Scenario 2	✓	×	✓	×	✓
Scenario 3	✓	×	×	✓	✓
Scenario 4	✓	×	✓	×	×
Scenario 5	✓	×	×	✓	×

From these scenarios, different evaluation results were produced. The following are the results of the evaluation shown in **Table 2**.

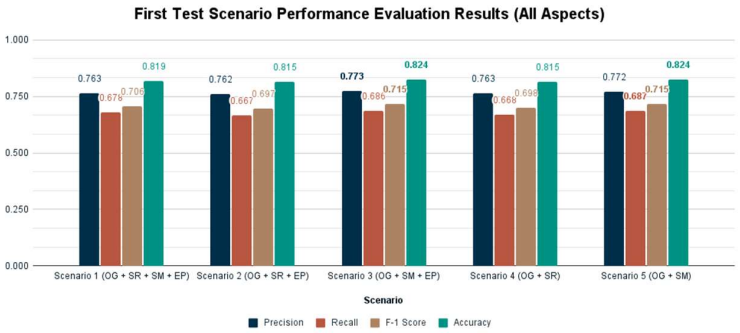

Fig. 2. First Test Scenario Performance Evaluation Results

Each scenario in the first test has a different pre-processing technique. Based on the table above, each value is generated from the average of the three aspects. The best F1-score was produced by scenarios 3 & 5 with a value of 0.715 and precision for scenario 3 with a value of 0.773. The highest accuracy was produced by scenarios 3 & 5 with a value of 0.824 and recall for scenario 5 with a value of 0.687. Both scenarios have similarities, namely they only use pre-processing stemming techniques and do not use stop words, the only difference is in the use of emoji processing. It can be interpreted that scenario 3 and scenario 5 are the models that most often predict results correctly because they produce the highest accuracy values.

Meanwhile, the model that most often predicts the positive class correctly is scenario 3 because it produces the highest precision, thereby reducing FP errors (false positives). Furthermore, the model that most often correctly identifies TPs (true positives) from all actual positive samples in the data set is scenario 5 because it has the highest recall. Finally, the models that best balance recall and precision are scenarios 3 and 5. However, please note that accuracy metrics cannot be used as a reference for determining model quality in this case because there is an imbalanced data problem, specifically in the activity labels. So, you can see the precision and recall results. In the original data, imbalanced data makes the classification biased and it is difficult to predict minority data, namely negative class (2) in the activity label. Apart from that, there is one class in the accessibility aspect that is in the minority, namely the positive class (1).

To find out how influential little data is in the negative class (2) on the activity label and the positive class (1) on the accessibility label, here are the results of the first test based on class.

Table 2. Performance Evaluation Results for Accessibility and Activity Aspects - First Test

Scenario	Accessibility Aspect			Activity Aspect		
	Positive Class			Negative Class		
	Precision	Recall	F-1 Score	Precision	Recall	F-1 Score
Scenario 1	0,735	**0,217**	**0,335**	**0,500**	0,061	0,108
Scenario 2	0,674	0,175	0,278	0,462	0,061	0,107
Scenario 3	**0,756**	0,205	0,322	0,429	0,061	0,106
Scenario 4	0,659	0,175	0,276	0,417	0,051	0,090
Scenario 5	**0,756**	0,205	0,322	0,438	**0,071**	**0,122**

Based on the table above, the recall numbers and f-1 scores are very low for both labels. Therefore, it is necessary to oversampling the data so that the f-1 score is not too low. This low recall score is because the positive class in the accessibility aspect and the negative class in the activity aspect have a small number compared to other classes or can be called the minority class. It should also be noted that class 0 does not need to be a priority because the results given are already quite good. Therefore, you only need to pay attention to the positive class (1) on the accessibility aspect and the negative class (2) on activities.

The final analysis is to check if the model used is classified as overfitting. The cross validation used is repeated k-fold with 3 repetitions and 10 folds. The following are the results of the cross validation which have been averaged and compared with the averaged accuracy values.

Table 3. Comparison of Average Cross Validation and Accuracy

Scenario	Cross Validation	Accuracy	Difference
Scenario 1	0,812	0,819	0,007
Scenario 2	0,812	0,814	0,003
Scenario 3	0,813	0,824	0,011
Scenario 4	0,811	0,815	0,004
Scenario 5	0,813	0,824	0,012

Based on a comparison of the average cross validation results and the accuracy of each scenario in the first test, the results are not much different from each other. The highest difference is in scenario 5, namely a difference of 0.012. However, this difference is not so significant that it can be concluded that in the first test, there was no overfitting in the model.

