JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

Multi-Aspect Sentiment Analysis Hotel Review Using RF, SVM, and Naïve Bayes based Hybrid Classifier

I Putu Ananda Miarta Utama, Sri Suryani Prasetyowati, Yuliant Sibaroni*

School of Computing, Informatics Study Program, Telkom University, Bandung, Indonesia
Email: [1] anandamiarta@student.telkomuniversity.ac.id, [2] srisuryani@telkomuniversity.ac.id,
[3,*]yuliant@telkomuniversity.ac.id
Correspondence Author Email: yuliant@telkomuniversity.ac.id

Abstrak−Dalam sektor pariwisata hotel tentunya tidak lepas dari peran media sosial karena wisatawan cenderung berbagi pengalaman tentang layanan dan produk yang ditawarkan oleh sebuah hotel, seperti menambahkan gambar, review, dan rating yang nantinya akan berguna sebagai referensi bagi wisatawan turis lain, misalnya di media online TripAdvisor. Namun, banyaknya pengalaman wisatawan mengenai hotel membuat sebagian orang bingung menentukan hotel yang tepat untuk dikunjungi. Oleh karena itu, dalam penelitian ini dilakukan analisis review berbasis aspek terhadap hotel, yang akan memudahkan wisatawan dalam menentukan hotel yang tepat berdasarkan aspek kategori terbaik. Dataset yang digunakan adalah dataset Ulasan Hotel TripAdvisor yang sudah ada di situs web Kaggle. Serta memiliki lima aspek yaitu Kamar, Lokasi, Kebersihan, Pendaftaran, dan Pelayanan. Analisis review dilakukan ke dalam kategori positif dan negatif menggunakan metode berbasis Random Forest, Support Vector Machine, dan Naive Bayes Hybrid Classifier untuk mengatasi masalah ini. Pada penelitian ini metode Hybrid Classifier mendapatkan akurasi yang lebih baik dibandingkan dengan klasifikasi yang menggunakan satu algoritma pada data multi aspek yaitu Hybrid Classifier mendapatkan rata rata akurasi 84%, Naïve Bayes mendapatkan rata rata akurasi 82.4%, Random Forest mendapatkan rata rata akurasi 82.2%, dan Support Vector Machine mendapatkan rata rata akurasi 81% .

Kata Kunci: Hotel, *Random Forest*, *Support Vector Machine*, *Hybrid Classifier*, Multi-aspek Analisis Sentiment, *Naïve Bayes*

Abstract−In the hotel tourism sector, of course, it cannot be separated from the role of social media because tourists tend to share experiences about services and products offered by a hotel, such as adding pictures, reviews, and ratings which will be helpful as references for other tourists, for example on the media online TripAdvisor. However, tourists' many experiences regarding a hotel make some people feel confused in determining the right hotel to visit. Therefore, in this study, an aspect-based analysis of reviews on hotels is carried out, which will make it easier for tourists to determine the right hotel based on the best category aspects. The dataset used is the TripAdvisor Hotel Reviews dataset which is already on the Kaggle website. And has five aspects, namely Room, Location, Cleanliness, Registration, and Service. A review analysis was carried out into positive and negative categories using the Random Forest, Support Vector Machine, and Naive Bayes Hybrid Classifier-based methods to solve this problem. In this study the Hybrid Classifier method gets better accuracy than the classification using one algorithm on multi-aspect data, namely the Hybrid Classifier gets an average accuracy of 84%, Naïve Bayes gets an average accuracy of 82.4%, Random Forest gets an average accuracy of 82.2%, and Support Vector Machine get an average accuracy of 81%.

Keywords: Hotel; Random Forest; Support Vector Machine; Hybrid Classifier; Multi-aspect Sentiment Analysis; Naïve Bayes

1. INTRODUCTION

In the current development of tourism, it cannot be separated from online media because it is beneficial for tourists in finding and referencing the tourism they want, one of which is tourism in hotel destinations. Hotels are essential in the world of tourism because every tourist uses hotels to rest after traveling. Travelers tend to share experiences or express themselves about a hotel's services and products on online media such as TripAdvisor. TripAdvisor is one of the online media or travel websites that can help tourists book hotels quickly. Tourists can also access and add images, reviews, and ratings of a hotel they have visited, which will later be helpful as a reference for other tourists.

Based on research [1], found that approximately 89% of global travelers and 64% of international hoteliers believe that online hotel reviews affect hotel bookings, based on the data that nearly 95% of travelers make a booking decision by first reading hotel reviews online, and this is one of the most important influences of a traveler's decision to choose a hotel. Online ratings and customer reviews can help customers make decisions, but studies provide a better insight into the hotel [2]. There are many reviews about too available hotels or do not contain certain aspects on the online media or the TripAdvisor website. It makes it difficult for other travelers to decide to book the best hotel.

