Aspect-based Sentiment Analysis on Indonesia's Tourism Destinations Based on Google Maps User Code-Mixed Reviews (Study Case: Borobudur and Prambanan Temples)

Dian Arianto
Faculty of Computer Science
Universitas Indonesia
Indonesia
dian.arianto@ui.ac.id

Indra Budi
Faculty of Computer Science
Universitas Indonesia
Indonesia
indra@cs.ui.ac.id

Abstract

In this paper, we conducted an aspect-based sentiment analysis using Google Maps user reviews of Indonesia's tourism destinations which are Borobudur and Prambanan Temple. The aspects we used are *Attractions*, *Amenities*, *Accessibility*, *Image*, *Price* and *Human Resources*. The dataset obtained is in code-mixed language. We applied five machine learning algorithms which are Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree (ET). The evaluation performed by making eight scenarios which are the combination of stopwords removal (SR), stemming (SM), emoji processing (EP), our own stopwords dictionary (OSD), and Suciati and Budi stopwords dictionary (SSD). The model performance was measured by ten folds cross-validation. The results suggest that SM without SR, and with or without EP, SSD, and OSD did not result in a significant difference for the F1-scores. However, the combination of SM and EP, and the combination of SR, EP, and SSD did improve the performance of models for classifying sentiments.

1 Introduction

Indonesia has many tourism destinations that can be visited by both domestic tourists and foreign tourists. There is an increasing number of tourists visiting Indonesia each year. Domestic tourists and foreign tourists continue to arrive every year to visit tourist attractions in Indonesia. Foreign tourists experienced a significant increase to visit Indonesia from 2009 to 2018. The average growth of foreign tourists in the 2009-2013 period was 9% per year and rose to 14% per year in the 2014-2018 period (Widowati, 2019).

Indonesia in 2015 set a program of 10 Priority Tourism Destinations or called "10 New Bali" to promote Indonesian tourism and increase foreign tourist visits. One of the 10 New Bali is Borobudur Temple, Central Java. The tourism destination of temples in Indonesia besides Borobudur Temple which is quite famous for tourists is Prambanan Temple.

Every year, reviews on online platforms experience significant competition. Since 2015, Google has experienced a significant increase in the number of reviews shared by internet users compared to other platforms such as Facebook, Yelp[1], TripAdvisor or Foursquare (Murphy, 2018). The increase in the number of reviews on the Google platform is supported by Google's program called Google Local Guides[2]. This program was originally launched in 2015 as a way to deal with Yelp Elites (Yelp contributors), which allows the most active Google Maps contributors to get rewards. In 2016, this program had 5 million contributors globally and increased every year to 120 million globally in 2019 (Sterling, 2019).

In Indonesia, research on text mining in the field of tourism had been done before to obtain

[1] https://www.yelp.com/

[2] https://maps.google.com/localguides

important information from tourism destinations. Prameswari et al. (2017) conducted a sentiment analysis for hotels in Bali and Labuan Bajo by using five aspects of the hotel domain, those are: *Accessibility*, *Activities & Entertainment*, *Food & Beverage Operations*, *Human Resources*, and *Physical Environment*. Herry et al. (2019) conducted sentiment analysis to obtain important information from user reviews on TripAdvisor on 10 Indonesian tourist destinations.

While research in the field of tourism using Google Maps reviews is still very limited both in Indonesia and globally. Munawir et al. (2019) conducted a study using Google Maps reviews to get a visitor's perspective on parks in the city of Bandung in Indonesia. They used TF-IDF to examine the term of reviews that had important value for the visitor.

In this paper, we conducted an aspect-based sentiment analysis using Google Maps user reviews of Indonesia's tourism destinations which are Borobudur Temple and Prambanan Temple. The aspects we used are *Attractions*, *Amenities*, *Accessibility*, *Image*, *Price* and *Human Resources* (World Tourism Organization, 2007). The dataset obtained from Google Maps is in code-mixed language (Indonesian and English).

The rest of this paper is arranged as follows: in Section 2, we review the related works with our study. Then we describe the research methodology applied in this work in section 3. In section 4, we discuss the results and analysis of the experiment. Finally, in section 5, we conclude our results and define future work.