3.2 Second Testing Scenario

The second test scenario in this research carried out experiments based on training data that was oversampled using the ROS (random oversampling) method on activity labels and a combination of stop word pre-processing (SR), stemming (SM), and emoji processing (EP). The stages carried out are split data with a ratio of 70% training data and 30% testing data, converting tokens into vectors (TF-IDF), preparing test data which is oversampled, oversampling the ROS method on activity label training data, multiclass & multilabel modeling, model evaluation, and data validation using repeated K-Fold cross validation. Oversampling is carried out on just one aspect, namely multiclass & multilabel, so it is possible if it is only one aspect.

Table 4. Second Testing Scenarios

Scenario	Dataset		Pre-Processing		
	Original	ROS	SR	SM	EP
Scenario 6	×	✓	✓	✓	✓
Scenario 7	×	✓	✓	×	✓
Scenario 8	×	✓	×	✓	✓
Scenario 9	×	✓	✓	×	×
Scenario 10	×	✓	×	✓	×

From these scenarios, different evaluation values are produced. The following are the results of the evaluation.

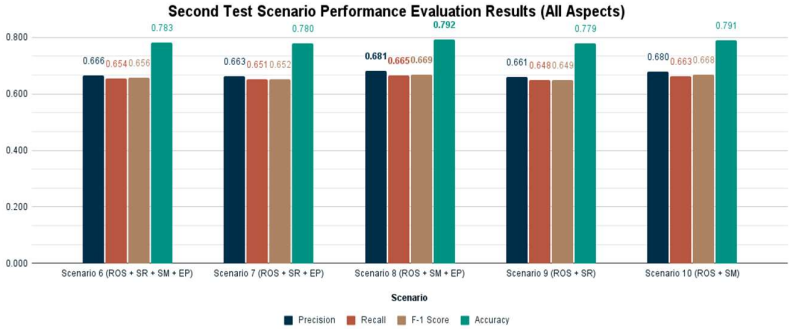

Fig. 3. Second Test Scenario Performance Evaluation Results

The table above shows the performance evaluation results based on the average of the three aspects. Precision with a value of 0.681, recall with a value of 0.665, f-1 score with a value of 0.669, and accuracy with a value of 0.792 are the highest values achieved by scenario 8. This scenario has similarities with the first test which uses pre-processing stemming (SM) and emoji techniques. processing (EP). It can be concluded that scenario 8 is the best model. This model most often correctly predicts outcomes (high accuracy), most often correctly predicts the positive class (high precision), and most often correctly identifies TPs (true positives) from all actual positive samples in the data set (high recall). When compared with the results in the first test, the evaluation value in the second test is lower. However, in the second test, the test data was oversampled. So, there is no longer any imbalance in data in minority classes. In this test, accuracy can be used as a valid evaluation metric, as can precision and recall.

Next, to find out whether random oversampling influences the results of the performance evaluation of these two aspects in the minority class, the following are the results of the performance evaluation in the accessibility and activity aspects in the first test based on the minority class.

Table 5. Performance Evaluation Results for Accessibility and Activity Aspects

Scenario	Accessibility Aspect			Activity Aspect		
	Positive Class			Negative Class		
	Precision	Recall	F-1 Score	Precision	Recall	F-1 Score
Scenario 6	0,528	**0,229**	**0,319**	**0,153**	**0,616**	**0,245**
Scenario 7	0,571	0,193	0,288	0,142	0,566	0,227
Scenario 8	**0,596**	0,205	0,305	0,152	0,545	0,237
Scenario 9	0,552	0,193	0,286	0,140	0,556	0,223
Scenario 10	0,593	0,211	0,311	**0,153**	0,556	0,240

From the table above, you can see a very significant increase in the activity aspect, especially in recall and f-1 score compared to the first test. If you take the highest recall in the activity aspect, namely scenario 6, the value reaches 0.616, very different from the first test in scenario 5 with the highest value of 0.071. Likewise with the accessibility aspect with the highest recall value is 0.229 in scenario 6. Meanwhile, for f-1 the highest score in the activity aspect is in scenario 6 with a value of 0.245. When compared with the highest f-1 score in the first test, it only reached a value of 0.122 in scenario 5. The F-1 score for accessibility decreased to the highest result. Apart from that, in the second test there was a decrease in precision from the first test. However, because the focus is on the f-1 score to get a balance between precision and recall, the second test is still better than the first test, especially on activity labels where the previous score was very low.