Previous research [3] resulted in an average F1 score of 91.4% on hotel review sentiment analysis, but an aspect-based analysis of hotel reviews has not been carried out. By using multi-aspect sentiment analysis, users can more easily determine something according to the aspect category that the user sees. There are five aspects of the hotel category in multi-aspect sentiment analysis research: location, food, service, comfort, and cleanliness, which get the highest accuracy results reaching 93% in these five aspects [1]. In contrast to the research [4], five elements of the category used: service, room, location, price, and food, which results in Precision, Recall, and F1-Score almost reaches above 70% in all aspects except the food aspect. In the study [5], the F1- Measure result was

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

0.840 in 5 hotel categories: location, food, service, comfort, and cleanliness. Sentiment classification yields F1-Size 0.946.

Referring to previous research that used 1000 data from the Amazon site with the Random Forest method and the Support Vector Machine based on the Hybrid Approach in sentiment analysis, the accuracy of 81% Random Forest classification, 82.4% Support Vector Machine and 83 Random Forest Supporting Vector Machines (Hybrid) was obtained. 83.4% [6]. Using Hybrid can increase the efficiency of sentiment analysis and get more accurate results than other algorithms. The study [7], which used the Support Vector Machine algorithm and Hybrid-based Artificial Neural Networks, obtained high accuracy results, namely 97.4% in film review data. Research using the Hybrid method also obtained 93% accuracy results using the movie review dataset, which combines two different machine learning algorithms, namely Naïve Bayes and Artificial Genetic Algorithm [8]. However, these three studies have not used multi-aspect sentiment analysis. It is an opportunity for further research. Naïve Bayes Classifier will add in the method study.

Based on these problems, this study conducted a Multi-Aspect Analysis of Hotel Review Sentiments Using Random Forest, Support Vector Machine, and Naïve Bayes based on Hybrid Classifier. Sentiment analysis will reveal both positive and negative reviews so that users can quickly determine the best judgments. The data set used 1,500 datasets for Hotel Reviews from TripAdvisor in English, which are already on the Kaggle website. This study will be Five aspects: rooms, location, cleanliness, check-in, and service. Sentiment labeling on multi-aspect data is done manually, where the sentiment will be labeled positive and negative based on each aspect of the hotel review.

This study aims to conduct multi-aspect sentiment analysis to help tourists choose the best hotel based on the desired category aspect. Then, conducted this research to determine the results of the classification and performance of the Random Forest, Support Vector Machine, and Naïve Bayes method based on Hybrid Classifier on several aspects of sentiment analysis based on hotel reviews.

The next section of this research is part 2, which will discuss the research methodology that will be used in this research, in section 3 discusses the discussion of the process carried out and the results obtained after conducting this research, and section 4 discusses the conclusions and suggestions after doing this research.

2. RESEARCH METHODOLOGY

2.1 Literature Review

In research [2], the Sentiment-Oriented Summarization-based aspects of Hotel Review revealed that assessing hotels based on aspects provides a better understanding than others according to user comments, and this can further help customers in the decision-making process which hotel to choose according to their needs and help management hotel because they will now realize which areas they need to improve and what their strengths are. A multi-aspect sentiment analysis regarding hotel reviews has been carried out in previous studies, where there are five category aspects in hotel reviews [1]. In this study, a multi-aspect sentiment analysis experiment was carried out using Latent Dirichlet Allocation (LDA) to determine hidden topics from the glossary, Semantic Similarity to categorize data, and a combination of Word Embedding and Long-short Term Memory (LSTM) for classification. In multi-aspect research using LDA + TF-ICF 100% + Semantic Similarity resulted in the highest F1- Measure 85% and Word Embedding + LSTM resulted in the F1- Measure 93% classification.

Research [4] conducted a sentiment analysis based on the hotel's aspects using the Support Vector Machine and Naive Bayes, which used the TF-IDF term weighting. Aspects reviewed in this study are location, room, food, price, and service, which are manually labeled positive and negative. In this research, using a Support Vector Machine results in better and more effective accuracy than Naive Bayes in all these aspects. The results are more than 70% accurate in all aspects except the food aspect. Multi aspects of sentiment analysis about hotels have also been carried out in this study [5]. Using PLSA + TF ICF 100% + Semantic Similarity, this research resulted in F1-Measure 84% on the five aspects of the hotel being reviewed, namely Location, Meal, Service, Comfort, and Cleanliness.