2 Related Work

In the field of Natural Language Processing (NLP), research on sentiment analysis has been carried out to extract information from the opinions of internet users in several domains. In the restaurant domain, Suciati and Budi (2019) conducted a sentiment analysis of internet user reviews for several restaurants in Indonesia. They used several machine learning methods such as Random Forest (RT), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree (ET) to classify aspects of internet user reviews in code-mixed languages (Indonesian-English). In addition, they used a combination of stemming and removing stopwords to discover what was the best scenario that was able to increase the performance of the models. They also used their own stopwords instead of using common Indonesian and English stopwords to gain better performance of the models.

In the tourism domain, Prameswari et al. (2017) conducted an aspect-based sentiment analysis for online reviews of TripAdvisor users in hotels in Bali using the Recursive Neural Tensor Network (RNTN) algorithm at the sentence level. They used eight aspects in their research. With an average accuracy of 85%, the proposed algorithm managed to predict well in classifying sentiments from words or aspects. While the average F1-score was 77% with the highest F1-score of positive sentiment was 90%.

Other research related to sentiment analysis in tourism is carried out by Kuhamanee et al. (2017). They conducted research to obtain information on sentiments from foreign tourists who aim to improve and develop the tourism industry in Bangkok. The dataset used was 10,000 tweets from the Twitter platform in 2017. The methods used were Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), and Artificial Neural Networks (ANN) using Rapidminer Studio 7.4. They gained insight that most tourists visiting Bangkok were for the purpose of nightlife activities, Thai culture, and shopping with a percentage of 65.54%, 16.07%, and 16.07%.

Kurniawan et al. (2019) conducted hierarchical sentiment analysis on hotel reviews taken from the Traveloka[3] website using Naïve Bayes. The data used were 1,720 reviews consisting of 430 positive reviews, 430 negative reviews, and 860 neutral reviews. The results of their study indicated that the use of hierarchical classifications for sentiment analysis was able to increase the average performance of the classification model by 2.3%.

Souza et al. (2018) performed a sentiment analysis on hotel reviews using CNN and compared it with other methods such as Lemmatization, Polarity inversion, and Laplace smoothing that had previously been done with the same dataset. The dataset used is 69,075 reviews from TripAdvisor about hotels located in the city of Rio De Janeiro, written in Brazilian Portuguese. The results showed that using only positive and negative reviews as in previous studies containing

[3] https://www.traveloka.com/

the same number of reviews in each class, the resulting accuracy of the model was 95.74%.

Khine and Aung (2019) conducted a sentiment analysis using the SenticNet MA-LSTM deep learning approach for restaurant reviews. The dataset consists of 20,000 sentences from TripAdvisor, which were categorised as positive, negative, and neutral sentences. The results of the experiments show that SenticNet MA-LSTM achieved the best results with an accuracy of 87.2% compared to the ordinary LSTM, which is equal to 82%.

From these studies, several approaches can be used to conduct sentiment analysis using either machine learning or deep learning. Machine learning is generally used for medium-sized data, and deep learning can be used for data with fairly large size.

3 Research Methodology

In this section, the research methodology that applied in this study will be described. This work consists of five steps, as shown in Figure 1.

Firstly, we collected the data from the website a using crawling tool. The second step was applying few text preprocessing techniques, the third one was doing feature extraction, the fourth was doing an experiment with machine learning models, and the last one, we evaluated the models.

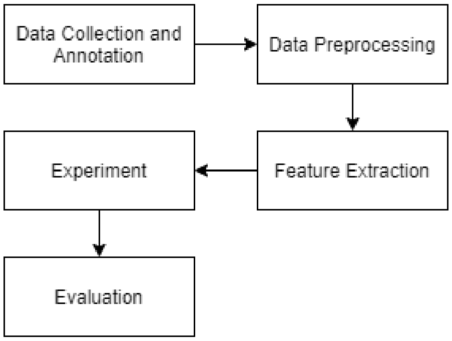

Figure 1. Research Methodology

3.1 Data Collection and Annotation

In the data collection step, we collected 5,592 reviews from Borobudur and Prambanan Temple by crawling them using BotSol's Google Maps Review Crawlers[4]. We collected 2,796 reviews from both Borobudur and Prambanan Temple. The reviews are in Indonesian, English, and mixed (Indonesian and English). The example of the data that we retrieved are as follows:

- Review in Indonesian: "*Disini adalah salah satu wisata di 7 keajaiban dunia, bahkan taman wisata borobudur ini sudah termasuk ke kawasan wisata 10 new bali, jadi semakin terkenal di kancah international, dan tiket masuknya ini 50.000 jika ingin sekaligus ke candi Prambanan bisa dengan harga tiketnya 75.000.*" (Here is one of the 7 wonders of the world tourist attractions, even Borobudur tourism park is already included in the 10 new Bali tourist area, so it is increasingly famous in the international arena, and the entry ticket is 50,000 if you want to go to Prambanan temple at the same time the ticket price is 75,000.)
- Review in English: "*The temple itself is beautiful but everything else about this place is terrible. The prices are extremely high, especially for sunset and sunrise, absolute rip off.*"
- Review in mixed Indonesian-English: "*A very nice temple. Sangat edukatif, penjelasan mengenai sejarah tiap candi dan kejadian terekam dengan baik. Harga yang memadai untuk pemandangan yang indah dan menyenangkan. Highly recommended for local and international tourist!*" (A very nice temple. Very educative, an explanation of the history of each temple and events are well recorded. Adequate prices for beautiful, pleasant scenery. Highly recommended for local and international tourists!)

After retrieving the data, we then annotated them manually with the help of volunteers. Each data annotated by three annotators and we used majority voting to decide the final label of each review.

The aspects that we used in this research are *attractions*, *amenities*, *accessibility*, *image*, *price*, and *human resources* as the important aspects of tourism recommended by World Tourism Organization (2007). Then we divided the sentiment polarities into "positive", "negative", and "neutral" based on Suciati and Budi (2019). In addition, we added the "none" sentiment polarity to those reviews that did not contain six aspects of tourism. Aspects are classified as "positive" if the review mentioned positive words or phrases such as "*beautiful*", "*nice park*", "*cukup bersih*" (clean enough), "*amazingly cheap*", "*bagus*" (good), etc. We classified the negative aspects if there are negatives words or phrases such as "*unattractive*

[4] https://www.botsol.com/Products/GmapsReviewCrawler

spot", "*nothing special*", "*expensive*", "*toilet gak ada air*" (toilet no water), "*kurang petugas*" (inadequate staff), etc. For the "neutral" aspects, we classified the reviews that are not both positive and negative. For example, the reviews that contain phrase like "*standard entry ticket*", "*part of the seven wonders of the world*", "*toilets are fairly up to standard*", "*one of unsesco world heritage*", etc. For the "none" aspects, we classified the reviews that do not contain the six aspects of tourism.

The following example shows how we annotated the data:

Review: "Borobodur is also known as Temple on the Hill. Breakfast and torch light included. Ticket was IDR 475,000 for the sunrise package. They'll give you a scarf as a souvenir when you return the torchlight. There were lots of people during the sunrise and even more after 6am, so it's still best to come at sunrise as there are lesser people. Nice place to see!" {*"attractions": "positive"; amenities: "positive"; "accessibility": "none"; "image": "positive"; "price": "neutral"; "human resources": "none"*}

After we annotated the data, we filtered the reviews and we only used 4,395 reviews of the data we obtained in this research.

3.2 Data Preprocessing

After we collected and annotated the data, we then performed several data preprocessing techniques in text mining to clean the data. These are the techniques we used:

1) *Emoji Processing*: at the first step, we processed the emojis that appeared in the text to string that represents its meaning which is in Indonesian. We used the top 10 positive, negative, and neutral representation of emojis based on Novak et al. (2015). We changed 10 positives emojis which are: "😂, ❤️, ♥, 😍, 😘, 😊, 👌, 💕, 👏, and 😁" to "*positif*" (positive); 10 negatives emojis which are: "😭, 😩, 😒, 😔, 😡, 😴, 🔫, 😞, 😪, and 😫" to "*negatif*" (negative); and 10 neutral emojis which are: "☯️, ✨, ★, █, 🔥, ♫, �, ©, 👀, and 🐓" to "*netral*" (neutral).

2) *Case Folding*: Next, we performed case folding to make the words in the text become lower case. For example, "Very beautyfulll" becomes "very beautyfulll".

3) *Remove Username, Numbers, and Punctuation*: the next step, we removed the usernames, numbers, and punctuation occurred in the data. For instance, "Amazing place open from 06.00 am until 17.00pm #visualine my instagram @antoniusandryano if you like please follow" converted to "Amazing place open from am until pm visualine my instagram if you like please follow".