Cross validation was also carried out in the second test to check whether the model used was classified as overfitting. The method used is the same as before, namely repeated k-fold with repetition 3 times and 10 folds. The following are the results of cross validation in the second test.

Table 6. Comparison of Average Cross Validation and Accuracy

Scenario	Cross Validation	Accuracy	Difference
Scenario 6	0,841	0,783	0,058
Scenario 7	0,835	0, 780	0, 055
Scenario 8	0,843	0, 792	0, 051
Scenario 9	0,835	0, 779	0, 056
Scenario 10	0,846	0, 791	0, 055

Almost the same as the first test, the results are not much different and are still in the range of 0.050 to 0.058. The highest difference is in scenario 6, namely 0.058. However, this not so big difference can be categorized as a model that is not overfitting.

4 Conclusion

The best combination of pre-processing techniques to carry out aspect-based sentiment analysis using the Multinomial Logistic Regression method based on the two test scenarios that have been carried out is pre-processing stemming (SM) and emoji processing (EP).

The influence of the oversampling method on the performance of aspect-based sentiment analysis using the Multinomial Logistic Regression method is very good for increasing the results of model performance evaluation values. This is proven by the increase in the f-1 score in the minority class in the aspects of accessibility and activities whose data is classified as imbalance data.

In the first test with an imbalanced dataset, there was a positive class (1) in the accessibility aspect and a negative class (2) in the activity aspect which had a small amount of data, causing a low f-1 score. In the first test, the highest f-1 score in the

positive class accessibility aspect (1) was 0.335 in scenario 1 (OG + SM + SR + EP), while in the negative class activity aspect (2) it was 0.122 in scenario 5 (OG + SM). Continuing in the second test by oversampling the ROS method, there was an increase in the f-1 score in the activity aspect of the negative class (2) from a value of 0.122 to 0.245 in scenario 6. Meanwhile, there was a decrease in the f-1 score in the accessibility aspect of the positive class (1) from 0.335 to 0.319 in scenario 6.

The effect of the oversampling method in producing the best performance is very good because it is proven to increase the model performance evaluation value, especially in the minority class. If West Java Tourism and Culture Department wants to include new reviews and finds imbalanced data, it can carry out oversampling using the ROS method.

This research can be used by West Java Tourism and Culture Department to implement aspect prediction with sentiment based on new review data entered by using the best model (with oversampling, stemming and EP pre-processing). After doing prediction on new reviews, West Java Tourism and Culture Department can analyze the result and decide what aspect they need to focus on in order to improve the quality of natural tourist attractions by presenting the result to managing party of each natural tourist attractions. This will help with effectiveness of analyzing data than doing it in traditional way, which is labelling every review one by one.

The advice for future research is to add another data collection sources from other platforms, for example TripAdvisor. The data used can be combined from TripAdvisor and Google Maps to see a comparison of data sources. This will add variety to the review and deeper analysis. Also, when imbalanced data occurs, researcher can look for other options besides oversampling, such as adding review data and re-labelling to add classes that have a small amount of data. Apart from that, researchers can perform undersampling if the dataset requires this method.

5 References

1. E. Elistia, "Perkembangan dan Dampak Pariwisata di Indonesia Masa Pandemi Covid-19," in Konferensi Nasional Ekonomi Manajemen dan Akuntansi (KNEMA), 2020.
2. D. Aldo, A. Syawitri, A. D. and K. Samosir, Data Mining, Insan Cendekia Mandiri, 2021.
3. D. Anand and D. Naorem, "Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering," in Intelligent Human Computer Interaction, 2016.
4. C. Hua, Y.-J. Choi and Q. Shi, Companion to BER 642: Advanced Regression Methods, Bookdown, 2021.
5. Team Ravedata, "Maths Behind ML- Multinomial Logistic Regression," 31 March 2020. [Online]. Available: https://ravedata.in/machine-learning/maths-multinomial-logistic-regression/.
6. S. Polamuri, "How Multinomial Logistic Regression Model Works In Machine Learning," 14 March 2017. [Online]. Available: https://dataaspirant.com/multinomial-logistic-regression-model-works-machine-learning/.
7. D. Arianto and I. Budi, "Analisis Sentimen Berbasis Aspek Dan Pemodelan Topik Pada Destinasi Pariwisata Berdasarkan Ulasan Pengguna Google Maps Dan Tripadvisor: Studi Kasus Candi Borobudur Dan Candi Prambanan," Depok, 2021.