A hybrid method based on Random Forest and Support Vector Machine has been used in this research [6]. This study using product review data from Amazon, where the Random Forest approach improves performance in a small review and the Support Vector Machine improves performance when an extensive review works as a single hybrid approach. Accuracy of Random Forest classification is 81%, Support Vector Machine 82.4%, and Random Forest Support Vector Machine (Hybrid) 83.4% [6]. Hybrid use was also used in a movie review in this study [8]. Hybrid Naive Bayes - Genetic Algorithm gets an excellent accuracy of 93%. The comparison between individual classifiers and hybrid classifiers in this study shows that the hybrid classifiers significantly improve the single classifier. In the study [9], the hybrid classifier improved accuracy and achieved a significant breakthrough in reducing GPU processing power. Researched the Rule-Based, Lexicon Based, Machine Learning Approach and Hybrid Approach classifier in this study, and the Hybrid Approach obtained the highest F-score, namely 61.81%.

The method proposed in this paper is a Hybrid classification method based on three classifications, namely Random Forest, Support Vector Machine, and Naïve Bayes, which will be used in multi-aspect sentiment analysis.

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

A study conducted by Savita Sangam and Subhash Shinde using the Hybrid Support Vector Machine and ANN in the movie review successfully improved classification performance in sentiment analysis [7]. And in research [4], multi-aspect sentiment analysis helps determine hotel services' customer reviews.

2.2 Random Forest

Random Forest, which was formally proposed in 2001 by Leo Breiman and Adèle Cutler, is part of an automated learning technique. The Random Forest has several decision trees, and each decision tree will be total growth. It does not need to cut processing. The more tree it has, the more accurate the result will be, not overfitting [10]. Random Forest is part of a family collection method that takes the decision tree as an individual predictor. They are based on the Bagging, Randomizing Outputs, and Random Subspace methods, which forgive boosting [6]. Figure 4 is an example of Random Forest's representation using bootstrap to extract k samples from original training sets with N samples for k times, establish k decision trees, and vote according to all decision trees' classification results. We can describe the voting effect called confidence score in (1):

$$confidence\ score = tree_{number}(positive)/tree_{number}(total) \quad (1)$$

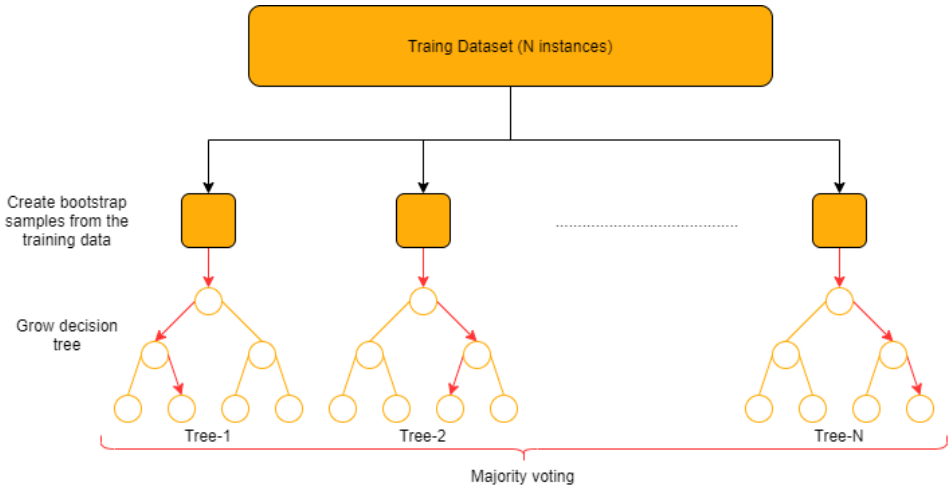

Figure 1. Representation of Random Forest

The Random Forest algorithm is one of the best among the classification algorithms [6] because it can accurately classify large amounts of data. Random forest uses a collection of decision trees to gather information. The calculation formula is presented in (2) and (3) [11]:

$$inf_{O_A}(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (2)$$

$$gain(A) = inf_O(D) - inf_{O_A}(D) \quad (3)$$

The output class is selected based on a majority vote i.e., the maximum number of similar courses produced by various trees is considered to be output from Random Forest [12].

2.3 Support Vector Machine

The Support Vector Machine method is a statistical classification approach based on maximizing the margin between instance and hyper-plane separation [6]. Support Vector Machine is one of a non-probabilistic binary linear classifier that can separate classes linearly by a large margin. It is one of the most potent classifiers capable of handling infinite-dimensional feature vectors. To maximization the margin in the Support Vector Machine, use formula (5):

$$min_w \frac{1}{2} (\|w\|)^2 \quad (4)$$

The limitation to maximalization the margin use formula (5):

$$y_i(x_i * w + b) - 1 \geq 0 \quad (5)$$

After that, calculate the elimination to get w and b use formula (6), (7), and (8):

$$x_i * w + b = 0 \quad (6)$$

$$x_i * w + b \geq +1 \qquad Untuk\ y_i = +1 \quad (7)$$

$$x_i * w + b \leq -1 \qquad Untuk\ y_i = -1 \quad (8)$$

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

x_i is the dimension vector, w is the weight vector, and b is the bias value, and y is the class. From equation (6) is the hyperplane to separation class positive and negative. Equation (7) is used in the study positive value and equation (7) to the negative class value. We will use the research a linear kernel because it is suitable and straightforward to use in data with many features.