4) *Text Normalization (part 1):* in this step, we normalised the spelling and abbreviation of the words into the formal words in Indonesian and English by using the modified dictionary from Suciati and Budi (2019). For example, "*wisata budaya yg sgt bagus*" changed to "*wisata budaya yang sangat bagus*" (the cultural tourist attraction which is very good).

5) *Stopwords Removal*: for this step, we used two dictionaries which are the stopwords dictionary used in Suciati and Budi (2019) and stopwords that we built and combined based on Tala (2003) for Indonesian and Countwordsfree Tools[5] for English. We used two dictionaries to investigate how the stopwords used in Suciati and Budi (2019) affect the performance of the models. Since they did not remove words such as "not" or "tidak" (not) in their dictionary in order to avoid missing information about the negation of positive words.

6) *Removing Duplicate Characters and Whitespace*: the next step, we removed the duplicate character occurred in text such as "beautifullll" changed to "beautiful". We also removed the whitespace in the text.

7) *Text Normalization (part 2)*: next, we performed normalization again to correct the spelling after we did duplicate character removal. We did this because there are some words result of removing duplicate characters like "peningalan" should be "peninggalan" (heritage) that can affect the stemming step after this step.

8) *Stemming*: in the last step, we performed stemming functions from libraries. We used two libraries, which are Snowball Stemmer from NLTK[6] library for English and Sastrawi library[7]

[5] https://countwordsfree.com/stopwords

[6] https://www.nltk.org/

[7] https://github.com/har07/PySastrawi

for Indonesian. For instance, the Indonesian review is "*tamannya tertata rapi*" (the park is neat) converted to "*taman tata rapi*" (neat park layout). For English review "*it's definitely an human patrimony, but it could have some more explanation (like small cards)*" converted to "*it definit an human patrimony, but it could have some more explan (like small cards)*"

3.3 Feature Extraction

In this step, we extracted the feature that would be used in the classification models. We used bigram term for the feature and its vector were extracted by vectorizing the words in the reviews. In addition, we also used the combination of stemming, stopwords removal, the use of our mixed stopwords dictionary, Suciati and Budi (2019) stopwords dictionary, and emoji processing steps to see whether they can increase the performance of the models.

3.4 Experiment

In our experiment, we performed eight scenarios and applied five machine learning algorithms that were used in Suciati and Budi (2019). Then, we measured and compared their performance using their F1-scores.

1) *Experiment Scenarios*: we examined eight scenarios for our experiments in this research. The objective is to see how the stopwords removal, stemming, the use of stopwords dictionaries, and emoji processing can affect the performances of the machine learning models when they applied to our dataset. In the first scenario, we built machine learning models by applying stopwords removal, emoji processing, and the use our stopwords dictionary, but we did not use stemming and the use of Suciati and Budi (2019) stopwords dictionary. For the second scenario, we applied stopwords removal, emoji processing, and the use of Suciati and Budi (2019) stopwords dictionary, but we did not use stemming and the use of our stopwords dictionary. The rest of scenarios can be seen in Table 1 with *SR = Stopwords Removal; SM = Stemming; EP = Emoji Processing; OSD = Our Stopwords Dictionary; SSD = Suciati and Budi Stopwords Dictionary*.

2) Dataset: we used all annotated data (4,395 reviews) for all scenarios. From Figure 2, we can see that "none" polarity had the highest number in *Amenities*, *Accessibility*, *Price*, and *Human Resources* (HR) aspects, while "positive" polarity had the highest number in *Attractions* and *Image* aspects. Positive reviews for *Attractions* aspect appeared almost in all reviews and there were only 855 reviews that were not positive. In *Image* aspect, the "positive" and "none" polarities had slightly different number of reviews. In contrast, *Amenities*, *Accessibility*, *Price*, and *HR* aspects had more than 3,500 reviews that had no polarity or "none".