2.4 Naïve Bayes

Naïve Bayes is one classification algorithm using probability and the statistical method by British scientist Thomas Bayes [4]. We will use this research Naïve Bayes Multinomial because Multinomial Naïve Bayes Classifier is a supervised learning method that uses probability and is focused on text classification cases [3]. The features of our data are in vector representation or discrete form. Also, this method strongly assumes the independence between the features themselves. In a document, the words distribution probability from word w_1 to w_n, for a given class c, can be calculated based on Bayes' formula [13], as shown in equation 6:

$$P(c|d) \propto P(c) \prod_{i=1}^{n_d} P(w_i|c) \quad (6)$$

$P(w_i|c)$ is the probability of some word in class c. $P(c)$ is prior probability in class c. $P(c|d)$is the probability class c in document d. The class determination is to compare the posterior probability results obtained, then the class with the most considerable posterior probability is the class chosen as the predicted result [3]. The prior probability formula:

$$P(c) = \frac{N_c}{N} \quad (7)$$

N_c is the sum of aspect c, and N is the sum of all aspects. The likelihood probability formula:

$$P(t_k|c) = \frac{T_{tc}}{\sum t' \in V^T ct'} \quad (8)$$

T_{tc} is the total probability of words in class c, and $\sum t' \in V^T ct''$ is the total probability of all words in class c.

2.5 Hybrid Classifier

The Hybrid Classifier method combines at least two techniques to improve the new technique's performance. The aim is to get the advantages of several combined techniques so that the Hybrid Classifier created has better accuracy. The Hybrid Classifier method is proposed in this paper, which uses ensemble learning from the Random Forest, Support Vector Machine, and Naïve Bayes method. The ensemble method is a meta-algorithm that combines many algorithms in classification to get a new model with better performance. Stacking is one of the ensemble models, which make a new model from combined predictions of 2 or more classifications.

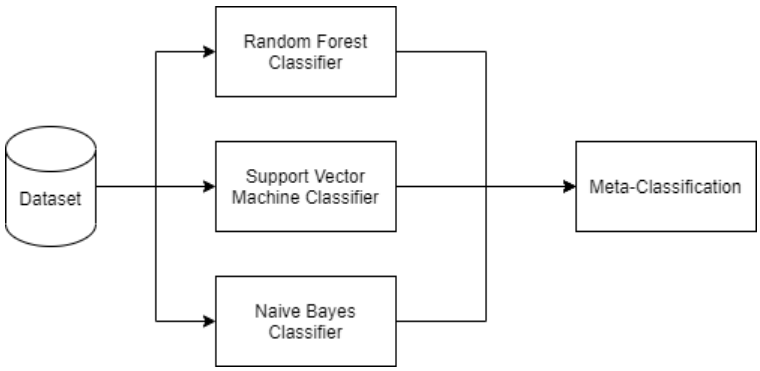

Figure 2. Design of Stacking Ensemble Method

2.6 Evaluation

Evaluate the performance of the system to measure the performance of the system. Measurement of system performance in this study will use accuracy, precision, and recall originating from the confusion matrix. The following is a formula for accuracy, precision, and recall:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (9)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (10)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (11)$$

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

For each combination, the element's presence is positive (P) or negative (N). The TP notation indicates True Positives: the number of samples predicted to be positive that is positive, FP indicates False Positives: the number of models expected to be positive that is negative, TN indicates True Negatives: the number of examples predicted to be opposing that are negative, and FN indicates False Negatives: the number of samples pessimistic predictions that is positive. Based on the TP value, FP, FN, and TN will get the accuracy, precision, and recall values. The accuracy value describes how accurately the system can classify data correctly. The precision value represents the accuracy between the input data and the prediction given by the system. Meanwhile, recall is the success rate of the system in recovering data information. The precision value describes the accuracy between the input data and the prediction given by the system. Meanwhile, recall is the success rate of the system in recovering data information.

3. RESULTS AND DISCUSSION

3.1 Design Scheme

In general, the steps for creating a system that performs sentiment analysis using Machine Learning are Dataset, Data Preprocessing, Feature Extraction, Modeling, and Evaluation. The course will produce a classification model that is used to classify the sentiment analysis. The following is the design scheme to be built:

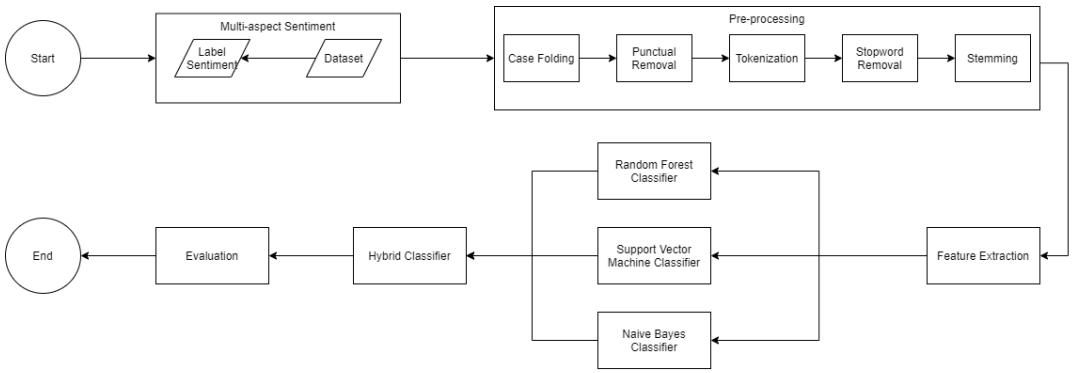

Figure 3. Design Scheme

3.2 Multi-aspect Sentiment

Sentiment Analysis or opinion mining is a computational study to identify and express opinions, sentiments, evaluations, attitudes, emotions, subjectivity, judgments, or views contained in a text [14]. Multi Aspects Sentiment Analysis is a detailed sentiment analysis technique using the category aspect of an opinion on each data. Using multi-aspect sentiment analysis, users can more easily determine something according to the category aspect that the user is desired. The main benefit of the multi-task framework is that it can mutually enhance aspect rating prediction between different aspects [15].

3.2.1 Dataset

The dataset used in this study is the dataset taken from the Kaggle website. This dataset is a dataset about Hotel Reviews based on reviews from hotel visitors from the *TripAdvisor* application. The amount of data from the dataset is 1500 reviews about hotels in English. Table 1 is an example of a hotel review data set that will be used.

Table 1. Dataset

Review
Very comfortable This is a very comfortable hotel located next to Universal. Walk or take the shuttle. The room on the 12th floor (we got an upgrade) was very nice with a great view. Would stay here again.
The hotel is a dirty and bad smell in the hotel. I will not recommend you to stay in this hotel!!!
Great service, location amazing, able walk Safeco field bit walk public market shopping, chose to walk space needle easily monorail couple blocks away. the rooms nice clean and comfortable. check-in also easy and professional

3.2.2 Label Sentiment

There will be 5 (five) kinds of aspects being reviewed in this research, namely Room, Location, Cleanliness, Check-in, and Service. Five aspects are based on hotel visitor reviews made on the TripAdvisor Hotel Review dataset from the Kaggle Website. Sentiment labeling is done manually on hotel review data based on the aspects

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

being reviewed. And in Table 2 are examples of data set labeling done where 1 represents Positive, and -1 represents Negative:

Table 2. Labeling Dataset

Review	Room	Location	Cleanliness	Check-in	Service
Very comfortable This is a very comfortable hotel located next to Universal. Walk or take the shuttle. The room on the 12th floor (we got an upgrade) was very nice with a great view. Would stay here again.	1	1			
The hotel is a dirty and bad smell in the hotel. I will not recommend you to stay in this hotel!!!			-1		
Great service, location amazing, able walk Safeco field bit walk public market shopping, chose to walk space needle easily monorail couple blocks away. the rooms nice clean and comfortable. check-in also easy and professional	1	1	1	1	1

3.3 Data Pre-processing

Pre-processing is done to get the best results by reducing noise in hotel review data. Pre-processing is a step taken to prepare data before analyzed in the sentiment classification process [4]. Pre-processing will handle imperfect data. This study's pre-processing process is case folding, punctual removal, tokenization, stop-word, and stemming.

Table 3. Description and Example of Pre-processing

Pre-processing	Description
Case Folding	Case Folding is the process of homogenizing letters into lowercase or lowercase letters. Examples on sentences " The hotel is a dirty and bad smell in the hotel. I will not recommend you to stay in this hotel!!!" becomes "the hotel is a dirty and bad smell in the hotel. i will not recommend you to stay in this hotel!!!"
Punctual Removal	Punctual removal is a process to get rid of punctuation marks. An Example of the sentence " the hotel is a dirty and bad smell in the hotel. i will not recommend you to stay in this hotel!!!" becomes "the hotel is a dirty and bad smell in the hotel i will not recommend you to stay in this hotel"
Tokenization	Tokenization is the process of separating a sentence into one part or word. Example of the sentence " the hotel is a dirty and bad smell in the hotel i will not recommend you to stay in this hotel " becomes "['the', 'hotel', 'is', 'a', 'dirty', 'and', 'bad', 'smell', 'in', 'the', 'hotel', 'i', 'will', 'not', 'recommend', 'you', 'to', 'stay', 'in', 'this', 'hotel']. "
Stop-word	Stop-word is the process of eliminating meaningless words. Example of the sentence "['the', 'hotel', 'is', 'a', 'dirty', 'and', 'bad', 'smell', 'in', 'the', 'hotel', 'i', 'will', 'not', 'recommend', 'you', 'to', 'stay', 'in', 'this', 'hotel']. " becomes "['hotel', 'dirty', 'bad', 'smell', 'hotel', 'recommend', 'stay', 'hotel']."
Stemming	Stemming is the process of removing affixes to words. Example of the sentence "['hotel', 'dirty', 'bad', 'smell', 'hotel', 'recommend', 'stay', 'hotel']." becomes "['hotel', 'dirty', 'bad', 'smell', 'hotel', 'recommend', 'stay', 'hotel']."