Scenarios	**SR**	**SM**	**EP**	**OSD**	**SSD**
Scenario 1	✓	×	✓	✓	×
Scenario 2	✓	×	✓	×	✓
Scenario 3	✓	×	×	✓	×
Scenario 4	✓	×	×	×	✓
Scenario 5	×	✓	✓	✓	×
Scenario 6	×	✓	✓	×	✓
Scenario 7	×	✓	×	✓	×
Scenario 8	×	✓	×	×	✓

Table 1. Experiment Scenarios

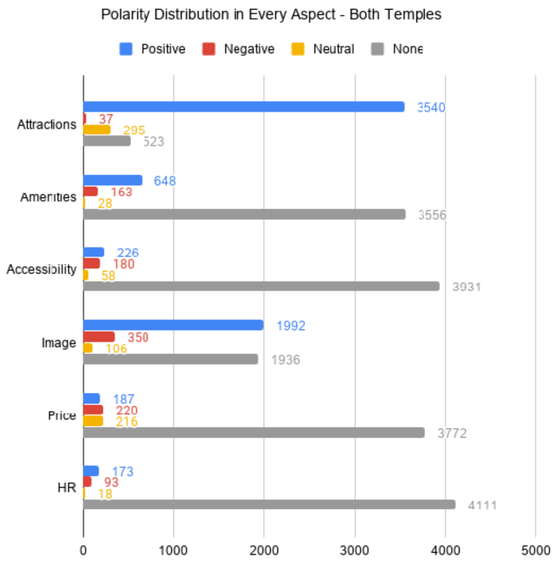

Figure 2. Polarity Distribution in Every Aspect

3) Classification Algorithms: for the experiments, we used five algorithms that were proposed by Suciati and Budi (2019) with different scenarios from the scenarios they used in their work. The machine learning algorithms we used are Decision Tree (DT), Random Forest (RF),

Logistic Regression (LR), Extra Tree (ET) or extremely Randomized Tree, and Multinomial Naïve Bayes (NB) classifiers. After we selected the machine learning algorithms and conducted the classification experiments, we then compared the obtained F1-scores of the models. Then, we can see the best machine learning algorithms for our dataset.

3.5 Evaluation

In this research, we used five classifiers which are NB, LR, RF, ET, and DT with cross-validation as the validation technique. After we did the experiments, we evaluated and compared the F1-scores of the machine learning models in all scenarios. In this experiment, we used ten folds for cross-validation technique. The performance of the models with the eight scenarios can be seen in Discussion section.

4 Discussion

In this section, we discussed the performance of the models in every aspect. For every aspect in Table 2 until Table 9 we use abbreviation in the table which are: *Att = Attractions; Amn = Amenities; Acc = Accessibility; Img = Image; Prc = Price; and HR= Human Resources*.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.655	0.744	0.796	0.479	0.773	0.818
LR	**0.733**	0.772	0.855	**0.516**	0.834	0.906
RF	0.729	0.763	0.855	0.455	0.841	0.905
ET	0.727	0.787	0.871	0.498	0.866	0.906
DT	0.717	**0.796**	**0.874**	0.500	**0.871**	**0.908**

Table 2. Result of First Scenario

Table 2 shows the F1-scores of the models for the first scenario which was applying SR, EP, and OSD, but without SM and SSD. From the table, we can see that LR had the highest F1-scores for Att and Img aspects which were 73.3% and 51.6% respectively. For the other aspects, it was led by DT by obtaining 79.6%, 87.4%, 87.1%, and 90.8% for Amn, Acc, Prc, and HR respectively.

Table 3 shows the F1-scores of the models for the second scenario which was applying SR, EP, and SSD, but without SM and OSD. From the table, we can see that LR had the highest F1-scores for Att aspects which were 73.4%. For the other aspects, it was led by DT excluding the Prc aspect that led by ET. Compared to Table 2, only NB had higher result in every aspect than other algorithms.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.672	0.748	0.801	0.490	0.782	0.826
LR	**0.734**	0.776	0.855	0.527	0.835	0.906
RF	0.733	0.747	0.860	0.477	0.839	0.906
ET	0.727	0.780	0.875	0.520	**0.867**	0.907
DT	0.720	**0.789**	**0.881**	**0.530**	0.866	**0.911**

Table 3. Result of Second Scenario

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.654	0.744	0.797	0.474	0.773	0.816
LR	**0.733**	0.772	0.854	**0.517**	0.834	0.906
RF	0.731	0.762	0.857	0.458	0.841	0.904
ET	0.724	0.785	0.871	0.502	0.865	0.906
DT	0.719	**0.794**	**0.876**	0.496	**0.872**	**0.908**

Table 4. Result of Third Scenario

Table 4 shows the F1-scores of the models for the third scenario which was applying SR and OSD, but without SM, EP, and SSD. From the table, we can see that LR had the highest F1-scores for Att and Prc aspects which were 73.3% and 51.7% respectively. For the other aspects, it was led by DT. Compared to Table 2 and 3, it can be concluded that result of all models were lower.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.670	0.746	0.803	0.486	0.782	0.827
LR	**0.734**	0.775	0.855	0.525	0.836	0.906
RF	0.728	0.755	0.857	0.481	0.845	0.906
ET	0.725	0.780	0.875	0.519	0.865	0.908
DT	0.719	**0.791**	**0.883**	**0.527**	**0.867**	**0.912**

Table 5. Result of Fourth Scenario

Table 5 depicts F1-scores of the models for the fourth scenario which was applying SR and SSD, but without SM, EP, and OSD. From the table, we can see that LR had the highest F1-score again for Att aspect which was 73.4%. For the other aspects, it surprisingly was all led by DT.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.731	0.746	0.827	0.520	0.810	0.868
LR	**0.744**	**0.810**	0.869	0.561	0.856	0.906
RF	0.732	0.778	0.863	0.525	0.840	0.906
ET	0.731	0.805	0.881	**0.571**	**0.879**	0.910
DT	0.713	0.799	**0.885**	0.539	0.871	**0.915**

Table 6. Result of Fifth Scenario

Table 6 depicts F1-scores of the models for the fifth scenario which was applying SM, EP, and OSD, but without SR and SSD. From the table, we can see that LR had the highest F1-scores again for Att aspect which were 74.4% and for the first time for Amn was led by LR which were 81%. For the Img and Prc aspects it was led by ET, and for Acc and HR it was led by DT.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.731	0.746	0.827	0.520	0.810	0.868
LR	**0.744**	**0.810**	0.869	0.561	0.856	0.906
RF	0.732	0.778	0.863	0.525	0.840	0.906
ET	0.730	0.807	0.878	**0.571**	**0.881**	0.912
DT	0.713	0.799	**0.885**	0.539	0.871	**0.915**

Table 7. Result of Sixth Scenario

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.731	0.746	0.828	0.519	0.810	0.868
LR	**0.742**	**0.810**	0.869	0.560	0.856	0.906
RF	0.734	0.773	0.860	0.524	0.846	0.905
ET	0.731	0.803	0.879	**0.572**	**0.880**	0.913
DT	0.711	0.794	**0.882**	0.542	0.876	**0.914**

Table 8. Result of Seventh Scenario

Table 7 depicts F1-scores of the models for the sixth scenario which was applying SM, EP, and SSD, but without SR and OSD. From the table, we can see that LR had the highest same F1-scores as the fifth scenario for Att and Amn. For the other aspects, it can be seen that every aspects had the same score as the fifth scenario except for the Prc aspect that had 0.02% difference.

Table 8 depicts F1-scores of the models for the seventh scenario which was applying SM and OSD, but without SR, EP, and SSD. From the table, we can see that LR had the highest F1-scores Att and Amn which were 74.2% and 81% respectively. For the other aspects, it can be seen that ET had the highest scores for Img and Prc aspects, and DT for Acc and HR aspects.

Model	Att	Amn	Acc	Img	Prc	HR
NB	0.731	0.746	0.828	0.519	0.810	0.868
LR	**0.742**	**0.810**	0.869	0.560	0.856	0.906
RF	0.734	0.773	0.860	0.524	0.846	0.905
ET	0.728	0.805	0.883	**0.571**	**0.877**	0.912
DT	0.711	0.794	**0.882**	0.542	0.876	**0.914**

Table 9. Result of Eighth Scenario

Table 9 shows F1-scores of the models for the last scenario which was applying SM and SSD, but without SR, EP, and OSD. From the table, we can see that every aspect had the highest score with the models same as seventh scenario which the result is in Table 8. However, for ET, the models had the lower scores compared to Table 8 which had 0.01% difference for Img aspect and 0.03% for Prc aspect.