3.4 Feature Extraction

At this stage, the researcher uses the Bag of Words. Bag of Words is a simplified representation of natural language processing, where the text will be represented as a number in the document classification which then becomes a vector [16]. Use of classifications with Bag of Words to practice categorizing features.

Table 4. Representation of Bag of Words

Review	hotel	dirty	bad	smell	recommend	stay	Vector
R1	1	0	0	0	0	1	[1,0,0,0,0,1]
R2	3	1	1	1	1	1	[3,1,1,1,1,1]
R3	0	0	0	0	0	0	[0,0,0,0,0,0]

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

3.5 Model Classification

From the model we test to use the dataset has a positive review for Hotel Room with 578 and negative review for Room Hotel with 191, a positive review for Hotel Location with 633 and negative review for Hotel Location with 156, a positive review for Hotel Cleanliness with 371 and negative review for Hotel Cleanliness 155, positive reviews for Hotel Check-In 332, and negative reviews for Hotel Check-In 116, positive reviews for Hotel Service 711, and negative reviews about Hotel Service 185. The dataset we split 30% to use data test and 70% to data train with random state 0.

3.5.1 Using Random Forest

The following this algorithm to train using the Random Forest method:

Table 5. Algorithm Random Forest

Algorithm 1 Random Forest
For b = 1 to B Make
1 Draw a bootstrap sample Z* of size N from the training data.
2 Grow a random forest tree T_b to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree until the minimum node size n_{min} is reached.
- Select m variables at random from the p variables.
- Pick the best variable/split-point among the m.
- Split the node into two daughter nodes.
Output the ensemble of trees $\{\boldsymbol{Tb_1^B}\}$

3.5.2 Support Vector Machine

The following this algorithm to train using the Support Vector Machine method:

Table 6. Algorithm Support Vector Machine

Algorithm 2 Support Vector Machine		
Input:	I:	Input data
Output:	V:	Support vectors set
Begin		
Step 1:		Divide the given data set into two sets of data items having different class labels assigned to them
Step 2:		Add them to support vector set V
Step 3:		Loop the divided n data items
Step 4:		If a data item is not assigned any of the class labels, then add it to set V
Step 5:		Break if insufficient data items are found
Step 6:		end loop
Step 7:		Train using the derived SVM classifier model and test for validating the unlabeled data items.
End		

3.5.3 Naïve Bayes

The following this algorithm to train using the Multinomial Naïve Bayes method:

Table 7. Algorithm Multinomial Naïve Bayes

Algorithm 3 Multinomial Naïve Bayes
Train Multinomial Naïve Bayes (C,D)
1 V ← ExtractVocabulary(D)
2 N ← CountDocs(D)
3 **For each** c ∈ C
4 **Do** N_c ← CountDocsInClass(D,c)
5 Prior[c] ← N_c/N
6 $text_c \leftarrow CouncateTextOfAllDocsInClass(D,c)$
7 $\boldsymbol{for\ each}\ t \in V$
8 $\boldsymbol{do}\ T_{ct} \leftarrow CountTokensOfTerm(text_x, t)$
9 $\boldsymbol{for\ each}\ t \in V$
10 $\boldsymbol{do}\ condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

11 **Return** V, prior, coundprob

Apply Multinomial Naïve Bayes(C,V,Prior,condprob,d)

1 $W \leftarrow ExtractTokensFromDoc(V, d)$
2 ***for each*** $c \in C$
3 ***do*** $score[c] \leftarrow \log prior[c]$
4 ***for each*** $t \in W$
5 ***do*** $score[c] += \log condprob[t][c]$
6 **Return** arg $max_{c \in C} score[c]$

3.5.4 Hybrid Classifier

The following this algorithm to train using the Hybrid Classifier method:

Table 8. Algorithm Hybrid RF-SVM-NB

Algorithm 4 Hybrid RF-SVM-NB

Input:
1 D, a set of tuples d.
2 k = 3, the number of models in the ensemble.
3 Basic Classifier (Random Forest, Support Vector Machine, Naïve Bayes)

Output: Hybrid RF-SVM-NB Model, M *

Procedure:
1 For i = 1 to k do // Create k models
2 Create a new training dataset, Di with sampling D with replacement. The same example of a given dataset D can occur more than once in the Di training dataset
3 Use D_i to get the model, M_i
4 Classify each sample d in the training data D_i and initialize its weight, W_i for the model, M_i based on the accuracy of the percentage of samples that are correctly classified in the training data D_i
5 Endfor

To use a hybrid model on a tuple, X:
1 If classification then
2 let each k model classify X and return the majority vote;
3 If the predictions then
4 let each model k predict the value for X and return the average predicted value;

3.5.4 Evaluation

The final result in this research, shown in Table 9 about the Precision, Recall, F1-Score, and Accuracy in every classification aspect from data. The result almost reached 80% in all aspects except Check-In using SVM Classification.

Table 9. Evaluation of Classifier

Aspect	Positive	Negative	Classification	Precision	Recall	F1-Score	Accuracy
Room	578	191	SVM	0.75	0.78	0.76	0.83
			RF	0.81	0.72	0.75	0.85
			NB	0.77	0.79	0.78	0,84
			Hybrid	0.76	0.77	0.77	0.84
Location	578	191	SVM	0.73	0.73	0.73	0.81
			RF	0.79	0.72	0.74	0.80
			NB	0.70	0.69	0.69	0.80
			Hybrid	0.77	0.70	0.72	0.84
Cleanliness	578	191	SVM	0.81	0.81	0.81	0.84
			RF	0.79	0.72	0.74	0.80
			NB	0.83	0.80	0.81	0.85
			Hybrid	0.84	0.79	0.81	0.85
Check-in	578	191	SVM	0.65	0.66	0.66	0.76
			RF	0.74	0.59	0.60	0.80
			NB	0.72	0.74	0.73	0.80
			Hybrid	0.74	0.74	0.74	0.82
			SVM	0.72	0.71	0.71	0.81
			RF	0.84	0.60	0.61	0.82

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

			NB	0.76	0.74	0.75	0.83
			Hybrid	0.79	0.75	0.77	0.85
Service	578	191					

From all table in every method is the result of classification we can average with equation (12):

$$Average = \frac{Total\ sum\ of\ all\ accuracy\ aspects}{Number\ of\ aspects} \quad (12)$$

The final result is represented in the table about the accuracy. Using Random Forest in multi-aspect data gets average accuracy of 81%. Using Support Vector Machine in multi-aspect data gets average accuracy of 82.2%. Naïve Bayes in multi-aspect data gets average accuracy of 82.4% and uses Hybrid Classifier in multi-aspect data get average accuracy of 84%. Using Random Forest in aspect data Rooms is better than Hybrid. Still, Hybrid classification using Random Forests, Support Vector Machine, and Naïve Bayes get a better result than using a single classification method if we average all accuracy result in every aspect of data. A hybrid Classifier can improve the accuracy and performance of classification. From Figure 4, it is evident that Random Forest Support Vector Machine Naive Bayes based on Hybrid Classifier shows the best performance compared to other studied algorithms.

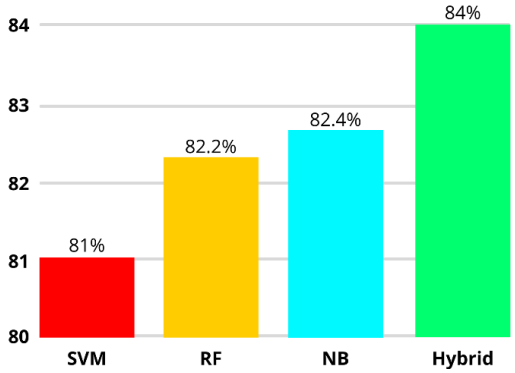

Figure 4. Average Classification Method Result

4. CONCLUSION

Multi-Aspect Sentiment Analysis is a technique of analyzing a person's opinion or judgment, specifically using certain aspects of each data category. In this study, there are five aspects of the hotel category in the data being reviewed: room, location, cleanliness, check-in, and service. The dataset has positive reviews for Hotel Room with 578 and negative reviews for Hotel Room with 191, positive review for Hotel Location with 633 and negative review for Hotel Location with 156, positive review for Hotel Cleanliness with 371 and negative review for Hotel Cleanliness 155, reviews positive reviews for Hotel Check-In 332, and negative reviews for Hotel Check-In 116, positive reviews for 711 Hotel Services, and negative reviews about Hotel Services 185. The methods used in this study are Random Forest, Support Vector Machine, Naïve Bayes based on Hybrid Classifier. The Random Forest method results produce an average accuracy of 82.2%, which is better than the Support Vector Machine method with an average accuracy of 81%. The best result from the three classification methods is Naïve Bayes, with an average accuracy of 82.4%. Hybrid use in this study gets better results than using a single classification method with an average accuracy of 84%. It can prove that hybrids can improve the final accuracy of the classification and performance in multi-aspect sentiment analysis data. This research also proves again from previous research that the hybrid classifier can improve accuracy and apply to the case study conducted in this study, namely using multi-aspect sentiment data. We can make some suggestions in further research; namely, the data can use Indonesian and use feature weighting with Term Frequency - Inverse Document Frequency (TF-IDF). Further research can also add new classification methods to be used in hybrid ways.