In summary, by seeing Figure 3 and Figure 4, DT was the best algorithm to predict the sentiment in almost six aspects for the first, second, third, and fourth scenarios, and the rest were LR and ET that obtained the highest F1-scores in the fifth, sixth, seventh, and eighth scenarios. It can be seen that the combination of SR, SM, EP, OSD, and SSD can affect the performance of models. The Att and Amn aspects obtained the highest scores (74.4% and 81% respectively) by LR in the fifth and sixth scenario which were applying SM, EP, and OSD without SR and SSD for the fifth scenario and applying SM, EP, and SSD without SR and OSD for the sixth scenario. Besides, Acc and HR aspects achieved the highest score by DT in the fifth and sixth scenarios as well, which were 88.5% and 91.5% respectively. It seems that SM and EP affect the models for Att, Amn, Acc, and HR aspects more than SR and EP because their scores were higher every time SM and EP used in the models. OSD and SSD did not affect the performance of models because SR was not used in the scenarios. While Img aspect obtained the highest score, which was 57.2% in the seventh scenario by ET, and Prc achieved the highest score, which was 88.1% in the sixth scenario by ET as well. It can be seen that the application of SM without EP in the seventh scenario did not result in a significant difference compared to the application of SM and EP in the fifth or sixth scenario since in the fifth and sixth scenarios Img aspect achieved 57.1% by ET.

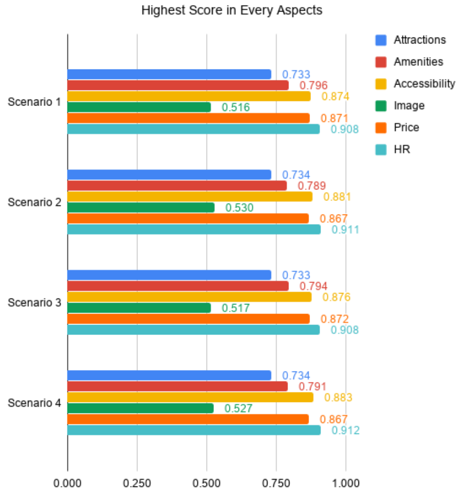

Figure 3. Highest Scores in Every Aspect

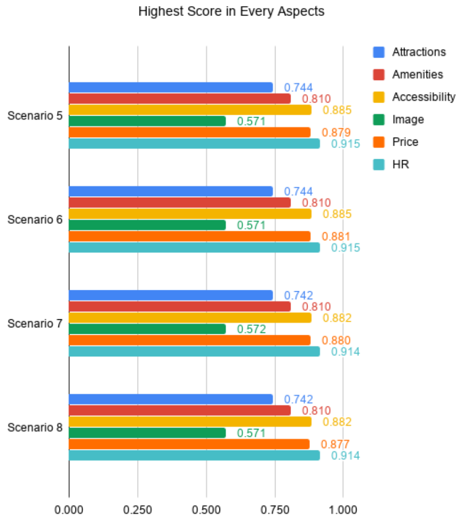

Figure 4. Highest Scores in Every Aspect

Furthermore, the F1-scores achieved by the models were various from 45.5% to 91.5%. However, the highest score obtained by Img aspect was only 57.2% while other aspects were above 70%. This was highly likely caused by the annotated data for the Img aspect had quite a significant difference annotation results between one annotator and another, leading to deletion of some data in this study and causing the models unable to predict sentiment well on the Img aspect. For example, the review was: "*Everything is good, but sometimes it gets very crowded. Avoid to go here when holidays or weekends*," and the results for the Img aspect for three annotators were "*negative, positive, and none*". So the data was removed and not included in the training dataset. This was also happened in other data that used for training dataset for Img aspect. For other aspects, the results of the data annotation were good enough so that the model managed to predict sentiment quite well by generating an F1-scores above 70%.

5 Conclusion

In this work, we have examined the performances of five machine learning algorithms and eight scenarios to classify the sentiment of Google Maps user reviews in Borobudur and Prambanan temples which is in code-mixed. The machine learning algorithms that we used are Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT), and Extra Tree (ET). The aspects are *Attractions* (Att), *Amenities* (Amn), *Accessibility* (Acc), *Image* (Img), *Price* (Prc), and *Human Resources* (HR). The evaluation performed by making eight scenarios which are the combination of stopwords removal (SR), stemming (SM), emoji processing (EP), our own stopwords dictionary (OSD), and Suciati and Budi stopwords dictionary (SSD). The model performance was measured by ten folds cross-validation, and the results show that LR achieved the highest score for Att (74.4%) and Amn (81%), DT achieved the highest score for Acc (88.5%) and HR (91.5%). While ET achieved the highest scores for Img (57.2%) and Prc (88.1%) aspects. By seeing the results, it can be concluded that SM without SR, and with or without EP, SSD, and OSD did not result in a significant difference for the F1-scores. However, the combination of SM and EP, and the combination of SR, EP, and SSD did improve the performance of models for classifying sentiments.

In this experiment, we only applied aspect-based sentiment analysis for the reviews, in the future we will conduct topic modelling to see what topics that people frequently talked in the reviews and will use deep learning for the larger dataset.

Acknowledgments

The authors would like to thank Universitas Indonesia for funding this work through International Indexed Publication (PUTI Proceedings) research grant No. NKB-870/UN2.RST/HKP.05.00/2020.

References

Herry Irawan, Gina Akmalia, R. A. M. (2019). *Mining Tourist's Perception toward Indonesia Tourism Destination Using Sentiment Analysis and Topic Modelling*. (1).

Joana Gabriela Ribeiro de Souza, A. de P. O., & Guidson Coelho de Andrade, A. M. (2018). A Deep Learning Approach for Sentiment Analysis Applied to Hotel's Reviews. In *NLDB 2018*. https://doi.org/10.1007/978-3-319-91947-8

Khine, W. L. K., & Aung, N. T. T. (2019). Applying Deep Learning Approach to Targeted Aspect-based Sentiment Analysis for Restaurant Domain. *2019 International Conference on Advanced Information Technologies, ICAIT 2019*, 206–211. https://doi.org/10.1109/AITC.2019.8920880

Kuhamanee, T., Talmongkol, N., Chaisuriyakul, K., San-Um, W., Pongpisuttinun, N., & Pongyupinpanich, S. (2017). Sentiment analysis of foreign tourists to Bangkok using data mining through online social network. *Proceedings - 2017 IEEE 15th International Conference on Industrial Informatics, INDIN 2017*, 1068–1073. https://doi.org/10.1109/INDIN.2017.8104921

Kurniawan, S., Kusumaningrum, R., & Timu, M. E. (2019). Hierarchical Sentence Sentiment Analysis of Hotel Reviews Using the Naïve Bayes Classifier. *2018 2nd International Conference on Informatics and Computational Sciences, ICICoS 2018*, 104–108. https://doi.org/10.1109/ICICOS.2018.8621748

Munawir, Koerniawan, M. D., & Dewancker, B. J. (2019). Visitor perceptions and effectiveness of place branding strategies in thematic parks in Bandung City using text mining based on google maps user reviews. *Sustainability (Switzerland)*, *11*(7). https://doi.org/10.3390/SU11072123

Murphy, R. (2018). Comparison of Local Review Sites: Which Platform is Growing the Fastest? Retrieved February 3, 2020, from https://www.brightlocal.com/research/comparison-of-local-review-sites/

Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PLoS ONE*, *10*(12), 1–22. https://doi.org/10.1371/journal.pone.0144296

Prameswari, P., Surjandari, I., & Laoh, E. (2017). Opinion mining from online reviews in Bali tourist area. *Proceeding - 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017*, *2018-Janua*, 226–230. https://doi.org/10.1109/ICSITech.2017.8257115

Prameswari, P., Zulkarnain, Surjandari, I., & Laoh, E. (2017). Mining online reviews in Indonesia's priority tourist destinations using sentiment analysis and text summarization approach. *Proceedings - 2017 IEEE 8th International Conference on Awareness Science and Technology, ICAST 2017*, *2018-Janua*(iCAST), 121–126. https://doi.org/10.1109/ICAwST.2017.8256429

Sterling, G. (2019). Google Maps becomes more 'social' with Local Guides follow feature. Retrieved June 27, 2020, from https://searchengineland.com/google-maps-becomes-more-social-with-local-guides-follow-feature-325322

Suciati, A., & Budi, I. (2019). Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia. *Proceedings of the 2019 International Conference on Asian Language Processing, IALP 2019*, 59–64. https://doi.org/10.1109/IALP48816.2019.9037689

Tala, F. Z. (2003). A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. *M.Sc. Thesis, Appendix D*, *pp*, 39–46.

Widowati, H. (2019). 5 Tahun Terakhir, Rerata Pertumbuhan Kunjungan Wisatawan Mancanegara 14%. Retrieved February 3, 2020, from Https://Databoks.Katadata.Co.Id website: https://databoks.katadata.co.id/datapublish/2019/07/17/5-tahun-terakhir-rerata-pertumbuhan-kunjungan-wisawatan-mancanegara-14

World Tourism Organization. (2007). A Practical Guide to Tourism Destination Management. In *A Practical Guide to Tourism Destination Management*. https://doi.org/10.18111/9789284412433