REFERENCES

[1] R. A. Priyantina and R. Sarno, "Sentiment analysis of hotel reviews using Latent Dirichlet Allocation, semantic similarity and LSTM," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 4, pp. 142–155, 2019, doi: 10.22266/ijies2019.0831.14.
[2] N. Akhtar, N. Zubair, A. Kumar, and T. Ahmad, "Aspect based Sentiment Oriented Summarization of Hotel Reviews," *Procedia Comput. Sci.*, vol. 115, no. May 2020, pp. 563–571, 2017, doi: 10.1016/j.procs.2017.09.115.
[3] A. A. Farisi, Y. Sibaroni, and S. Al Faraby, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012024.
[4] F. A. Bachtiar, W. Paulina, and A. N. Rusydi, "Text Mining for Aspect Based Sentiment Analysis on Customer Review : a Case Study in the Hotel Industry," *5th Int. Work. Innov. Inf. Commun. Sci. Technol.*, no. March, 2020.
[5] D. A. K. Khotimah and R. Sarno, "Sentiment analysis of hotel aspect using probabilistic latent semantic analysis, word embedding and LSTM," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 4, pp. 275–290, 2019, doi: 10.22266/ijies2019.0831.26.

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 5, Nomor 2, April 2021, Page 630-639
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i2.2959

[6] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Comput. Sci.*, vol. 127, pp. 511–520, 2018, doi: 10.1016/j.procs.2018.01.150.
[7] S. Sangam and S. Shinde, "Sentiment classification of social media reviews using an ensemble classifier," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 1, p. 355, 2019, doi: 10.11591/ijeecs.v16.i1.pp355-363.
[8] M. Govindarajan, "Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm," *Int. J. Adv. Comput. Res.*, no. 4, pp. 2277–7970, 2013, [Online]. Available: http://imdb.com.
[9] A. Sharma, A. Sharma, R. K. Singh, and M. D. Upadhayay, "Hybrid Classifier for Sentiment Analysis Using Effective Pipelining," *Int. Res. J. Eng. Technol.*, vol. 4, no. 8, pp. 2276–2281, 2017, [Online]. Available: https://irjet.net/archives/V4/i8/IRJET-V4I8411.pdf.
[10] Z. Wu, W. Lin, Z. Zhang, A. Wen, and L. Lin, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," *Proc. - 2017 IEEE Int. Conf. Comput. Sci. Eng. IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. CSE EUC 2017*, vol. 1, pp. 531–536, 2017, doi: 10.1109/CSE-EUC.2017.99.
[11] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, "A novel hybrid classification approach for sentiment analysis of text document," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, pp. 4554–4567, 2018, doi: 10.11591/ijece.v8i6.pp4554-4567.
[12] M. M and S. Mehla, "Sentiment Analysis of Movie Reviews using Machine Learning Classifiers," *Int. J. Comput. Appl.*, vol. 182, no. 50, pp. 25–28, 2019, doi: 10.5120/ijca2019918756.
[13] S. L. Mahfiz and A. Romadhony, "Aspect-based Opinion Mining on Beauty Product Reviews," *2020 3rd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2020*, pp. 488–493, 2020, doi: 10.1109/ISRITI51436.2020.9315350.
[14] Y. Lin, X. Wang, and A. Zhou, "Opinion spam detection," *Opin. Anal. Online Rev.*, no. May, pp. 79–94, 2016, doi: 10.1142/9789813100459_0007.
[15] J. Li, H. Yang, and C. Zong, "Document-level Multi-aspect Sentiment Classification by Jointly Modeling Users, Aspects, and Overall Ratings," *Proc. 27th Int. Conf. Comput. Linguist.*, pp. 925–936, 2018, [Online]. Available: https://www.tripadvisor.com/.
[16] F. F. Rahmawati and Y. Sibaroni, "Multi-Aspect Sentiment Analysis pada Destinasi Pariwisata Yogyakarta Menggunakan Support Vector Machine dan Particle Swarm Optimization sebagai Seleksi Fitur," 2019.

Submitted: **27/03/2021**; Accepted: **19/04/2021**; Published: **25/04/2021**